



Quantification post-hoc de l'incertitude prédictive : méthodes avec applications à la prévision des prix de l'électricité

Margaux Zaffran

► To cite this version:

Margaux Zaffran. Quantification post-hoc de l'incertitude prédictive : méthodes avec applications à la prévision des prix de l'électricité. Statistics [math.ST]. Institut Polytechnique de Paris, 2024. English. NNT : 2024IPPAX033 . tel-04720002

HAL Id: tel-04720002

<https://theses.hal.science/tel-04720002v1>

Submitted on 3 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2024IPPAX033

Thèse de doctorat



Post-hoc predictive uncertainty quantification: methods with applications to electricity price forecasting

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Paris, le 25 juin 2024, par

MARGAUX ZAFFRAN

Composition du Jury :

Florence Forbes Directrice de recherche, INRIA (STATIFY)	Présidente
Pierre Pinson Professeur, Imperial College London (Dyson School of Design Engineering)	Rapporteur
Étienne Roquain Maître de conférences HdR, Sorbonne Université (LPSM)	Rapporteur
Emmanuel Candès Professeur, Stanford (Departments of Mathematics and Statistics)	Examineur
Éric Moulines Professeur, École polytechnique (CMAP)	Examineur
Aaditya Ramdas Professeur Assistant, Carnegie Mellon University (Departments of Statistics and Machine Learning)	Examineur
Julie Josse Directrice de recherche, INRIA (PreMeDICaL)	Directrice de thèse
Aymeric Dieuleveut Professeur, École polytechnique (CMAP)	Co-directeur de thèse
Olivier Féron Chercheur sénior, EDF R&D (OSIRIS)	Invité
Yannig Goude Chercheur sénior, EDF R&D (OSIRIS)	Invité

*À Joy,
On croyait en toi
probablement plus que toi,
Tu croyais en moi
assurément plus que moi.*

Abstract

The increasing use of renewable intermittent energy leads to more dependent and volatile energy markets. Therefore, an accurate electricity price forecasting is required to stabilize energy production planning, thus reducing the associated carbon emissions. The surge of more and more powerful statistical algorithms and machine learning offers promising prospects for tackling this problem. However, these methods provide ad hoc forecasts, with no indication of the degree of confidence to be placed in them. To ensure the trust of key actors in energy markets with regard to such decision-support tools, it is crucial to quantify their predictive uncertainty. This thesis focuses on developing predictive intervals for any underlying algorithm, including neural networks, without assumptions on the latter. While motivated by the electrical sector, the methods developed are generic: they can be applied in many sensitive fields.

Split Conformal Prediction (SCP, [Vovk et al., 2005](#); [Papadopoulos et al., 2002](#); [Lei et al., 2018](#)) is a versatile procedure associating predictive intervals with any prediction model. Unlike existing probabilistic prediction methods, SCP is highly promising as it offers theoretical guarantees with finite sample size, under the sole distributional assumption that the data are exchangeable (i.e. the data distribution is invariant to permutation, a weaker assumption than independency with identical distribution).

Formally, suppose we have n data $(X_i, Y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$, where Y is the response variable (e.g., electricity price) and $X \in \mathbb{R}^d$ the d covariates (e.g., productions). The user sets a *miscoverage rate* $\alpha \in [0, 1]$ (typically 0.1 or 0.05). SCP constructs a predictive interval $\mathcal{C}_{n,\alpha}$ such that $\mathbb{P}\{Y_{n+1} \in \mathcal{C}_{n,\alpha}(X_{n+1})\} \geq 1 - \alpha$: $\mathcal{C}_{n,\alpha}$ is said to be marginally *valid*. Its length must be as small as possible to be informative (*efficient*). An example of such an interval is given in Figure A.

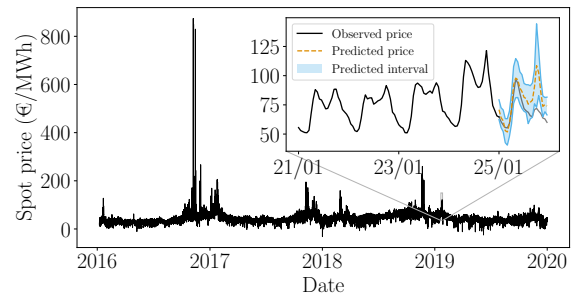


Figure A: predictive intervals for electricity prices.

However, SCP is not applicable on time series (such as electricity prices) as they are not exchangeable due to temporal dependence. To address this limitation, a first approach ([Gibbs and Candès, 2021](#)) relies on using an adaptive miscoverage rate α_t , that is updated according to previous performances and to an hyper-parameter $\gamma > 0$, playing the role of a learning rate. Using Markov Chain theory, the first contribution of this thesis analyzes the influence of γ on the efficiency of the resulting intervals. It allowed to propose a novel method

not requiring the choice of γ —and therefore usable in practice—based on online expert aggregation. Following the electricity prices explosion in 2021, the second contribution of this thesis investigates the impact of this higher non-stationarity on probabilistic forecasts, and the improvements brought by different adaptive post-hoc layers such as SCP and online aggregation.

Still, to improve electricity price point forecasts, one could leverage the emergence of open data platforms to integrate more explanatory variables such as commodity prices, or prices from other correlated markets. However, aggregating different data sources comes with methodological challenges, such as dealing with missing values, as time frequencies and market horizons may differ. Missing data are common in statistical practice and, paradoxically, their occurrence increases with the quantity of data.

A usual way to get point predictions is to replace the missing values (NAs) by plausible values and then apply any learning algorithm on the completed data. Yet, there was no method for quantifying predictive uncertainty with NAs. The third and forth contributions of this thesis show that SCP applied on an imputed data set enjoys the exact same marginal *validity* guarantees it would on a complete dataset. The strength of this result lies in its generality: it implies that the user can choose any imputation, even a naive one, without affecting the validity of the intervals, even for informative NAs (a complex and rarely studied scenario). However, The third and forth contributions of this thesis identify that NAs generate heteroskedasticity: the validity of the intervals depends on which covariates are observed. They propose the first algorithms to solve this problem, that are extremely simple to implement. Theoretically grounded, the assumptions on which they rely are nearly minimal according to hardness results.

Résumé

L'utilisation croissante d'énergies renouvelables intermittentes rend les marchés de l'énergie plus dépendants et plus volatils. Par conséquent, une prévision précise du prix de l'électricité est nécessaire afin de stabiliser la planification de la production d'énergie et réduire ainsi les émissions de carbone associées. L'essor d'algorithmes statistiques et de l'apprentissage automatique de plus en plus puissants offre des perspectives prometteuses pour traiter ce problème. Cependant, ces méthodes fournissent des prévisions ad hoc, sans indication du degré de confiance à leur accorder. Pour garantir la confiance des acteurs des marchés de l'énergie à l'égard de ces outils d'aide à la décision, il est crucial de quantifier leur incertitude prédictive. Cette thèse porte sur le développement d'intervalles prédictifs pour tout algorithme de prédiction, y compris les réseaux neuronaux, sans hypothèses sur ce dernier. Bien que motivées par le secteur électrique, les méthodes développées sont génériques : elles peuvent être appliquées dans de nombreux autres domaines sensibles.

La prédiction conforme par partition (SCP, [Vovk et al., 2005](#); [Papadopoulos et al., 2002](#); [Lei et al., 2018](#)) est une procédure polyvalente associant des intervalles prédictifs à tout modèle de prédiction. Contrairement aux méthodes de prédiction probabilistes existantes, CP est hautement prometteuse car elle offre des garanties théoriques à taille d'échantillon finie, sous la seule hypothèse distributionnelle que les données sont échangeables (c'est-à-dire que la distribution des données est invariante par permutation, ce qui est plus faible que des données indépendantes et identiquement distribuées).

Formellement, supposons que nous disposons de n données $(X_i, Y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ où Y est la variable à prédire (e.g., le prix de l'électricité) et $X \in \mathbb{R}^d$ les d covariables (e.g., les productions). L'utilisateur fixe un *taux de non-couverture* $\alpha \in [0, 1]$ (typiquement 0.1 ou 0.05). SCP construit un intervalle prédictif $\mathcal{C}_{n,\alpha}$ tel que $\mathbb{P}\{Y_{n+1} \in \mathcal{C}_{n,\alpha}(X_{n+1})\} \geq 1 - \alpha$: on dit que $\mathcal{C}_{n,\alpha}$ est *valide* marginalement. Sa longueur doit être la plus petite possible pour qu'il soit informatif (*efficace*). Un exemple de tel intervalle est donné en Figure B.

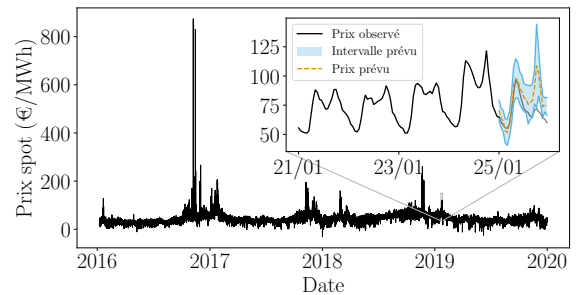


Figure B: intervalles prédictifs pour les prix de l'électricité.

Cependant, SCP n'est pas applicable sur une séries temporelles (telles que les prix de l'électricité) car elles ne sont pas échangeables en raison de leur dépendance temporelle. Pour remédier à cette limitation, une première approche ([Gibbs and Candès, 2021](#)) repose

sur l'utilisation d'un taux de non-couverture adaptatif α_t , qui est mis à jour en fonction des performances passées et d'un hyperparamètre $\gamma > 0$, jouant le rôle d'un taux d'apprentissage. En utilisant la théorie des chaînes de Markov, la première contribution de cette thèse analyse l'influence de γ sur l'efficacité des intervalles prédictifs associés. Cela a permis de proposer une nouvelle méthode ne nécessitant pas le choix de γ —et donc utilisable en pratique—basée sur l'agrégation d'experts en ligne. Suite à l'explosion des prix de l'électricité en 2021, la deuxième contribution de cette thèse étudie l'impact de cette non-stationnarité accrue sur les prévisions probabilistes, et les améliorations apportées par différentes surcouches adaptatives telles que SCP et l'agrégation en ligne.

Néanmoins, pour améliorer les prévisions des prix de l'électricité, nous pourrions tirer parti de l'émergence de plateformes de données ouvertes pour intégrer davantage de variables explicatives telles que les prix des matières premières ou les prix d'autres marchés corrélés. Cependant, l'agrégation de différentes sources de données s'accompagne de défis méthodologiques, tels que le traitement des valeurs manquantes, comme les fréquences temporelles et les horizons de marché peuvent différer. Les données manquantes sont courantes dans la pratique statistique et, paradoxalement, leur nombre augmente avec la quantité de données.

Une approche traditionnelle pour obtenir des prédictions ponctuelles consiste à remplacer (imputer) les valeurs manquantes (NAs) par des valeurs plausibles, puis à entraîner n'importe quel algorithme d'apprentissage sur les données complétées. Cependant, il n'existe aucune méthode permettant de quantifier l'incertitude prédictive avec les NAs. Les troisième et quatrième contributions de cette thèse montrent que SCP appliquée à un jeu de données imputé bénéficie exactement des mêmes garanties de *validité* marginales que sur des données complètes. La force de ce résultat réside dans sa généralité : il implique que l'utilisateur peut choisir n'importe quelle imputation, même naïve, sans affecter la validité des intervalles, même pour des NAs informatives (un scénario complexe et rarement étudié). Cependant, Les troisième et quatrième contributions de cette thèse constatent que les NA génèrent de l'hétéroscédasticité : la validité des intervalles dépend de quelles variables explicatives sont observées. Ils proposent les premiers algorithmes pour résoudre ce problème, qui sont extrêmement simples à mettre en pratique. Théoriquement valides, les hypothèses sur lesquelles ils reposent sont presque minimales d'après de nouveaux résultats d'impossibilité.

Remerciements

La recherche, et en particulier la thèse, est finalement une aventure grandement collective. Évidemment, mentors et collaborateurs façonnent les travaux de recherche, mais pas seulement : les cafés, les trajets en train pour se rendre en conférence, les inspirations prises de lectures ou d'exposés pédagogiques, les étudiants dont les questions défient toutes attentes, et les amitiés nouées en chemin également. Alors, pour clôturer le chapitre de la thèse, commençons par les canoniques remerciements. Ce chapitre sera sans doute le plus lu de cette thèse, je présente donc d'avance mes excuses si quelqu'un se sent oublié : paradoxalement ce sera également le moins relu de cette thèse !

Au commencement, il y avait Yannig. Professeur quand j'étais sur les bancs de l'ENSTA, je ne pouvais rêver mieux que l'entrée dans le milieu de la recherche que tu m'as permis d'effectuer à Bristol. J'ai ensuite eu la chance d'apprendre à tes côtés à EDF ce qui m'a définitivement orientée vers l'objectif d'une thèse CIFRE avec EDF. Et là encore, tu as insisté : non convaincue initialement par l'idée d'un sujet proche de la finance, tu es revenu m'en parler et tu m'as convaincue. Pour tout ça, déjà, merci. Merci également pour ta vision scientifique fine, toujours concrète mais rigoureuse. Enfin, merci pour ta bonne humeur et ta positivité en toutes circonstances, tes conseils avisés, tes coups de boost, et ta tranquillité inébranlable.

Le binôme EDF était au complet avec Olivier. Olivier, ton écoute et ton attention m'ont profondément touchée tout au long de cette thèse. Merci pour le soin et la protection que tu accordes à tes doctorants. Tu as toujours été vigilant à prendre des décisions dans mon intérêt, et à réfléchir sur le long terme. Merci également de m'avoir initiée aux marchés de l'énergie, d'avoir répondu à toutes mes questions même quand elles se répétaient ou qu'elles étaient naïves. Sans ton apport expert, cette thèse serait à mes yeux incomplète car déconnectée de considérations concrètes.

Julie, merci pour tout. Tu es un vrai modèle sur tous les aspects, et je suis plus que ravie de t'avoir fait confiance dès notre première discussion. Scientifiquement, ta vision interface parfaitement théorie et pratique utile, avec une précision mathématique et expérimentale fine, qui se retranscrivent dans ton amour de la rédaction. Tout ça en créant une équipe dynamique comme aucune autre, et en accordant sans faille du temps à tes étudiants. Je suis encore ébahie par ta capacité de concentration en visio et en talks. Au-delà de tout ça, ton énergie débordante, ton organisation indéfectible, ton envie de transmission, ton écoute et ta compréhension (sur-)humaines, ta présence malgré ta distance, sont des qualités que j'admire (et je ne pense pas être la seule !). Merci.

Last but not least, Aymeric. Parfois, je me dis que tu ne t'attendais pas à passer autant de temps sur cette thèse quand tu en as accepté la co-direction. J'espère que tu ne regrettes pas, et je te remercie, vraiment. Merci de m'avoir accompagnée au plus près quand tout a débuté, et de m'avoir rassurée quand je doutais. Merci de transmettre ta passion, ton émerveillement, et ton excitation sur des sujets (mathématiques mais pas que) qui t'enthousiasment : c'est contagieux. Merci aussi pour les longues sessions au tableau, à se casser les dents sur des problèmes dont on oublie la formulation quand on est au milieu du chemin. Merci de ne jamais laisser tomber et de toujours y croire. Merci pour la méta-science et les réflexions sur la recherche, sur la gestion d'équipe ou de labos, et sur notre place en tant que chercheurs dans la société. Merci pour tous les à-côtés et les bons moments passés ensemble.

Pour reprendre les mots de Julie, "ça ne pouvait pas être mieux". Sans trop le savoir quand le montage de la thèse s'est fait, la complémentarité entre vous quatre était vraiment parfaite. Merci d'avoir créé cette équipe de choc et de m'y avoir accueillie. Merci aussi de m'avoir fait confiance quand on a ré-orienté cette thèse. J'ai hâte de voir l'avenir de nos échanges !

I would like to deeply thank my jury members. I am truly honored to have all of you in my jury, and, to be honest, a bit impressed too. First, I am indebted to you, Pierre and Étienne, for your precious reports. Pierre, thanks for your broad point of view and important connections to reality. Étienne, thanks for your extremely careful reading of (all of) my manuscript and precise comments. Then, many thanks to the examiners: Emmanuel, my whole PhD would be critically different if I did not attend one of your (excellent) talks; Florence, I am grateful to have you as President and in-person; Éric, you supported me in my professional evolution as a researcher; and Aaditya, I have to say that I always try to follow your advice and spread the word around me about them. For all of this, I feel really lucky.

Ryan and Aaditya, I can't wait to work with you on exciting topics! Many thanks for this amazing opportunity. I am also looking forward to join your teams: you both seem to be creating fruitful environments where everyone flourishes as human being.

Indeed, I had the amazing chance to meet your students during a visiting trip to the US or in conferences. These events definitely opened my mind to new horizons. Thanks Madeleine and your team for hosting me in Stanford. Thanks to all the people who spent time with me, whether in Stanford, Berkeley, or Pittsburgh, all of our discussions shaped my perspectives (professionally and personally).

A year before, I met Yaniv who hosted me for three wonderful months. Again, a huge thanks to you and your students for integrating me in the team, I came back with new vision on research. It was really an immense pleasure to work with you, and I hope we have more opportunities in the future to pursue our collaboration. I also came back with new friends (to quote one of them "it is not because it fades away that it did not exist"), and, sincerely, this trip have been a life changer on so many aspects. I am grateful for it to have

existed. Thank you Meshi for being the best sergeant I could have hoped for, and thanks to the office mates (or their convex hull): Omer, Hagay, Noa, Naseem, Uri, Noam, and Nimrod. Morning breaks, ice creams, matches, trips and pomodoros, among other things, are cherished moments. Gilboa 23, what a team of roommates. In particular, Yeela, thank you for letting me your room; Hen, thank you for welcoming me, for deep discussions, and for sweet chocolate; Johanna, it was too short but I am so glad I met you, and Michael too. Il y a eu quelques français lors de ce séjour, et pas des moindres. Merci à Ruthy et Robert pour votre accueil à bras ouvert du début, qui s'est transformé en deuxième famille tout du long. Merci à Muriel, Gérard et toute votre famille pour ces quelques jours de fêtes ensemble, Muriel, nos discussions ont impacté ma fin de thèse sans aucun doute.

Ces voyages ont en partie été possibles grâce à la Fondation L'Oréal – UNESCO, que je remercie de nouveau pour son soutien, et à l'association Séphora Berrebi. Emmanuelle et Gérard, je tiens encore à vous remercier chaleureusement, au-delà du voyage que vous avez permis de réaliser. Vous rencontrer et discuter avec vous a été une source d'inspiration et de confiance en l'avenir inébranlables. Je vous admire sincèrement, et je suis d'autant plus touchée et ravie d'avoir eu votre soutien. Merci.

La transmission auprès de plus jeunes est quelque chose qui me tient à coeur, et cela a – je crois – un effet miroir : merci aux lycéennes rencontrées et aux étudiants auprès de qui j'ai eu la chance d'enseigner, vos sourires, vos questions, et votre confiance sont des vecteurs de motivation et de réflexion impressionnants. Merci également à Grégoire, premier stagiaire formidable, c'était un vrai plaisir de travailler avec toi.

Dans l'autre sens, ce bout de chemin a pu exister car j'ai rencontré des enseignants et mentors incroyables. En particulier, merci à Marseilleveyre et Monsieur Khelladi, Madame De Redon, Monsieur Torregrosa, Monsieur Jego et Monsieur Milliard. Ensuite, un très grand merci encore à Jeanne Nguyen, excellente chargée de TD de statistiques, sans qui je ne ferais vraisemblablement pas des stats aujourd'hui ! Then, a thousand of thanks to Matteo, from whom I learned so much during my first research experience. Enfin, merci à Philippe, pour un stage de master particulier (2020 oblige) et pourtant une super expérience : ton encadrement était plus que parfait pour moi, merci pour m'avoir partagé tes savoirs, points de vue, et ton enthousiasme serein.

Il y a les professeurs, il y a les étudiants, mais il y a également les collègues qui font partie de cet environnement enrichissant et épanouissant. Alors, sans vous citer car pour sûr j'en oublierai, d'abord merci à toutes les personnes rencontrées en conférences, ici ou là, autour d'un café, d'un poster, d'un exposé ou d'un verre. Toutes nos discussions sont des trampolines pour avancer, pour apprécier la vie de chercheur, et pour démarrer la journée avec le sourire. Merci aux "plus vieux" qui prennent du temps avec nous. Merci aussi au groupe Jeunes Statisticien.ne.s de la SFdS et à la SFdS toute entière, et même plus encore, pour la bonne humeur et pour tout ce dont la communauté bénéficie grâce à vous, et notamment pour nous permettre de donner une voix à des sujets primordiaux qui me tiennent à coeur. Merci au groupe de travail conformal d'Orsay de m'avoir adoptée,

et en particulier à Pierre et Jean-Baptiste. Merci Claire pour le temps passé à travailler ensemble et pour avoir canalisé mes couleurs extravagantes.

Ensuite, les collègues du quotidien. Mes remerciements s'éternisent, et ça n'est pas prêt de s'arrêter : une thèse entre trois instituts, ça fait du monde !

Au tout début, bien sûr, EDF. Bulle d'oxygène au début de ma thèse confinée, R33, R39 et OSIRIS, merci. Pour la suite également, pour les sujets passionnants que vous offrez, pour le recul que vous apportez, et pour les habitudes. Quelques encarts à la non-citation : merci d'abord à Édouard d'avoir cru en moi et de m'avoir montré la valeur d'un bon chef, merci à Clémence pour les conseils avisés et le smile, merci à Maximilien, la machine, nos routes se sont surprenamment suivies de la prépa jusqu'au bureau à R33 (et ça va s'arrêter ici ;)), et enfin merci à Laura, Thomas et Pierre, accolytes doctorants d'Olivier, pour des discussions toujours chouettes.

Ensuite, depuis le début aussi, l'INRIA et l'équipe de Julie, devenue PreMeDiCaL depuis. Thank you all for creating such a dynamic team in which it is a pleasure to work in. Aude, Imke et Bénédicte en particulier, vous avoir côtoyées en distanciel, au jour le jour, ou en conférences, a été une chance. Réfléchir avec vous sur comment faire de la recherche et quelle est notre place a été précieux. Merci aux doctorants (ou pas) de Zénith pour leur accueil dès que je passais sur Montpellier.

Finalement, le CMAP. J'y suis arrivée un peu sur le tard, mais j'y suis pour de bon maintenant. Merci à tous, que vous soyez en thèse, en post-doc ou permanents. Pour le coup, ici, je ne vais vraiment pas vous citer, on est bien trop nombreux pour que je ne fasse pas de faux pas. Merci à chacun d'entre vous pour tout ce que nous avons pu partager, et notamment pour les repas (trop tôt). Mention spéciale à l'équipe SIMPAS. Pour les moments partagés avec des chaussures de rando, de ski, un maillot, des cartes ou des cordes vocales bien utilisées, de jour comme de nuit, merci (au moins) à Achille, Antoine, Aymeric, Baptiste, Baptiste, Benjamin, Christoph, Clément, Clément, Constantin, Erwan, Erwan, Guillaume, Jean, Louis, Mahdi, Manon, Marylou, Maxence, Pablo, Paul, Pierre, Renaud, Solange, Thomas, Tom, Vincent. Merci tout particulier à l'équipe d'Aymeric (Constantin, Baptiste, Alexis, Renaud, Rémi, Damien, Mahmoud, Jean-Baptiste) pour l'entraide permanente, les relectures nombreuses et les répétitions.

Parce qu'en recherche il y a une certaine proximité que je ne saurai expliquer, certains d'entre vous sont devenus bien plus que des compagnons de route. Manon, tu m'impressionnes de courage, de détermination, et de verre à moitié plein. Solange, colocationner avec toi n'est pas la meilleure idée pour suivre les exposés le lendemain mais qu'importe. Baptiste, nos rendez-vous du lundi soir me manquent déjà. Arthur, je recharge mes batteries quand je vois l'énergie que tu mets dans les combats qui comptent, qu'ils soient personnels ou sociétaux. Marie, boule de gentillesse pleine de pep's, je t'attends pour le prochain karaoké. Constantin, nounours à la grosse voix, tu as vraisemblablement une vocation pour raconter des histoires loufoques qui font du bien. Baptiste, tu es le meilleur remonte-morale qui existe et une machine à idées et conseils en tout genre. Alexis, je suis heureuse d'avoir pu découvrir ton univers et tes petits plats. Renaud, co-organisateur

de GM, je suis ravie que la relève soit assurée également s'agissant des valeurs à défendre.

Merci à mes coupaings d'ailleurs, qui supportent ma nullité en messages (et ce n'est pas peu dire). Promis, j'essaie de m'améliorer ! Mention spéciale ici aux filles de la Palaisienne : vous m'avez adoptée il y a maintenant 5 ans, et c'est non sans un pincement au coeur que cette aventure marque une pause. En rentrant dans le gymnase, je vous retrouvais les mardi et jeudi soirs, et ça n'avait pas de prix. Merci pour la nécessaire déconnexion, pour les rires et la sueur. Merci Marine d'avoir été une coach en or. Merci à la meilleure équipe : Elena, copine bien avant d'être co-équipière, Flore, Rire et Chansons sur pattes, Marie-Ange, duotte au top, et Marion, au dévouement inébranlable.

Pour finir, merci à Carl pour tes traits d'esprit, ton humour fin, et les petites choses qui ont fait le quotidien plus rose. Merci également à Théo pour les dessins malgré le fait qu'on ne se comprenait pas toujours, ils sont magnifiques. Merci à Sylvie et Pierre pour votre accueil toujours à bras ouverts et votre écoute attentive. Tous ces moments m'ont ressourcée et restent gravés.

Évidemment, merci à ma famille, qui m'a permis d'en arriver là, et me soutient malgré la distance. Merci de croire (beaucoup trop) en moi. Merci d'être là dans ce moment important, à Paris ou ailleurs, ça veut dire beaucoup pour moi.

La vie est plus douce avec une bouillotte. Merci de faire exister ce nous.

Contents

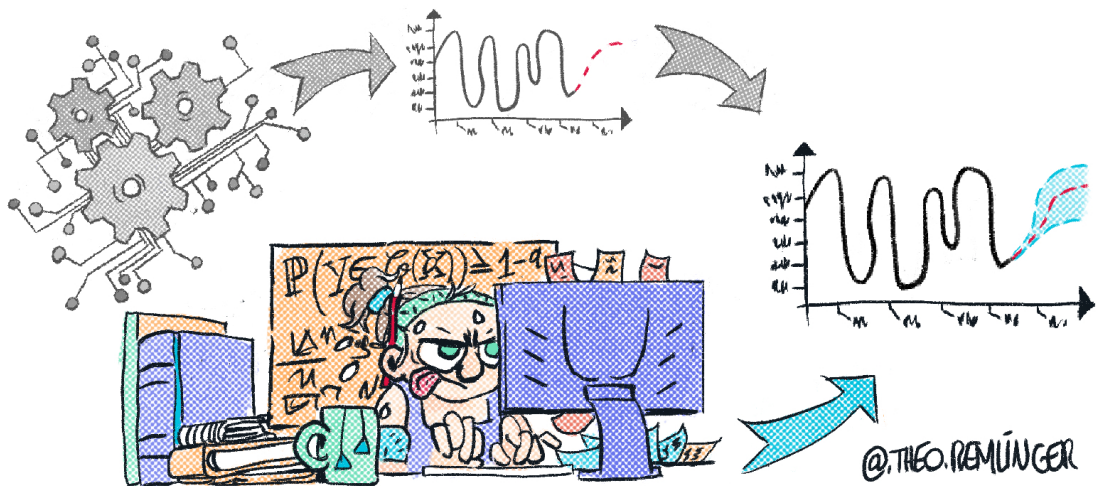
I	Introduction	1
1	Forecasting Electricity Spot Prices	3
1.1	Energy and electricity transition	3
1.2	Electricity markets	4
1.3	Electricity price forecasting	5
1.4	Probabilistic electricity price forecasting	6
2	Thesis Outline and Main Contributions	7
3	Introduction to Conformal Prediction	9
3.1	Supervised learning context and predictive uncertainty	9
3.2	Split Conformal Prediction (SCP)	17
3.3	On the design choices of CP and (empirical) conditional guarantees	33
3.4	Avoiding data splitting: full CP and out-of-bags approaches	41
3.5	Beyond exchangeability	49
4	Technical Summary of the Contributions	54
4.1	Contributions' summary of Part II – Time Series	54
4.2	Contributions' summary of Part III – Missing Values	56
II	Time Series	59
5	Adaptive Conformal Predictions for Time Series	60
5.1	Introduction	62
5.2	Setting: ACI for time series	64
5.3	Impact of γ on ACI efficiency	65
5.4	Adaptive strategies based on ACI	68
5.5	Numerical evaluation on synthetic data sets	69
5.6	Forecasting French electricity spot prices	74
5.7	Conclusion	76
5.A	Details on Split Conformal Prediction	77
5.B	Proof of the results presented in Section 5.3 and additional numerical experiments	81
5.C	Experimental details.	94
5.D	Additional experiments on synthetic data sets	97
5.E	Forecasting French electricity spot prices	100

6	Adaptive Probabilistic Forecasting of French Electricity Spot Prices in 2020 and 2021	104
6.1	Introduction	106
6.2	Data presentation and insightful new explanatory variables	108
6.3	Probabilistic forecasting methods	110
6.4	Adaptiveness as a wrapper around individual forecasts	114
6.5	Application and results	119
6.6	Conclusion and perspectives	122
<hr/>		
6.A	Results on the CRPS	124
III	Missing Values	126
7	Conformal Prediction with Missing Values	127
7.1	Introduction	129
7.2	Background	131
7.3	Warm-up: marginal coverage with NAs	132
7.4	Challenge: NAs induce heteroskedasticity	133
7.5	Achieving mask-conditional-validity (MCV)	134
7.6	Towards asymptotic individualized coverage	138
7.7	Empirical study	139
7.8	Conclusion and perspectives	143
<hr/>		
7.A	Detailed perspective discussion	144
7.B	Illustration and details on CQR (Romano et al., 2019) procedure	145
7.C	Impute-then-predict+conformalization	147
7.D	Gaussian linear model	148
7.E	Finite sample algorithms	151
7.F	Infinite data results	156
7.G	Experimental study	159
8	Predictive Uncertainty Quantification with Missing Covariates	165
8.1	Introduction	167
8.2	When is Mask-Conditional-Validity (MCV) a too lofty goal?	173
8.3	Links between missing covariates and predictive uncertainty	177
8.4	Principled unified Missing Data Augmentation (MDA) framework: CP-MDA-Nested*	184
8.5	A practical glimpse on the impacts of breaking the distribution’s assumptions	191
<hr/>		
8.A	Hardness results	199
8.B	Link between missing covariates and uncertainty	204
8.C	Leave-one-out predictive sets for randomized algorithms	208
8.D	Theory on CP-MDA-Nested* and CP-MDA-Nested	210

IV Conclusion and perspectives	215
Conclusion	216
Open directions	217
Résumé long en français	220
Bibliography	230

Part I

Introduction



Contents

1	Forecasting Electricity Spot Prices	3
1.1	Energy and electricity transition	3
1.2	Electricity markets	4
1.3	Electricity price forecasting	5
1.4	Probabilistic electricity price forecasting	6
2	Thesis Outline and Main Contributions	7
3	Introduction to Conformal Prediction	9
3.1	Supervised learning context and predictive uncertainty	9
3.1.1	Probabilistic modeling	10
3.1.2	Statistical learning	10
3.1.3	On the importance of predictive uncertainty	12
3.1.4	Quantile Regression	12
3.1.5	Framework of interest, its limits and use cases	15
3.2	Split Conformal Prediction (SCP)	17
3.2.1	Standard mean-regression case and exchangeability	18
3.2.2	Conformalized Quantile Regression (CQR)	24
3.2.3	Generalization of SCP: going beyond regression	27
3.2.4	Some examples of SCP in classification	29
3.3	On the design choices of CP and (empirical) conditional guarantees	33
3.3.1	What choices for the conformity scores?	33
3.3.2	On distribution-free X -conditional validity	35
3.3.3	Y -conditional validity	39
3.3.4	Impact of the calibration set on the coverage	40
3.4	Avoiding data splitting: full CP and out-of-bags approaches	41
3.4.1	Full Conformal Prediction	41
3.4.2	Jackknife+ and leave-one-out CP	44
3.4.3	CV+	47
3.5	Beyond exchangeability	49
3.5.1	Weighting strategies	49
3.5.2	Online setting	51
4	Technical Summary of the Contributions	54
4.1	Contributions' summary of Part II – Time Series	54
4.2	Contributions' summary of Part III – Missing Values	56

Chapter 1

Forecasting Electricity Spot Prices

This PhD thesis has been conducted under a CIFRE contract (industrial agreement for training through research) with EDF (Electricité de France, French main producer and supplier of electricity).

1.1 Energy and electricity transition

“Who could have foreseen the climate crisis?”

There is no need here to remind that according to IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services) in 150 years, 83% of wildlife biomass and 41.5% of plant biomass have disappeared due to human activities; that the IPCC (Intergovernmental Panel on Climate Change) was created more than 35 years ago to ring the alarm; and that despite all of this only insufficient measures have been taken at the political and governmental levels ([HCC-2021](#)). Yet, this question is the tree that hides the actual forest: **what can we actually do to limit the climate crisis, or at least adapt to it?**

Starting from the highest level, a partial natural answer is to reduce anthropogenic greenhouse gas emissions: this is necessary to meet the Paris Agreement requiring to ensure that the earth’s average temperature does not increase by more than 2°C before 2100, compared to 1850. Obviously, reducing our production and consumption would have a great quick impact on this. However, how to achieve it and whether we want to enforce it might be beyond the scope of an academic debate and most likely seems to belong to the citizens’ sphere. Closer to our concrete scope of applications, yet highly relevant, is how we produce energy and everything that lies around it.

The last decades have witnessed important changes in the energy panorama, with an increasing integration of non-fossil fuels based energy production. For instance, major research and operational efforts have been deployed to develop renewable energies ([RTE, 2022](#); [IEA, 2022a](#))¹. Especially, France did commit to reach carbon neutrality by 2050, and in particular by attaining 1/3 of renewable energies in gross final energy consumption by 2030. France also decided to support the development of nuclear plants ([France-2023-491](#)) to attain a decarbonized energy mix. In parallel, many usages have been electrified, or

¹RTE is the French Electricity Transmission Network, while IEA is the International Energy Agency.

are in their way to be, such as electric vehicles and distributed storages. Self-consumption (also known as consumer-producer, i.e. consuming the energy we produce) or even demand response programs (i.e. adapting the demand in concordance with the production, and not the traditional contrary) are also greatly incentivized (Bakare et al., 2023).

The proliferation of these new uses of electricity and the growing importance of intermittent renewable energies are profoundly changing the energy landscape in Europe, and are at the root of major transformations in European electricity markets. In particular, they are becoming more dependent and volatile. **Therefore, an accurate electricity price forecasting is required to stabilize energy production planning and thus reduce the associated carbon emissions by increasing the investments in renewable energies and storage solutions. In this thesis, we focus on short-term prices.**

1.2 Electricity markets

There are 4 main short-term markets in France, and more generally in Europe.

- i) The first one, on which we will focus, is the *spot* market. The spot electricity market is a blind auction market in which producers and suppliers offer bids and offers for each hour, or for a block of hours, of the following day. The market closes at 12am of the day before the delivery. The 24 hourly prices are defined by a “pay-as-clear” principle: all players will exchange MegaWatt-hours at the same price, which, at first glance, can be seen as the cross between global supply and global demand. However, defining the price is more complex, as it takes into account interconnections between different countries, as well as so-called “block” offers.
- ii) The second one is the *intraday* market. It is a continuous trading market, offering hourly, half-hour and quarter-hour products. In contrast to the spot market, the prices are fixed on the fly in order to match the orders as soon as possible, with a closing time 5 to 15min before the delivery.
- iii) Finally, the last two markets are the *system services* and *balancing* markets. These markets are handled by the transport system operator and are the ones responsible to ensure the perfect equilibrium between supply and offer at any time.

These short-term markets are impacted by the transition described in Section 1.1. On the one hand, the need for greater security of electricity supply on different timescales is leading to an overhaul of system services, with the creation of new markets for these services at European level, notably in the new “Electricity balancing” regulatory framework adopted by the European Commission in 2017 (EU-2017/2195). On the other hand, the growing penetration of renewable energies has accentuated uncertainty over a short-term horizon of electricity production, affecting the operation of intraday markets, which are becoming the indispensable tool for managing forecasting errors in renewable production. In the German market, we are already seeing strong correlations between prices and wind generation, and it is only a matter of time before these phenomena appear in France. The presence of storage assets, whose price is steadily falling—even if quite high at the moment—, means that new market strategies can be put in place to stabilize supply and reduce costs.

1.3 Electricity price forecasting

In this fast-changing context, it is essential to have high-performance price forecasting methods for all short-term markets.

Indeed, good price forecasts on successive markets enable us to *better anticipate the financial flows linked to renewable production and optimize the placement of production on the various markets*. It is one of the essential elements for a good valuation of these production assets, which will *encourage investments in these low-carbon assets*.

Moreover, an accurate price forecast, both on successive markets and on the different hourly prices of a same market, enables to *optimize the management of flexibilities* (physical battery or short-term consumption effacement contract, upward and downward adjustment flexibility of thermal power plants, etc.). In particular, raising the value of these flexibilities will encourage players to *invest in these assets, leading to a more secure power system*.

Yet, forecasting electricity prices is highly challenging due to all the aforementioned specificities of electricity: matching demand and production at all times, non-storable nature of electricity, exchanges between different countries via interconnections, the variable nature of generating facilities, etc. Specifically, these characteristics lead to negative or extremely high prices of non-negligible occurrence (see Figure 1.1). This was before recent unfortunate fortuitous events that affected tremendously the markets, making them highly non-stationary, such as Covid-19 pandemic in 2020-2021 (IEA, 2021), the stress corrosion issue which affected French nuclear power plants in 2022 or the crisis of the gas markets triggered by Russia's invasion of Ukraine (IEA, 2022b). Despite the increasing number of available historical data, state-of-the-art models (Weron, 2014; Lago et al., 2021) (from classical times series forecasting to deep learning methods), along with internal studies at EDF R&D², did not obtain forecasts' errors lower than 10% of the realized price³. As a reference, national consumption's forecasts achieve errors around only 1% of the realized consumption.

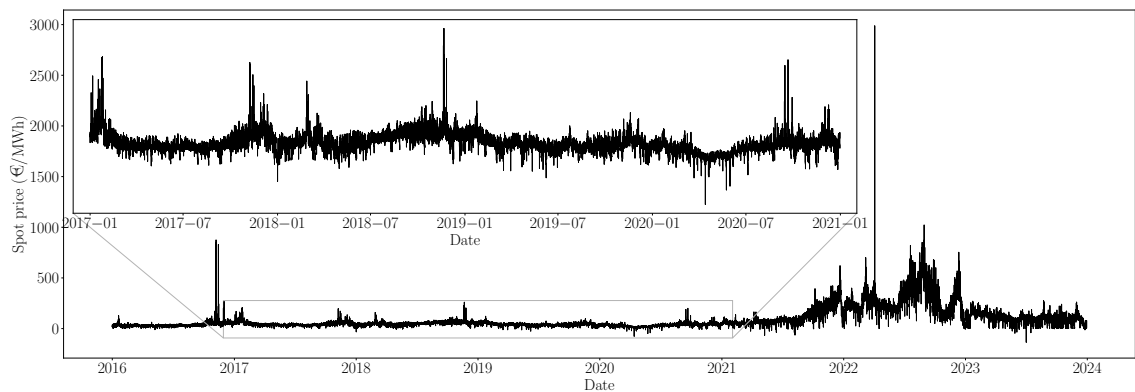


Figure 1.1: Temporal evolution of the French electricity spot prices between 2016 and 2021.

²We note here that operational forecasting tools available at EDF-Trading may be more efficient but they are using real time information that are not available as historical data.

³Surprisingly, this holds for forecasts before 2020 as well as after 2020: the errors are more important after 2020, but as the prices are also higher, the relative error stays at the same order of magnitude.

Leveraging the emergence of open data platforms such as ENTSO-E⁴ Transparency Platform, or Eco2Mix Platform powered by RTE would likely improve electricity price forecasts. However, aggregating different data sources introduces a new demanding setting: the occurrence of **missing values** that comes along with computational and statistical challenges. For instance, it can be caused by different time frequencies or market horizons between fundamentally different explanatory variables. Also, the quality of the data evolves with time (as processes get consolidated) and anomalies can be observed.

1.4 Probabilistic electricity price forecasting

Crucially, these forecasting methods provide ad hoc predictions, with no indication of the degree of confidence to be placed in them. To ensure the trust of key actors in energy markets with regard to such decision-support tools, it is crucial to **quantify their predictive uncertainty**.

Furthermore, trading and energy management decisions (such as the ones mentioned in Section 1.3) require risk management tools which are based on probabilistic electricity price forecasting, leading to a rapid expansion of the literature in this area (see the review of Nowotarski and Weron, 2018). However, traditional probabilistic forecasts are only valid asymptotically or upon strong assumptions on the data that are typically not met by electricity prices (Gaussianity, stationarity).

This supports the advancement of adaptive probabilistic approaches for forecasting prices, which can continuously learn and adjust to the evolving behaviors of electricity prices, resulting in accurate and reliable probabilistic forecasts even on **non-stationary time series**.

In this PhD thesis, we propose to provide **theoretically grounded tools** able to quantify predictive uncertainty under **light assumptions on the underlying data distribution** and whose guarantees are agnostic to the prediction algorithm. We consider **post-hoc** methods, in order to allow their use in a plug-in fashion: any energy markets' actor could keep its preferred operational pipeline and still turn the resulting predictions into guaranteed probabilistic forecasts.

⁴ENTSO-E is the European Network of Transmission System Operators for Electricity.

Chapter 2

Thesis Outline and Main Contributions

This manuscript is divided in three main parts. The rest of this introductory Part I is organized as follows. This chapter 2 provides a quick overview of the outline and main contributions. Chapter 3 is a pedagogical introduction to Conformal Prediction methods (see Table 2.1 for a reading guide), based on a tutorial designed during the completion of this PhD. Finally, in Chapter 4 we give a more technical and detailed summary of our contributions.

Part II studies post-hoc predictive uncertainty quantification for **time series**. The first bottleneck to apply conformal methods in order to obtain guaranteed probabilistic electricity price forecasting in a post-hoc fashion is the highly non-stationary temporal aspect of electricity prices, breaking the exchangeability assumption. In Chapter 5 (based on a joint work with Olivier Féron, Yannig Goude, Julie Josse and Aymeric Dieuleveut) we propose a parameter-free algorithm tailored for time series, which is based on theoretically analysing the efficiency of Adaptive Conformal Inference (Gibbs and Candès, 2021). To investigate deeper how adaptive post-hoc probabilistic electricity prices forecast can be obtained, in Chapter 6 (based on the internship of Grégoire Dutot, co-supervised with Olivier Féron and Yannig Goude) we conduct an extensive application study on novel data set of recent turbulent French spot prices in 2020 and 2021.

Another challenge that predictive uncertainty quantification for electricity prices forecasting faces is the occurrence of **missing values**. Therefore, in Part III (based on joint works with Aymeric Dieuleveut, Julie Josse and Yaniv Romano) we analyse the interplay between missing values and predictive uncertainty quantification. In Chapter 7 we highlight that missing values induce heteroskedasticity, leading to uneven coverage depending on which features are observed. We design two algorithms that recover equalized coverage for any missingness under distributional assumptions on the missingness mechanism. In Chapter 8 we push forwards the theoretical

analysis to understand precisely which distributional assumptions are unavoidable for theoretical informativeness. We also unify the previously proposed algorithms into a general framework that demonstrates empirical robustness to violations of the supposed missingness distribution.

All these contributions are implemented with open source code available on [this GitHub](#). The tutorial on which Chapter 3 is based has also been made openly available on [this website](#).

Each chapter is self-contained, thus the notations may slightly vary from chapter to chapter.

Related contribution		Relevant sections of Chapter 3			
Ch. 3	Tutorial at:	Split CP, Section 3.2	Conditional validity, Section 3.3	Full and K -fold CP, Section 3.4	Non exchangeable, Section 3.5
	► <i>MASPIN days 2023</i> (national), with C. Boyer				
	► <i>ENBIS 2023</i> (European)				
	► <i>UAI 2024</i> (international), with A. Dieuleveut				
	► <i>ICML 2024</i> (international), with A. Dieuleveut				
Ch. 5	M. Zaffran , O. Féron, Y. Goude, J. Josse and A. Dieuleveut <i>ICML 2022</i> ¹	✓			✓
Ch. 6	G. Dutot*, M. Zaffran *, O. Féron and Y. Goude submitted to <i>Applied Energy</i> ²	✓			✓
Ch. 7	M. Zaffran , A. Dieuleveut, J. Josse and Y. Romano <i>ICML 2023</i> ³	✓	(✓)	(✓)	
Ch. 8	M. Zaffran , J. Josse, Y. Romano and A. Dieuleveut submitted to <i>JMLR</i> ⁴	✓	✓	✓	

Table 2.1: Summary of the scientific production (* denotes equal contribution), with indications for a parsimonious reading of Chapter 3.

¹“Adaptive Conformal Predictions for Time Series”.

²“Adaptive Probabilistic Forecasting of French Electricity Spot Prices”.

³“Conformal Prediction with Missing Values”.

⁴“Predictive Uncertainty Quantification with Missing Covariates”.

Chapter 3

Introduction to Conformal Prediction

This chapter is a pedagogical introduction to conformal prediction. Therefore, some proofs are included in the body of the text as they are informative, and might have been modified or detailed with respect to the original papers.

3.1 Supervised learning context and predictive uncertainty

The goal of supervised learning is to predict a *label* $Y \in \mathcal{Y}$ (also known as *response* or *target* or *outcome*), given some *features* $X \in \mathcal{X}$ (also known as *explanatory variables* or *covariates*). We assume that the features and label spaces are measurable and that $\mathcal{X} \subseteq \mathbb{R}^d$, where $d \in \mathbb{N}^*$ is the *problem's dimension*, i.e. the number of features. The nature of \mathcal{Y} defines the type of supervised learning task at hand.

Example 3.1.1 (regression).

In *regression* problems, the label to be predicted is continuous, i.e. $\mathcal{Y} \subseteq \mathbb{R}$.

e.g., electricity prices

Example 3.1.2 (classification).

In *classification* problems, the label to be predicted is categorical, i.e. the label set is finite, typically $\mathcal{Y} \subseteq \mathbb{N}$ or $\mathcal{Y} = \{-1, 1\}$ for the specific case of binary classification.

In other words, predicting $Y \in \mathcal{Y}$ given $X \in \mathcal{X}$ corresponds to looking for a measurable function $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y}) \subseteq \mathcal{Y}^{\mathcal{X}}$ called a *predictor*, such that $f(X)$ “is close to” Y , in a sense that remains to be defined.

Definition 3.1.1 (loss function).

A measurable *loss function*, noted $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_+$, compares two points of \mathcal{Y} , typically by being such that for any $(y, y') \in \mathcal{Y}^2$, $\ell(y, y')$ gets smaller as y and y' gets more similar. Usually, y and y' are the prediction of the studied predictor and the ground truth value.

Example 3.1.3 (quadratic loss—regression).

In *regression*, a standard loss function is $\ell(y, y') = (y - y')^2$.

Example 3.1.4 (0-1 loss—classification).

In *classification*, a natural loss function is $\ell(y, y') = \mathbb{1}\{y \neq y'\}$.

3.1.1 Probabilistic modeling

Modeling the labels Y and the features X as random variables whose joint distribution is denoted \mathcal{D} , the goal of supervised learning is to find a function f^* that minimizes the expectation of the loss ℓ over \mathcal{D} , referred to as the ℓ -risk.

Definition 3.1.2 (ℓ -risk).

The ℓ -risk of a predictor is the expectation of the loss ℓ evaluated on the labels and the predictor outputs under the distribution \mathcal{D} :

$$\mathcal{R}_\ell : \begin{cases} \mathcal{M}(\mathcal{X}, \mathcal{Y}) & \rightarrow \mathbb{R}_+ \\ f & \mapsto \mathbb{E}_{\mathcal{D}}[\ell(Y, f(X))] \end{cases}.$$

Any f^* minimizing the ℓ -risk is a ℓ -Bayes predictor and achieves the ℓ -Bayes risk.

Definition 3.1.3 (ℓ -Bayes predictor).

An ℓ -Bayes predictor is a minimizer of the ℓ -risk:

$$f^* \in \arg \min_{f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})} \mathcal{R}_\ell(f).$$

Moreover, the ℓ -Bayes risk is defined as $\mathcal{R}_\ell^* := \mathcal{R}_\ell(f^*)$ for any ℓ -Bayes predictor f^* .

Example 3.1.5 (quadratic loss Bayes predictor—regression).

In regression, the quadratic-Bayes predictor is $f^*(X) = \mathbb{E}[Y|X]$.

Example 3.1.6 (0-1 loss Bayes predictor—classification).

In classification, the 0-1-Bayes predictor is $f^*(X) \in \arg \max_{k \in \{-1, 1\}} \mathbb{P}(Y = k|X)$.

3.1.2 Statistical learning

In practice, the distribution \mathcal{D} is unknown. Computing explicitly the ℓ -risk and a fortiori exhibiting the ℓ -Bayes predictor is therefore impossible. However, we typically have access to $n \in \mathbb{N}^*$ independent and identically distributed (i.i.d.) random variables drawn from \mathcal{D} , forming a data set $(X_i, Y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n \sim \mathcal{D}^{\otimes(n)}$. One can leverage this data set in order to learn a predictor based on the *historical/training data*.

Definition 3.1.4 (statistical learning algorithm).

A *statistical learning algorithm* is a measurable function

$$\mathcal{A} : \begin{cases} \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n & \rightarrow \mathcal{M}(\mathcal{X}, \mathcal{Y}) \\ (X_i, Y_i)_{i=1}^n & \mapsto \hat{f}_n. \end{cases}$$

More generally, a *stochastic* statistical learning algorithm is a measurable function

$$\mathcal{A} : \begin{cases} \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \times [0, 1] & \rightarrow \mathcal{M}(\mathcal{X}, \mathcal{Y}) \\ (X_i, Y_i)_{i=1}^n \times \xi & \mapsto \hat{f}_n, \end{cases}$$

where ξ encodes the randomness of \mathcal{A} .

One goal of such a *statistical learning algorithm* could be to attain a risk close to the ℓ -Bayes risk \mathcal{R}_{ℓ}^* . However, again, without information on \mathcal{D} the true ℓ -risk of a predictor can not be computed. Nonetheless, we can use the *training data* as a surrogate for \mathcal{D} to estimate the ℓ -risk by computing the so-called empirical ℓ -risk.

Definition 3.1.5 (empirical ℓ -risk).

The *empirical ℓ -risk* of a predictor is the empirical average of its loss on the training data set:

$$\hat{R}_{n,\ell} : \begin{cases} \mathcal{M}(\mathcal{X}, \mathcal{Y}) & \rightarrow \mathbb{R}_+ \\ f & \mapsto \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)). \end{cases}$$

Remark 3.1.1 (consistency of the empirical ℓ -risk).

The empirical ℓ -risk is a consistent estimator of the ℓ -risk.

Many statistical learning algorithms are built so as to minimize the *empirical risk*. By doing so, they aim at using *historical/training data* to infer a predictor that should provide accurate prediction on any $X \in \mathcal{X}$, even non-observed ones. To ensure this *generalization*, the predictor has to be constrained to a fixed family of functions $\mathcal{F} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$, called a *model*.

Definition 3.1.6 (empirical risk minimizer).

A *minimizer of the empirical risk* over $\mathcal{F} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$ is a statistical learning algorithm \mathcal{A} such that:

$$\mathcal{A} : \begin{cases} \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n & \rightarrow \mathcal{F} \\ (X_i, Y_i)_{i=1}^n & \mapsto \arg \min_{f \in \mathcal{F}} \hat{R}_{n,\ell}(f). \end{cases}$$

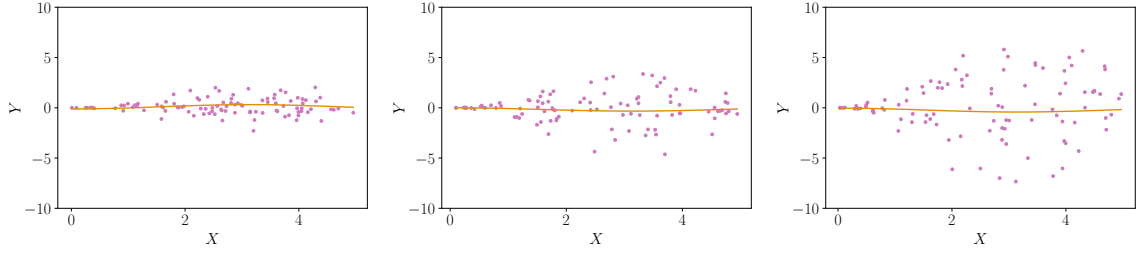


Figure 3.1: Three distinct data distributions with the same quadratic-Bayes predictor (regression setting).

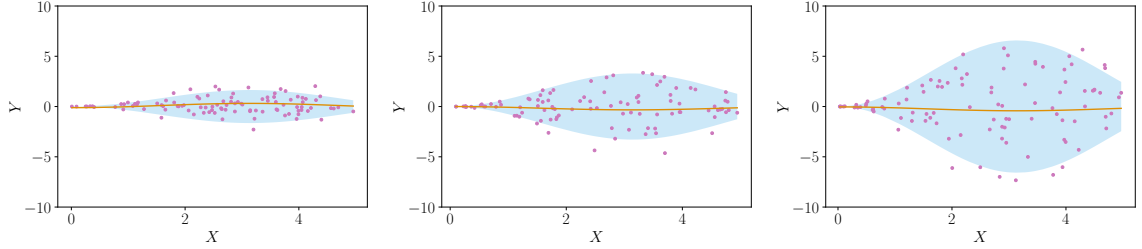


Figure 3.2: Three distinct data distributions with the same conditional expectation, and different conditional quantiles in blue (regression setting).

3.1.3 On the importance of predictive uncertainty

In the previous sections, we have explored the paradigm where one aims to predict a *single* value, also referred to as *point prediction*, without any indication of the degree of confidence that can be given to these predictions. By leveraging increasingly large data sets, statistical algorithms and machine learning methods are now frequently used to support high-stakes decision-making problems such as autonomous driving, medical or civic applications, among others. Yet, as it can be observed in Figure 3.1, an important drawback of this approach is that even the Bayes predictor does not allow to characterize the underlying distribution of $Y|X$. Therefore, the same perfect predictions cover up different underlying phenomena. Quantifying uncertainty (e.g., as illustrated in Figure 3.2 through perfect predictive intervals based on the conditional quantiles of $Y|X$) conveys the information of the predictive uncertainty. To ensure the safe deployment of predictive models¹ it is crucial to quantify the inherent uncertainty of the resulting predictions, communicating the limits of predictive performance.

3.1.4 Quantile Regression

In this subsection, we focus solely on the regression setting.

An approach to take the predictive uncertainty into account is to consider the *quantiles* of a random variable Y , as they encapsulate the overall distribution of Y . First, let's consider the *univariate* or *marginal quantiles*, which do not take into account any link between Y and some features X (i.e. they take the expectation over X).

¹By a slight abuse of language, we commonly use “model” instead of “statistical learning algorithm”.

Definition 3.1.7 (univariate quantile).

The *quantile* of level $\beta \in [0, 1]$ of Y , denoted $Q_Y(\beta)$, is defined as:

$$\begin{aligned} Q_Y(\beta) &:= \inf\{y \in \mathbb{R}, \mathbb{P}(Y \leq y) \geq \beta\} \\ &:= \inf\{y \in \mathbb{R}, F_Y(y) \geq \beta\}. \end{aligned}$$

$Q_Y(\cdot)$ is the *quantile function*, which is the generalized inverse of the cumulative distribution function F_Y .

It can be estimated through the *empirical quantile* of level β :

$$q_\beta(Y_1, \dots, Y_n) := \lceil \beta \times n \rceil \text{ smallest value of } (Y_1, \dots, Y_n).^a$$

^aSimilarly, let $q_{\beta, \inf}(Y_1, \dots, Y_n) := \lfloor \beta \times n \rfloor \text{ smallest value of } (Y_1, \dots, Y_n)$.

Example 3.1.7 (median).

The quantile of level $\beta = 0.5$ is better known as the *median*.

$\hookrightarrow q_{0.5}(Y_1, \dots, Y_n)$ is the $\lceil 0.5 \times n \rceil$ smallest value of (Y_1, \dots, Y_n) , i.e. the smallest value of (Y_1, \dots, Y_n) which is larger than at least half of (Y_1, \dots, Y_n) , known as the *empirical median* of (Y_1, \dots, Y_n) ;

$\hookrightarrow Q_Y(0.5)$ is the *median* of the distribution of Y .

Just like the expectation is the natural minimizer of the quadratic loss, the quantiles minimize the *pinball loss* described below, and widely used to estimate quantiles in practice. This is formalized in Remark 3.1.2.

Definition 3.1.8 (pinball loss).

The *pinball loss* of level $\beta \in [0, 1]$ is defined as:

$$\ell_\beta : \begin{cases} \mathcal{Y} \times \mathcal{Y} & \rightarrow \mathbb{R}_+ \\ (y, y') & \mapsto \beta |y - y'| \mathbb{1}\{y - y' \geq 0\} + (1 - \beta) |y - y'| \mathbb{1}\{y - y' \leq 0\}. \end{cases}$$

Remark 3.1.2 (minimizing the pinball loss retrieves the quantile).

Let $\beta \in [0, 1]$. Assume $\arg \min_{q \in \mathcal{Y}} \mathbb{E}_{\mathcal{D}_Y} [\ell_\beta(Y, q)] \neq \emptyset$.

Set $q_\beta^* \in \arg \min_{q \in \mathcal{Y}} \mathbb{E}_{\mathcal{D}_Y} [\ell_\beta(Y, q)]$.

Then if F_Y is continuous and strictly increasing, we have that $q_\beta^* = F_Y^{-1}(\beta)$.

Proof. First, as $q_\beta^* \in \arg \min_{q \in \mathcal{Y}} \mathbb{E}_{\mathcal{D}_Y} [\ell_\beta(Y, q)]$, we have:

$$\begin{aligned} 0 &= \left(\frac{d}{dq} \mathbb{E}_{\mathcal{D}_Y} [\ell_\beta(Y, q)] \right) (q_\beta^*) \\ \Leftrightarrow 0 &= \left(\frac{d}{dq} \int_{-\infty}^{+\infty} \ell_\beta(y, q) dF_Y(y) \right) (q_\beta^*). \end{aligned}$$

Then, let $q' \in \mathbb{R}$. Remark that for any $y \neq q'$, $\frac{\partial \ell_\beta}{\partial q}(y, q')$ does exist. Furthermore, for any $y \neq q'$, we have that:

$$\left| \frac{\partial \ell_\beta}{\partial q}(y, q') \right| = |(\beta - 1)\mathbb{1}\{y < q'\} + \beta\mathbb{1}\{y > q'\}| \leq 1.$$

As F_Y is continuous, $Y \neq q_\beta^*$ almost surely.

Therefore, by differentiation under the integral sign, we obtain:

$$\begin{aligned} 0 &= \int_{-\infty}^{+\infty} \frac{\partial \ell_\beta}{\partial q}(y, q_\beta^*) dF_Y(y) \\ 0 &= (\beta - 1) \int_{-\infty}^{q_\beta^*} dF_Y(y) + \beta \int_{q_\beta^*}^{+\infty} dF_Y(y) \\ 0 &= (\beta - 1)F_Y(q_\beta^*) + \beta(1 - F_Y(q_\beta^*)) \\ (1 - \beta)F_Y(q_\beta^*) &= \beta(1 - F_Y(q_\beta^*)) \\ \beta &= F_Y(q_\beta^*). \end{aligned}$$

Finally, as F_Y is also strictly increasing, we get:

$$q_\beta^* = F_Y^{-1}(\beta).$$

□

Building on the marginal quantiles Q_Y , an interesting notion is the conditional quantiles $Q_{Y|X}$: it leverages the information of the features X to describe the distribution of Y . Formally, the conditional quantiles portray the conditional distribution of $Y|X$. This is essential when the underlying distribution is heteroskedastic and the predictive uncertainty varies depending on X , such as in Figures 3.1 and 3.2.

Definition 3.1.9 (conditional quantile).

The *conditional quantile* of level $\beta \in [0, 1]$ of $Y|X$, denoted $Q_{Y|X}(\beta)$, is defined as:

$$\begin{aligned} Q_{Y|X}(\beta) &:= \inf\{y \in \mathbb{R}, \mathbb{P}(Y \leq y|X) \geq \beta\} \\ &:= \inf\{y \in \mathbb{R}, F_{Y|X}(y) \geq \beta\}. \end{aligned}$$

Armed with this definition of conditional quantiles, one can perform *quantile regression* by considering the pinball loss—in place of mean regression based on the quadratic loss—in order to learn the predictive uncertainty of $Y|X$.

Definition 3.1.10 (quantile regression).

Quantile regression for the level $\beta \in [0, 1]$ aims at minimizing the associated pinball risk, that is solving:

$$f_\beta^* \in \arg \min_{f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})} \mathcal{R}_{\ell_\beta}(f) := \arg \min_{f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})} \mathbb{E}_{\mathcal{D}} [\ell_\beta(Y, f(X))].$$

Such a f_β^* satisfies $\mathbb{P}(Y \leq f_\beta^*(X)|X) = \beta$ if $F_{Y|X}$ is continuous.

Example 3.1.8 (median regression).

Minimizing the risk associated to the *absolute error* $\ell(y, y') := |y - y'| = \ell_{0.5}(y, y')$ corresponds to *median regression*:

$$\text{median}[Y|X] = Q_{Y|X}(0.5) \in \arg \min_{f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})} \mathbb{E}_{\mathcal{D}} [|Y - f(X)| | X].$$

An illustration of quantile regression for various levels β along with the associated pinball losses is provided in Figure 3.3.

Remark 3.1.3 (no theoretical guarantees in general).

One may consider building a predictive intervals (such as the ones illustrated in Figure 3.2) through the conditional quantiles of $Y|X$. Indeed, using the exact quantiles, we have for any $\beta \in [0, 1]$:

$$\mathbb{P} \left(Y \in \left[Q_{Y|X} \left(\frac{\beta}{2} \right); Q_{Y|X} \left(1 - \frac{\beta}{2} \right) \right] \right) = 1 - \beta.$$

However, as discussed in Section 3.1.2, in practice we do not have access to $Q_{Y|X}(\cdot)$ and we have to estimate it to obtain a $\hat{Q}_{Y|X}(\cdot)$ e.g., by minimizing the empirical risk. Then, with a finite number of observations n , in general:

$$\mathbb{P} \left(Y \in \left[\hat{Q}_{Y|X} \left(\frac{\beta}{2} \right); \hat{Q}_{Y|X} \left(1 - \frac{\beta}{2} \right) \right] \right) \neq 1 - \beta.$$

Consequently, without further assumptions such as consistency and infinite data or distributional assumptions, quantile regression is not sufficient for providing guaranteed predictive uncertainty quantification.

3.1.5 Framework of interest, its limits and use cases

Our goal is to predict $Y \in \mathcal{Y}$ given its covariates $X \in \mathcal{X}$ with a notion of **confidence**, i.e. with a quantification of the predictive uncertainty. Formally, given a *miscoverage rate*

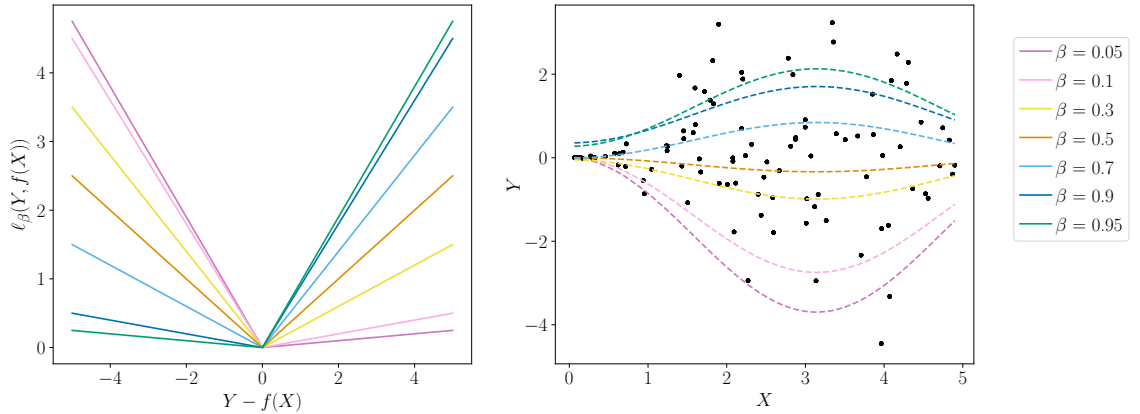


Figure 3.3: Illustration of quantile regression for various quantile levels β represented by the colors. Left: pinball losses. Right: estimated quantile regressions.

$\alpha \in [0, 1]$, typically small, we aim at building a *predictive set* \mathcal{C}_α such that:

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(X)) \geq 1 - \alpha,$$

where \mathcal{C}_α should be as small as possible in order to be informative. Indeed, the predictive set given in Example 3.1.9 which outputs \mathcal{Y} with probability $1 - \alpha$, and the empty set otherwise, is *exactly* valid, yet it is critically uninformative.

Example 3.1.9 (uninformative yet always valid predictive set).

$$\mathcal{C}_\alpha(\cdot; \xi) \equiv \mathcal{Y} \mathbb{1}\{\xi \leq 1 - \alpha\} + \emptyset \mathbb{1}\{\xi > \alpha\},$$

where $\xi \sim \mathcal{U}([0, 1])$.

We remind that in practice we only access a data set $(X_i, Y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, and aim at predicting on an unseen individual X_{n+1} . Therefore, we build an estimator $\hat{\mathcal{C}}_{n,\alpha}$ of the predictive sets using a statistical learning algorithm on the training data set, in the objective that it satisfies Equation (MV) (we then say that $\hat{\mathcal{C}}_{n,\alpha}$ is *marginally valid*) while being as small as possible (we then say that $\hat{\mathcal{C}}_{n,\alpha}$ is *efficient*).

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{C}}_{n,\alpha}(X_{n+1})) \geq 1 - \alpha \quad (\text{MV})$$

In this thesis, we study estimators satisfying Equation (MV), to quantify predictive uncertainty in the statistical learning setting. Yet, several constraints typically arise:

- i) The learner generally has access only to a finite number of data points;
- ii) Data set from the real world derives from unknown distributions. If large deviations are sometimes easy to check, smaller ones can still lead to important statistical failure;
- iii) The multiplicity and heterogeneity of used models as well as the complexity to finely analyse some of them ask for generic methods that do not assume any specific learning algorithm and can be plugged-in on top of any existing pipeline.

To answer these concerns, we focus on methods satisfying Equation (MV) on *i*) finite sample data sets, in opposition to asymptotic guarantees, *ii*) without relying on distributional assumptions with respect to \mathcal{D} , and *iii*) which can be used with any learning algorithm.

On the importance of the post-hoc design

Let us pause here to underline the importance of the last point *iii*). We see the estimation of $\hat{\mathcal{C}}_{n,\alpha}$ as an add-on to an existing learning pipeline \mathcal{A} , which turns the (point) predictions of \mathcal{A} into predictive sets with guaranteed coverage, irrespectively of the quality of \mathcal{A} on the considered data set. In other words, $\hat{\mathcal{C}}_{n,\alpha}$ **can be plugged-in in a post-hoc fashion on top of any \mathcal{A} , with no impact of the choice of \mathcal{A} on the validity of Equation (MV)**. Of course, even if the choice of \mathcal{A} does not affect Equation (MV), it will nonetheless impact the shape of the predictive sets: the lower the performances of \mathcal{A} , the larger the predictive sets will be. This is in fact a good property of our framework as the final user can analyze the quality of the predictive sets to understand how reliable \mathcal{A} is on the current task at hand.

Conformal prediction (CP, [Vovk et al., 2005](#)) is a versatile framework achieving Equation (MV) in finite sample with no assumption on the distribution \mathcal{D} , and in a post-hoc fashion. Therefore, we focus in this PhD thesis on CP approaches, and the subsequent sections of this introductory chapter are devoted to provide a detailed overview of CP. Before diving into this introduction, let us first pause to highlight exactly the statement of Equation (MV) and how it should be understood.

Remark 3.1.4 (no free lunch).

► **What type of predictive uncertainty quantification would we like to have?**

Given historical data $(X_i, Y_i)_{i=1}^n$ and new features X_{n+1} , we would like to find $\hat{C}_{n,\alpha}(X_{n+1})$ such that:

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \mid X_{n+1}, (X_i, Y_i)_{i=1}^n\right) \geq 1 - \alpha, \quad (\text{UQ dream})$$

which means that the coverage does not vary with *i*) the training sample $(X_i, Y_i)_{i=1}^n$ (i.e. no under/over-covering depending on the training set draw) nor *ii*) the covariates X_{n+1} (e.g., whether we predict on week end or week day).

► **What can we have?**

In Equation (MV), the probability is taken not only on the new label Y_{n+1} , but also on the new features X_{n+1} as well as on the training set $(X_i, Y_i)_{i=1}^n$ (through $\hat{C}_{n,\alpha}$). In fact, as developed in Section 3.3.2, the previous wish Equation (UQ dream) is impossible to achieve under our set of assumptions. On the one hand, we will see that it achieving conditional validity on the covariates X_{n+1} in an informative distribution-free fashion is impossible. On the other hand, we will also see that some CP approaches still manage to ensure some form of conditional validity on the training set.

Given Remark 3.1.4, CP predictive sets do not have to be understood as a “magic wand” to probabilistic prediction and predictive uncertainty quantification. We believe that CP should be used as a last protective layer to be plugged-in after the best learning pipeline that can be designed, tailored for the application at hand. The strength of CP is precisely that it can be used in combination with any learning pipeline and still provide a valid marginal guarantee, leading to robust prediction if the underlying pipeline is corrupted, and achieving stronger guarantees than expected otherwise.

However, developing extensions of CP that refines the guarantee is of great interest. It constitutes a branch of the literature, that we will discuss hereafter.

3.2 Split Conformal Prediction (SCP)

We start this introductive overview of CP by presenting Split CP (SCP, [Vovk et al., 2005](#); [Papadopoulos et al., 2002](#); [Lei et al., 2018](#)). Historically, SCP was introduced after Full CP, and is in fact a particular case of it. However, we find it more pedagogical to start with SCP.

3.2.1 Standard mean-regression case and exchangeability

Let us begin by explaining SCP in the very simple case where the base learning algorithm \mathcal{A} performs mean-regression and outputs a function $\hat{\mu}$ based on some training data.

SCP first splits the n points of the training set into two disjoint sets $\text{Tr}, \text{Cal} \subset \llbracket 1, n \rrbracket$, to create a *proper training set*, Tr of size $\#\text{Tr}$, and a *calibration set*, Cal of size $\#\text{Cal}$. On the proper training set, a mean-regression model $\hat{\mu}$ (chosen by the user) is fitted, and then used to predict on the calibration set. *Conformity scores* $s(x, y; \hat{\mu}) := |y - \hat{\mu}(x)|$ are computed to assess how well the fitted model $\hat{\mu}$ predicts the response values of the calibration points, forming the set $\mathcal{S} = \{(S_i := s(X_i, Y_i; \hat{\mu}))_{i \in \text{Cal}}\} \cup \{+\infty\}$. Finally, the $(1 - \alpha)$ -th quantile of these scores $q_{1-\alpha}(\mathcal{S})$ is computed to define the size of the predictive interval: $\hat{C}_{n,\alpha}(\cdot) := [\hat{\mu}(\cdot) \pm q_{1-\alpha}(\mathcal{S})]$.

Remark 3.2.1 ($\hat{\mu}$ can be independent of Tr).

When we say that “ $\hat{\mu}$ is fitted on the proper training set”, we include the extreme case where $\hat{\mu}$ is in fact independent of Tr , e.g., when obtaining a model from a third party. The important point is that $\hat{\mu}$ has to be independent of the calibration set.

Remark 3.2.2 (on the $+\infty$ in \mathcal{S}).

When forming the set of scores \mathcal{S} , we have cautiously added $+\infty$. This is crucial to ensure finite sample guarantees: ideally we would like to use the true quantile of the scores’ distribution but once again, this quantity is unknown, and to estimate it we apply a finite-sample correction. See Lemma 3.2.1 for a formal derivation. One can think of it as including a worst-case scenario for the unknown value of $s(X_{n+1}, Y_{n+1}; \hat{\mu})$. In fact, this is strictly equivalent to taking the $\left((1 - \alpha) \left(1 + \frac{1}{\#\text{Cal}}\right)\right)$ empirical quantile of $\{(s(X_i, Y_i; \hat{\mu}))_{i \in \text{Cal}}\}$.

An illustration is provided in Figure 3.4 in the case where $d = 1$, i.e. when there is only one explanatory variable. We present in Algorithm 1 the pseudo-code of SCP in the particular case explained above.

Let us now state formally the theoretical guarantees enjoyed by Algorithm 1. As the calibration set is used to estimate the quantiles of the “errors” made by $\hat{\mu}$ and infer their order of magnitude at test time, we intuit that if the calibration and test points are i.i.d. then we could show that this method achieves Equation (MV). In fact, we only need a weaker notion than i.i.d. which allows for some dependence structure: *exchangeability*.

Definition 3.2.1 (exchangeability).

$(X_i, Y_i)_{i=1}^n$ are *exchangeable* if, for any permutation σ of $\llbracket 1, n \rrbracket$:

$$((X_1, Y_1), \dots, (X_n, Y_n)) \stackrel{d}{=} ((X_{\sigma(1)}, Y_{\sigma(1)}), \dots, (X_{\sigma(n)}, Y_{\sigma(n)})).$$

Toy case: Z_1 and Z_2 are exchangeable if $(Z_1, Z_2) \stackrel{d}{=} (Z_2, Z_1)$.

Exchangeability implies that the $(X_i, Y_i)_{i=1}^n$ are identically distributed. Denoting \mathcal{D} their

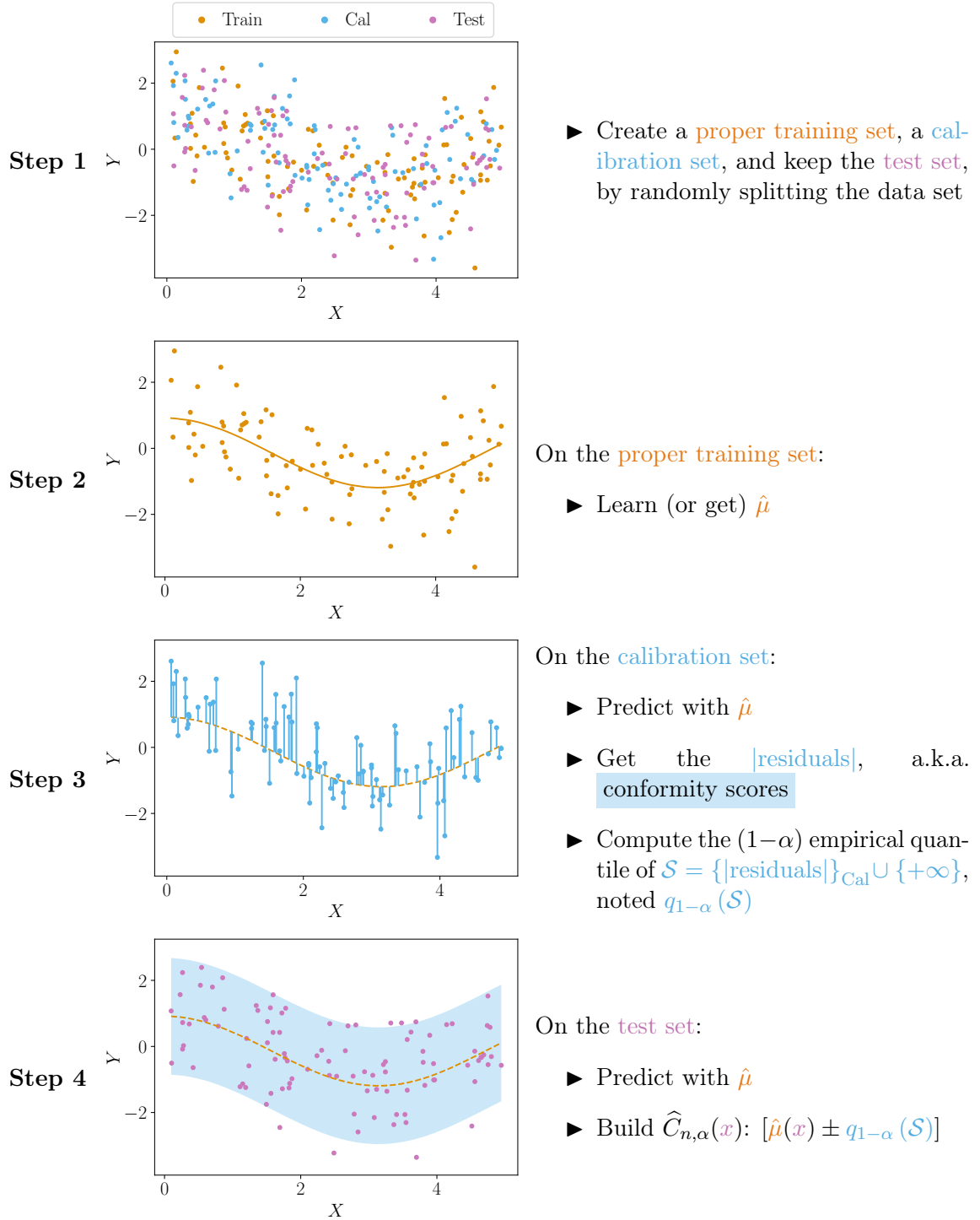


Figure 3.4: Schematic illustration of the Split Conformal Prediction procedure. Special case of a mean-regression task, with the absolute value of the residuals as conformity scores.

Algorithm 1 SCP in mean-regression using the absolute value of the residuals as conformity scores

Input: Mean-regression algorithm \mathcal{A} , miscoverage rate α , training set $(X_i, Y_i)_{i=1}^n$

Output: Prediction interval $\hat{C}_{n,\alpha}$



- 1: Randomly split the training data $(X_i, Y_i)_{i=1}^n$ into a **proper training set** (size $\#Tr$) and a **calibration set** (size $\#Cal$)
- 2: Get $\hat{\mu}$ (by training \mathcal{A} on the **proper training set** $(X_i, Y_i)_{i \in Tr}$)
- 3: On the **calibration set**, get prediction values with $\hat{\mu}$
- 4: Obtain a set of $\#Cal + 1$ **conformity scores**:

$$\mathcal{S} = \{S_i = s(X_i, Y_i; \hat{\mu}), i \in \text{Cal}\} \cup \{+\infty\}, \text{ with } s(x, y; \hat{\mu}) := |y - \hat{\mu}(x)|$$

Obtain a set of $\#Cal$ conformity scores: $\mathcal{S} = \{S_i, i \in \text{Cal}\}$

- 5: Compute the $1 - \alpha$ quantile of these scores: $q_{1-\alpha}(\mathcal{S})$

Compute the $\left((1 - \alpha) \left(1 + \frac{1}{\#Cal}\right)\right)$ quantile of these scores: $q_{1-\alpha}(\mathcal{S})$

- 6: For a new point X_{n+1} , return

$$\hat{C}_{n,\alpha}(X_{n+1}) = [\hat{\mu}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}); \hat{\mu}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$$

marginal distribution as earlier, we note $\mathcal{D}^{\text{exch}(n)}$ the set of exchangeable joint distributions of marginal \mathcal{D} .

Example 3.2.1 (i.i.d.).

An i.i.d. sequence is exchangeable.

Example 3.2.2 (sampling without replacement).

A sequence (U_1, \dots, U_n) obtained through sampling without replacement from $\{u_1, \dots, u_n\}$ is exchangeable (but not i.i.d.).

Example 3.2.3 (multivariate gaussian).

The components of $\mathcal{N}\left(\begin{pmatrix} m \\ \vdots \\ m \end{pmatrix}, \begin{pmatrix} \sigma^2 & & \\ & \ddots & \gamma^2 \\ & \gamma^2 & \ddots \\ & & & \sigma^2 \end{pmatrix}\right)$ with $m \in \mathbb{R}$, $(\sigma, \gamma) \in \mathbb{R}_+^2$, are exchangeable even when $\gamma \neq 0$ (thus even when they are not independent).

Equipped with the notion of exchangeability, we can now show that SCP for mean-regression with absolute value of the residuals as conformity score (Algorithm 1) achieves Equation (MV) for any sample size, whatever the learning algorithm \mathcal{A} is and for any distribution \mathcal{D} as long as $(X_i, Y_i)_{i=1}^{n+1} \sim \mathcal{D}^{\text{exch}(n+1)} \in \mathcal{D}^{\text{exch}(n+1)}$ (Vovk et al., 2005; Papadopoulos

et al., 2002; Lei et al., 2018).

Theorem 3.2.1 (marginal validity of SCP—mean-regression, absolute residuals).

SCP for mean-regression with absolute value of the residuals as conformity score (Algorithm 1) outputs $\hat{C}_{n,\alpha}$ such that for any distribution \mathcal{D} , for any associated exchangeable joint distribution $\mathcal{D}^{\mathcal{E}(\text{Cal} \cup \{n+1\})} \in \mathcal{D}^{\text{exch}(\text{Cal} \cup \{n+1\})}$:

$$\mathbb{P}_{\mathcal{D}^{\mathcal{E}(\text{Cal} \cup \{n+1\})}} \left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \right) \geq 1 - \alpha.$$

Additionally, if the scores $\{S_i\}_{i \in \text{Cal}} \cup \{S_{n+1}\}$ are almost surely (a.s.) distinct:

$$\mathbb{P}_{\mathcal{D}^{\mathcal{E}(\text{Cal} \cup \{n+1\})}} \left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \right) \leq 1 - \alpha + \frac{1}{\#\text{Cal} + 1}.$$

We defer the remarks of Theorem 3.2.1 after its proofs, which relies on the following quantile lemma 3.2.1 (see also Tibshirani et al., 2019).

Lemma 3.2.1 (quantile lemma).

If $(U_1, \dots, U_n, U_{n+1})$ are exchangeable, then for any $\beta \in]0, 1[$:

$$\mathbb{P}(U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty)) \geq \beta.$$

Additionally, if U_1, \dots, U_n, U_{n+1} are almost surely distinct, then:

$$\mathbb{P}(U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty)) \leq \beta + \frac{1}{n+1}.$$

Proof. Let $\beta \in]0, 1[$.

First, observe that $\{U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty)\} \iff \{U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})\}$.

By exchangeability, and using Lemma 3.2.2 with the function

$$g : \begin{cases} \bigcup_{n \geq 0} & \rightarrow \{0, 1\} \\ \mathcal{Z} = (Z_i)_{i=1}^{n+1} & \mapsto \mathbb{1}\{\mathcal{Z}_{n+1} \leq q_\beta(\mathcal{Z})\}, \end{cases}$$

we obtain that for any $i \in \llbracket 1, n+1 \rrbracket$: $\{U_{n+1} \leq q_\beta(U_1, \dots, U_{n+1})\} \stackrel{d}{=} \{U_i \leq q_\beta(U_1, \dots, U_{n+1})\}$.

Therefore, for any $i \in \llbracket 1, n+1 \rrbracket$, it holds that $\mathbb{P}(U_{n+1} \leq q_\beta(U_1, \dots, U_{n+1})) \stackrel{d}{=} \mathbb{P}(U_i \leq q_\beta(U_1, \dots, U_{n+1}))$. Thus:

$$\begin{aligned} \mathbb{P}(U_{n+1} \leq q_\beta(U_1, \dots, U_{n+1})) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}(U_i \leq q_\beta(U_1, \dots, U_{n+1})) \\ &= \frac{1}{n+1} \mathbb{E} \left[\sum_{i=1}^{n+1} \mathbb{1}\{U_i \leq q_\beta(U_1, \dots, U_{n+1})\} \right] \\ &\geq \frac{1}{n+1} \mathbb{E}[\lceil \beta(n+1) \rceil] \\ &= \frac{\lceil \beta(n+1) \rceil}{n+1} \\ &\geq \beta, \end{aligned}$$

proving the first statement.

For the second statement, remark that by definition of q_β :

$$\{U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})\} \iff \{\text{rank}(U_{n+1}) \leq \lceil \beta(n+1) \rceil\}.$$

By exchangeability and the fact that there are no ties (U_1, \dots, U_n, U_{n+1} are a.s. distinct), $\text{rank}(U_{n+1}) \sim \mathcal{U}(\{1, \dots, n+1\})$. Thus:

$$\begin{aligned} \mathbb{P}(\text{rank}(U_{n+1}) \leq \lceil \beta(n+1) \rceil) &= \frac{\lceil \beta(n+1) \rceil}{n+1} \\ &\leq \frac{1 + \beta(n+1)}{n+1} = \beta + \frac{1}{n+1}. \end{aligned}$$

□

Proof of Theorem 3.2.1. When $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, the scores $\{S_i\}_{i \in \text{Cal}} \cup \{S_{n+1}\}$ are exchangeable, due to Lemma 3.2.2 with the function $g((X_i, Y_i)_{i=1}^{n+1}) := (|Y_i - \hat{\mu}(X_i)|)_{i=1}^{n+1}$. Therefore, applying the quantile lemma to the scores concludes the proof, as:

$$\begin{aligned} \{Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1})\} &= \{\hat{\mu}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}) \leq Y_{n+1} \leq \hat{\mu}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})\} \\ &= \{|Y_{n+1} - \hat{\mu}(X_{n+1})| \leq q_{1-\alpha}(\mathcal{S})\} \\ &= \{S_{n+1} \leq q_{1-\alpha}(\mathcal{S})\}. \end{aligned} \tag{3.1}$$

□

Lemma 3.2.2 (function of exchangeable sequences).

Let $(U_1, \dots, U_n, U_{n+1})$ be exchangeable. Let σ be a permutation on $\llbracket 1, n+1 \rrbracket$. For any random function g such that $g(\cdot) = h(\cdot; \xi)$ with h a deterministic function, and ξ encoding the randomness of g and independent of $(U_1, \dots, U_n, U_{n+1})$, it holds:

$$g(U_1, \dots, U_n, U_{n+1}) \stackrel{d}{=} g(U_{\sigma(1)}, \dots, U_{\sigma(n)}, U_{\sigma(n+1)}).$$

This includes the particular case where g is a deterministic function.

The strength of Theorem 3.2.1 is that the coverage holds for any finite sample size, for any data distribution \mathcal{D} as long as the data set is exchangeable, and whatever the quality of the fitted model $\hat{\mu}$ is. Again, if $\hat{\mu}$ is a bad predictor (e.g., predicting constantly 10 when the data is distributed as in Figure 3.4) then the length of the predictive interval is critically large. But precisely, this can be used as a diagnostic tool indicating that the modelisation is not tailored for the underlying problem.

Remark 3.2.3 (the upper bound is not sufficient for efficiency).

Talking about efficiency, the upper bound in Theorem 3.2.1 decreases with the calibration size. This is a positive result as an efficient predictive interval will achieve exactly $1 - \alpha$ coverage, but it is not a sufficient condition for efficiency. Indeed, again, the naive predictor presented in Example 3.1.9 has a probability of coverage of exactly $1 - \alpha$ but is critically inefficient.

Remark 3.2.4 (the guarantee is conditional on Tr).

Importantly, remark here that the probability is taken over $\mathcal{D}^{\mathcal{E}(\text{Cal} \cup \{n+1\})}$, excluding the proper training set Tr : the validity is *conditional* on Tr , thus the validity holds conditionally on the fitted model $\hat{\mu}$, regardless of its accuracy.

Remark 3.2.5 (the guarantee is not conditional on X).

However, we insist again here that the probability controlled in Theorem 3.2.1 is not conditional neither on the training data nor on the test features X_{n+1} . In particular, for $x \in \mathcal{X}$, $\mathbb{P}\left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \mid X_{n+1} = x\right) \geq 1 - \alpha$.

Through SCP with absolute value of the mean-residuals, we move from a situation where two quantile regressions do not have any form of validity in finite sample and could under-cover drastically (Figure 3.5a), to a setting where we do achieve marginal validity in finite sample for any distribution (Figure 3.5b). However, in practice, one usually aims at X -conditional coverage (Figure 3.5c), a guarantee that is not achieved by SCP in mean-regression using the absolute value of the residuals as conformity scores. X -conditionally valid predictive sets are such that the random variable which is *the indicator of coverage is independent of X* , i.e. a point is equally likely to be covered whatever is the X -draw.

While marginal coverage allows the distribution of the indicator of coverage to vary across regions of the features space, i.e. the predictive sets can be non-adaptive, the stronger notion of X -conditional coverage ensures that the indicator of coverage is evenly distributed, i.e. the predictive sets are fully adaptive. These differences are illustrated in Figure 3.5. Therefore, a X -conditionally valid estimator of the predictive sets is necessarily adaptive to X .

However, SCP as described in the previous Section 3.2.1 is critically non-adaptive as its predictive intervals depend on the features x only through the intervals' location, but their shape is constant (symmetric and constant length accross the features space). One could think that a better methodology's design would then lead to guaranteed X -conditional coverage. Unfortunately, this is all the more wrong. As shown in Vovk (2012); Lei and Wasserman (2014); Barber et al. (2021a) and detailed later in Section 3.3.2, it is impossible to achieve *informative* X -conditional validity under our set of assumptions.

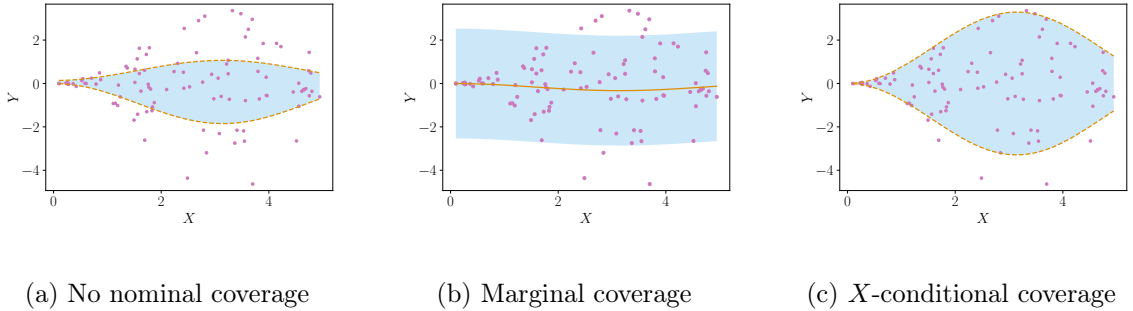


Figure 3.5: Illustration of various notion of coverage.

Informal theorem

Without distribution assumption, in finite sample, a perfectly X -conditionally valid $\widehat{C}_{n,\alpha}$ is such that $\mathbb{P} \left\{ \text{measure} \left(\widehat{C}_{n,\alpha}(x) \right) = \infty \right\} \geq 1 - \alpha$ for any non-atomic x .

In practice, as one can not accept to only have marginal coverage even empirically (see Figure 3.5b), there have been important research effort to get closer to X -conditional coverage. We can separate this line of work into two different branches: one trying to achieve some form of approximate conditional coverage in finite sample, i.e. they target $\mathbb{P}_{\mathcal{D}^{(n+1)}} \left(Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1}) \mid X_{n+1} \in \mathcal{V}(x) \right) \geq 1 - \alpha$ with $\mathcal{V}(x)$ representing some region or neighbourhood around x (Romano et al., 2020a; Guan, 2022; Jung et al., 2023; Gibbs et al., 2023, to name just a few), relying on the fact that the impossibility result naturally only holds for non-atomic points x , and on any atomic x an instinctive idea is to only calibrate with calibration points for which $X_i = x, i \in \text{Cal}$ (this is related to Mondrian CP which groups data points according to a family of groups \mathcal{G} to achieve \mathcal{G} -conditional validity, Vovk et al., 2005); and the other one aiming at asymptotic (with the sample size) X -conditional coverage based on the intuition that enjoying asymptotic theoretical guarantees goes hand in hand with enhanced empirical performances, these works are usually based on estimating the overall c.d.f. or p.d.f. of the data using consistent estimators (Romano et al., 2019; Kivaranovic et al., 2020; Cauchois et al., 2021; Chernozhukov et al., 2021; Sesia and Romano, 2021; Izbicki et al., 2022, among others). In the next subsection, we present one of them (Conformalized Quantile Regression, Romano et al., 2019) as it played a key role in the growing interest of the machine learning community towards CP, and is one of the most used CP algorithm in practice.

3.2.2 Conformalized Quantile Regression (CQR)

Conformalized Quantile Regression (CQR, Romano et al., 2019) first splits the n points of the training set into two disjoint sets $\text{Tr}, \text{Cal} \subset \llbracket 1, n \rrbracket$, to create a *proper training set*, Tr , and a *calibration set*, Cal . On the proper training set, two quantile regression algorithms (chosen by the user) are fitted ($\widehat{\text{QR}}_{\text{lower}}$ and $\widehat{\text{QR}}_{\text{upper}}$), and then used to predict on the calibration set. *Conformity scores* $s(x, y; \widehat{\text{QR}}_{\text{lower}}, \widehat{\text{QR}}_{\text{upper}}) := \max(\widehat{\text{QR}}_{\text{lower}}(x) - y, y - \widehat{\text{QR}}_{\text{upper}}(x))$ are computed to assess how well the fitted interval predicts the response values of the calibration points, forming the set $\mathcal{S} = \left\{ \left(S_i := s(X_i, Y_i; \widehat{\text{QR}}_{\text{lower}}, \widehat{\text{QR}}_{\text{upper}}) \right)_{i \in \text{Cal}} \right\} \cup \{+\infty\}$. Finally, the $(1 - \alpha)$ -th quantile of these scores $q_{1-\alpha}(\mathcal{S})$ is computed to define the correction of the predictive interval: $\widehat{C}_{n,\alpha}(\cdot) := \left[\widehat{\text{QR}}_{\text{lower}}(\cdot) - q_{1-\alpha}(\mathcal{S}); \widehat{\text{QR}}_{\text{upper}}(\cdot) + q_{1-\alpha}(\mathcal{S}) \right]$.

An illustration of CQR is given in Figure 3.6 for $d = 1$. Contrary to Figure 3.4, we use a heteroskedastic distribution to illustrate the impact and interest of the quantile regressions. The idea behind the new conformity scores is the following: the score is negative for any point that belongs to the initial interval, and positive otherwise (see also Step 2 in Figure 3.6). Hence, if the initial interval is too sharp (resp. overly conservative) then more (resp. less) than α of the scores will be positive, leading to a positive (resp. negative) $q_{1-\alpha}(\mathcal{S})$ and thus the final interval will be enlarged (resp. shrunked) in comparison with the initial interval, when adding $q_{1-\alpha}(\mathcal{S})$ to its bound. Finally, the value of the scores

Algorithm 2 CQR**Input:** Quantile regression algorithm \mathcal{A} , miscoverage rate α , training set $(X_i, Y_i)_{i=1}^n$ **Output:** Prediction interval $\hat{C}_{n,\alpha}$ 

- 1: Randomly split the training data $(X_i, Y_i)_{i=1}^n$ into a **proper training set** (size $\#Tr$) and a **calibration set** (size $\#Cal$)
- 2: Get \widehat{QR}_{lower} and \widehat{QR}_{upper} (by training \mathcal{A} on the **proper training set** $(X_i, Y_i)_{i \in Tr}$)
- 3: On the **calibration set**, get prediction values with \widehat{QR}_{lower} and \widehat{QR}_{upper}
- 4: Obtain a set of $\#Cal + 1$ **conformity scores**:

$$\mathcal{S} = \left\{ S_i = s \left(X_i, Y_i; \widehat{QR}_{lower}, \widehat{QR}_{upper} \right), i \in \text{Cal} \right\} \cup \{+\infty\},$$

$$\text{with } s \left(x, y; \widehat{QR}_{lower}, \widehat{QR}_{upper} \right) := \max \left(\widehat{QR}_{lower}(x) - y, y - \widehat{QR}_{upper}(x) \right)$$

Obtain a set of $\#Cal$ conformity scores: $\mathcal{S} = \{S_i, i \in \text{Cal}\}$

- 5: Compute the $1 - \alpha$ quantile of these scores: $q_{1-\alpha}(\mathcal{S})$

Compute the $\left((1 - \alpha) \left(1 + \frac{1}{\#Cal} \right) \right)$ quantile of these scores: $q_{1-\alpha}(\mathcal{S})$

- 6: For a new point X_{n+1} , return

$$\hat{C}_{n,\alpha}(X_{n+1}) = \left[\widehat{QR}_{lower}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}); \widehat{QR}_{upper}(X_{n+1}) + q_{1-\alpha}(\mathcal{S}) \right]$$

reflect how far the point is from the initial interval bound, conveying the information of how much enlargement or shrinkage is required to ensure marginal validity. Algorithm 2 provides a formal description.

Note that, exactly as for SCP for mean-regression with absolute values of the residuals as conformity scores, Remarks 3.2.1 and 3.2.2 (stating that the fitted model can in fact be independent of Tr , and explaining the reason behind the $+\infty$) apply to CQR. Due to the very same reason, one can start to feel that in fact SCP in mean-regression with absolute values of the residuals as conformity scores and CQR share the same construction. We will formalize this intuition further in Section 3.2.3. For now we state the theoretical validity of CQR, from Romano et al. (2019).

Theorem 3.2.2 (marginal validity of CQR).

CQR (Algorithm 2) outputs $\hat{C}_{n,\alpha}$ such that for any distribution \mathcal{D} , for any associated exchangeable joint distribution $\mathcal{D}^{\otimes(\text{Cal} \cup \{n+1\})} \in \mathcal{D}^{\text{exch}(\text{Cal} \cup \{n+1\})}$:

$$\mathbb{P}_{\mathcal{D}^{\otimes(\text{Cal} \cup \{n+1\})}} \left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \right) \geq 1 - \alpha.$$

Additionally, if the scores $\{S_i\}_{i \in \text{Cal}} \cup \{S_{n+1}\}$ are almost surely (a.s.) distinct:

$$\mathbb{P}_{\mathcal{D}^{\otimes(\text{Cal} \cup \{n+1\})}} \left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \right) \leq 1 - \alpha + \frac{1}{\#Cal + 1}.$$

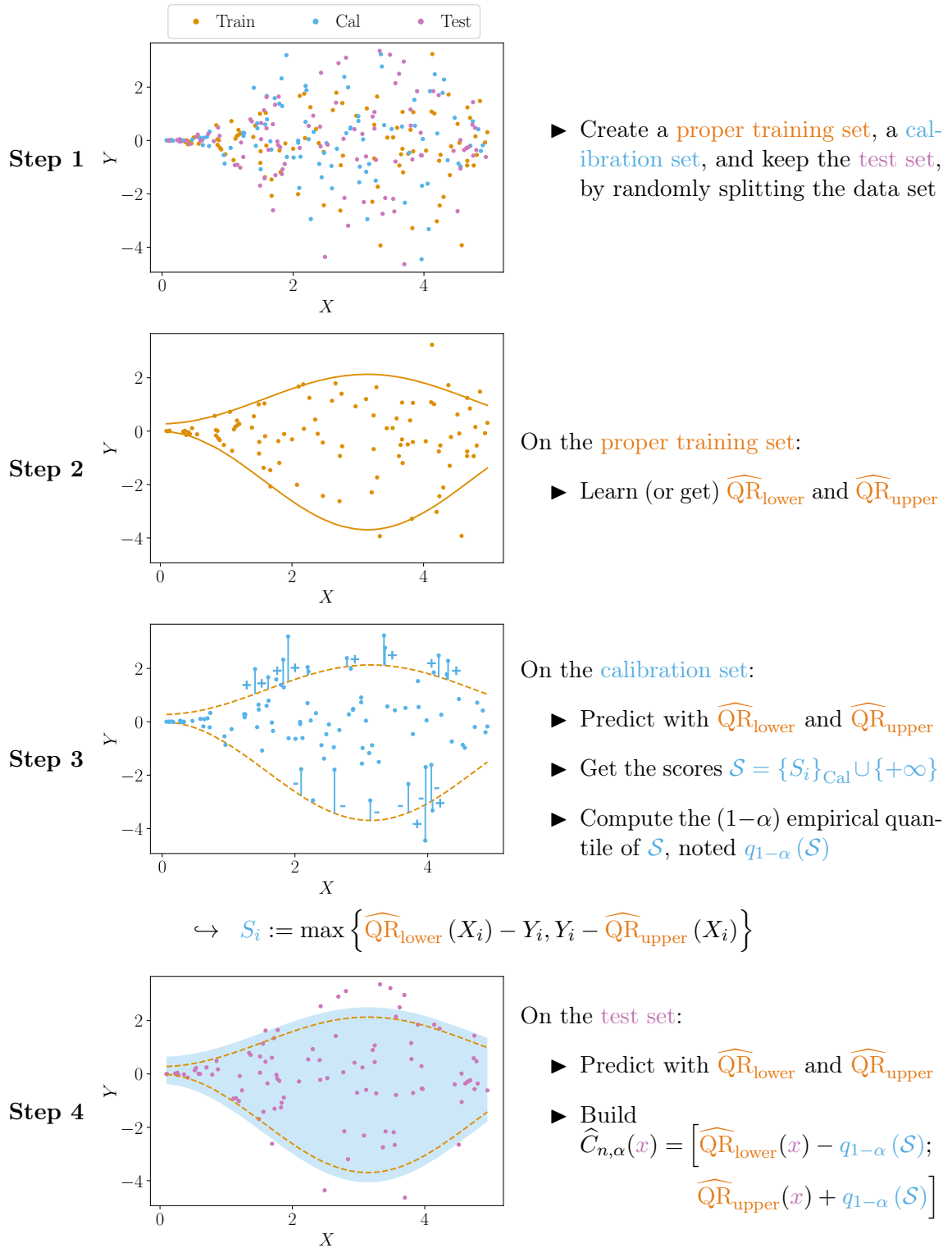


Figure 3.6: Schematic illustration of the Conformalized Quantile Regression procedure.

Proof. First, on any $(X_i, Y_i)_{i=1}^{n+1}$ exchangeable sequence, CQR builds scores $\{S_i\}_{i \in \text{Cal}} \cup \{S_{n+1}\}$ that are exchangeable due to Lemma 3.2.2. Then, observe that:

$$\begin{aligned}
\{Y_{n+1} \notin \widehat{C}_{n,\alpha}(X_{n+1})\} &= \left\{ Y_{n+1} < \widehat{\text{QR}}_{\text{lower}}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}) \right. \\
&\quad \left. \text{or } Y_{n+1} > \widehat{\text{QR}}_{\text{upper}}(X_{n+1}) + q_{1-\alpha}(\mathcal{S}) \right\} \\
&= \left\{ \widehat{\text{QR}}_{\text{lower}}(X_{n+1}) - Y_{n+1} > q_{1-\alpha}(\mathcal{S}) \right. \\
&\quad \left. \text{or } Y_{n+1} - \widehat{\text{QR}}_{\text{upper}}(X_{n+1}) > q_{1-\alpha}(\mathcal{S}) \right\} \\
&= \left\{ \max\left(\widehat{\text{QR}}_{\text{lower}}(X_{n+1}) - Y_{n+1}, Y_{n+1} - \widehat{\text{QR}}_{\text{upper}}(X_{n+1})\right) \right. \\
&\quad \left. > q_{1-\alpha}(\mathcal{S}) \right\} \\
&= \{S_{n+1} > q_{1-\alpha}(\mathcal{S})\} \\
\{Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1})\} &= \{S_{n+1} \leq q_{1-\alpha}(\mathcal{S})\}.
\end{aligned}$$

Note that this last equation is equivalent to Equation (3.1) above. Now, it only remains to apply the quantile lemma 3.2.1 to conclude the proof. \square

Remarks 3.2.4 and 3.2.5 apply to CQR as well: it is valid conditionally to Tr , and, importantly, even though it does improve X -conditional coverage in practice, it does not enjoy theoretical guarantees on this. This is expected given our discussion on the impossibility of X -conditional coverage in Section 3.3.2. It is even more expected as CQR is adaptive on X only through the quantile regression, while the conformal scores and correction are independent of X : the key step that is devoted to retrieving validity is independent of X , thus there was no hope for finite sample distribution-free X -conditional validity by design. However, [Sesia and Candès \(2020\)](#) provides asymptotic guarantees on X -conditional validity of CQR under consistency of the quantile regression algorithm.

Remark 3.2.6 (CQR validity holds regardless of the quantile regression levels).

The marginal validity of CQR holds for any quantile regression algorithm. This means that in particular, the levels of these quantile regressions can be picked arbitrarily. While a natural choice might be lower = $\alpha/2$ and upper = $1 - \alpha/2$, [Romano et al. \(2019\)](#) suggest to choose them via cross-validation as it seems to enhance the resulting intervals' efficiency.

3.2.3 Generalization of SCP: going beyond regression

As hinted by the design of Algorithms 1 and 2 and the proofs of the associated Theorems 3.2.1 and 3.2.2, SCP with absolute value of mean-regression residuals and CQR are in fact two particular instances of a global algorithm, SCP, that is general enough to even tackle the classification problems. SCP is a wrapper around any learning algorithm \mathcal{A} (e.g., any mean regressor for Algorithm 1, or any quantile regressor for Algorithm 2) that is fitted on an

Algorithm 3 General SCP

Input: Learning algorithm \mathcal{A} , conformity score function s , miscoverage rate α , training set $(X_i, Y_i)_{i=1}^n$

Output: Prediction set $\hat{C}_{n,\alpha}$



- 1: Randomly split the training data $(X_i, Y_i)_{i=1}^n$ into a **proper training set** (size $\#Tr$) and a **calibration set** (size $\#Cal$)
- 2: Get \hat{A} (by training \mathcal{A} on the **proper training set** $(X_i, Y_i)_{i \in Tr}$)
- 3: On the **calibration set**, obtain a set of $\#Cal + 1$ **conformity scores**:

$$\mathcal{S} = \{S_i = s(X_i, Y_i; \hat{A}), i \in \text{Cal}\} \cup \{+\infty\}$$

► SCP absolute value of the mean-residuals (Algorithm 1): $s(x, y; \hat{A}) := |y - \hat{\mu}(x)|$

► CQR (Algorithm 2): $s(x, y; \hat{A}) := \max(\widehat{QR}_{\text{lower}}(x) - y, y - \widehat{QR}_{\text{upper}}(x))$

- 4: Compute the $1 - \alpha$ quantile of these scores: $q_{1-\alpha}(\mathcal{S})$
- 5: For a new point X_{n+1} , return

$$\hat{C}_{n,\alpha}(X_{n+1}) = \{y \in \mathcal{Y} \text{ such that } s(X_{n+1}, y; \hat{A}) \leq q_{1-\alpha}(\mathcal{S})\}$$

In particular cases, this set boils down to:

- $\hat{C}_{n,\alpha}(X_{n+1}) = [\hat{\mu}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S})]$ in SCP absolute value of the mean-residuals
- $\hat{C}_{n,\alpha}(X_{n+1}) = [\widehat{QR}_{\text{lower}}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}), \widehat{QR}_{\text{upper}}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$ in CQR

independent training set to produce \hat{A} : given a conformity score function tailored to the learning algorithm \mathcal{A} (e.g., absolute value of the residuals in Algorithm 1, or the signed score of Algorithm 2), used to construct $\mathcal{S} = \left\{s(X_i, Y_i; \hat{A})\right\}_{i \in \text{Cal}} \cup \{+\infty\}$ in order to assess how well the fitted model \hat{A} predicts the response values of the calibration points, it builds a predictive set containing only the labels leading to a score on X_{n+1} which is smaller than a $1 - \alpha$ fraction of the calibration scores, i.e. $\{y \in \mathcal{Y} \text{ such that } s(x, y; \hat{A}) \leq q_{1-\alpha}(\mathcal{S})\}$. A pseudo-code of SCP is provided in Algorithm 3.

Theorem 3.2.3 (marginal validity of SCP).

SCP (Algorithm 3) outputs $\hat{C}_{n,\alpha}$ such that for any distribution \mathcal{D} , for any associated exchangeable joint distribution $\mathcal{D}^{\mathcal{S}(\text{Cal} \cup \{n+1\})} \in \mathcal{D}^{\text{exch}(\text{Cal} \cup \{n+1\})}$:

$$\mathbb{P}_{\mathcal{D}^{\mathcal{S}(\text{Cal} \cup \{n+1\})}}(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})) \geq 1 - \alpha.$$

Additionally, if the scores $\{S_i\}_{i \in \text{Cal}} \cup \{S_{n+1}\}$ are almost surely (a.s.) distinct:

$$\mathbb{P}_{\mathcal{D}^{\mathcal{S}(\text{Cal} \cup \{n+1\})}}(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})) \leq 1 - \alpha + \frac{1}{\#Cal + 1}.$$

Proof. First, on any $(X_i, Y_i)_{i=1}^{n+1}$ exchangeable sequence, SCP builds scores $\{S_i\}_{i \in \text{Cal}} \cup \{S_{n+1}\}$ that are exchangeable due to Lemma 3.2.2. Then, it only remains to apply the

quantile lemma 3.2.1. □

Remark 3.2.7 (randomized SCP).

To ensure that the upper bound always holds, even when ties among scores occur with non-zero probability, one can add a randomization in SCP algorithm.

Formally, before introducing this tie-breaking randomization, let us first rewrite the predictive set:

$$\begin{aligned}\hat{C}_{n,\alpha}(x) &= \left\{ y \in \mathcal{Y}, s(x, y; \hat{A}) \leq q_{1-\alpha}(\mathcal{S}) \right\} \\ &= \left\{ y \in \mathcal{Y}, s(x, y; \hat{A}) \leq q_{(1-\alpha)(1+1/\#\text{Cal})}((S_i)_{i \in \text{Cal}}) \right\} \\ &= \left\{ y \in \mathcal{Y}, \sum_{i \in \text{Cal}} \mathbb{1}\{s(x, y; \hat{A}) > S_i\} < (1-\alpha)(1+\#\text{Cal}) \right\} \\ &= \left\{ y \in \mathcal{Y}, \frac{1 + \sum_{i \in \text{Cal}} \mathbb{1}\{s(x, y; \hat{A}) \leq S_i\}}{1 + \#\text{Cal}} > \alpha \right\}\end{aligned}$$

The randomization consists instead in drawing $U \sim \mathcal{U}([0, 1])$ and outputting:

$$\hat{C}_{n,\alpha}^r(x) = \left\{ y \in \mathcal{Y}, \frac{\sum_{i \in \text{Cal}} \mathbb{1}\{s(x, y; \hat{A}) < S_i\} + U \left(1 + \sum_{i \in \text{Cal}} \mathbb{1}\{s(x, y; \hat{A}) = S_i\} \right)}{\#\text{Cal} + 1} > \alpha \right\}.$$

3.2.4 Some examples of SCP in classification

The generalized framework introduced in the previous section does not make any assumption on the label space \mathcal{Y} . Indeed, while we have only introduced regression-tailored algorithm so far, SCP—and more generally CP—is general enough to encapsulate classification tasks. Let us focus here in presenting two traditional SCP algorithms for classification.

The framework is the following. Assume that the label space is $\mathcal{Y} = \{1, \dots, C\} \subseteq \mathbb{N}^*$ where $C = \#\mathcal{Y}$ is the number of classes. We consider that the learning algorithm fits a model $\hat{A} \stackrel{\text{not.}}{=} \hat{p}$, which is a function that outputs a vector of estimated probabilities for each class (e.g., after a softmax layer).

A first idea of tailored conformity scores is $\mathbf{s}(x, y; \hat{p}) = 1 - \hat{p}(x)_y$. Indeed, by doing so the score is large (resp. small) when the model predicts a low (resp. high) estimated probability on the true class. Note that now, the predictive set $\hat{C}_{n,\alpha}(X_{n+1}) = \{y \text{ such that } \mathbf{s}(X_{n+1}, y; \hat{p}) \leq q_{1-\alpha}(\mathcal{S})\}$ does not boil down to any explicit expression, and we have to try all the possible y . As \mathcal{Y} is finite, unlike in the regression setting, this task is doable. Examples 3.2.4 and 3.2.5 provide a toy example of how such an algorithm would work in practice. In these examples, we emphasize that (i) the quality of the fitted model impacts the size of the predictive set (to see this, compare the predictive set of Example 3.2.4 to the one of Example 3.2.5), as discussed previously; (ii) the level of difficulty to predict on test point is poorly reflected in the predictive set (to see this, the text in gray shows that the final predictive sets stay constant on a different prediction). Point (ii) is due to

the design of the conformity score, see Remark 3.2.8.

Remark 3.2.8 (efficiency yet non-adaptivity of the simplest classification scores).

While this conformity score function allows to output the most efficient set possible (i.e. achieving the smallest average set size, [Sadinle et al., 2018](#)), it does not allow to discriminate between “easy” and “hard” test point. In practice, it leads to predictive sets that under-cover (resp. over-cover) on “hard” (resp. “easy”) subgroups. This is due to the fact that the same threshold $q_{1-\alpha}(\mathcal{S})$ is applied to any test point.

Example 3.2.4 (toy use case of classification SCP with the simplest score).

Let consider a toy use case where we want to classify households according to the best electricity tariff to propose them in order to align electricity production and consumption (this is a simplified example of demand-side management). In this context, assume $\mathcal{Y} = \{\text{“N”}, \text{“B”}, \text{“D”}\}$ where “N” stands for neutral (constant standard tariff), “B” stands for bitariff (such as (off)-peak hours, with lower and higher tariffs) and “D” stands for dynamic (i.e. the price switches between low, standard and high tariffs depending of the day with 2 days early notice to the consumers).

We want to build predictive sets at the level $\alpha = 0.1$, and we have access to a calibration data set with $\#\text{Cal} = 10$ points.

► **Unconfident fitted model**

1. Compute the scores on the calibration set using $\mathbf{s}(x, y; \hat{p}) = 1 - \hat{p}(x)_y$.

$Y_i, i \in \text{Cal}$	“N”	“N”	“N”	“B”	“B”	“B”	“B”	“D”	“D”	“D”
$\hat{p}_N(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_B(X_i)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_D(X_i)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
S_i	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

Define $\mathcal{S} = \{S_i, i \in \text{Cal}\} \cup \{+\infty\}$.

2. Compute their empirical quantile: $q_{1-\alpha}(\mathcal{S}) = 0.65$.

3. Predict on a new point X_{n+1} : $\hat{p}(X_{n+1}) = (0.05, 0.60, 0.35)$.

(or less predictable: $\hat{p}(X_{n+1}) = (0.25, 0.4, 0.35)$)

4. For each possible label, evaluate the scores on this new point

$$\begin{aligned}
 &\hookrightarrow \mathbf{s}(X_{n+1}, \text{“N”}; \hat{p}) = 0.95 \text{ (or } 0.75) && \text{“N”} \notin \hat{C}_{n,\alpha}(X_{n+1}) \\
 &\hookrightarrow \mathbf{s}(X_{n+1}, \text{“B”}; \hat{p}) = 0.40 \leq q_{1-\alpha}(\mathcal{S}) \text{ (or } 0.6 \leq q_{1-\alpha}(\mathcal{S})) && \text{“B”} \in \hat{C}_{n,\alpha}(X_{n+1}) \\
 &\hookrightarrow \mathbf{s}(X_{n+1}, \text{“D”}; \hat{p}) = 0.65 \leq q_{1-\alpha}(\mathcal{S}) \text{ (or } 0.65 \leq q_{1-\alpha}(\mathcal{S})) && \text{“D”} \in \hat{C}_{n,\alpha}(X_{n+1})
 \end{aligned}$$

5. Form the predictive set associated to X_{n+1} : $\hat{C}_{n,\alpha}(X_{n+1}) = \{\text{“B”}, \text{“D”}\}$.

Example 3.2.5 (toy use case of classification SCP with the simplest score).

Let consider again the demand-side management toy use case where $\mathcal{Y} = \{\text{"N"}, \text{"B"}, \text{"D"}\}$, and we wish to build predictive sets at the level $\alpha = 0.1$. Assume we have access to a calibration data set with $\#\text{Cal} = 10$ points.

► **Confident fitted model**

1. Compute the scores on the calibration set (compared to the previous example above, the subsequent scores are less uniform as we illustrate the case where the underlying model is more truthfully confident).

$Y_i, i \in \text{Cal}$	"N"	"N"	"N"	"B"	"B"	"B"	"B"	"D"	"D"	"D"
$\hat{p}_N(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.05	0.10	0.10	0.15
$\hat{p}_B(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.70	0.25	0.30	0.30
$\hat{p}_D(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.25	0.65	0.60	0.55
S_i	0.05	0.1	0.15	0.15	0.20	0.25	0.30	0.35	0.40	0.45

Define $\mathcal{S} = \{S_i, i \in \text{Cal}\} \cup \{+\infty\}$.

2. Compute their empirical quantile: $q_{1-\alpha}(\mathcal{S}) = 0.45$.
3. Predict on a new point X_{n+1} : $\hat{p}(X_{n+1}) = (0.05, 0.60, 0.35)$.
(or more predictable: $\hat{p}(X_{n+1}) = (0.05, 0.9, 0.05)$)
4. For each possible label, evaluate the scores on this new point X_{n+1} .
 - ↪ $\mathbf{s}(X_{n+1}, \text{"N"}; \hat{p}) = 0.95$ (or 0.95) "N" $\notin \hat{C}_{n,\alpha}(X_{n+1})$
 - ↪ $\mathbf{s}(X_{n+1}, \text{"B"}; \hat{p}) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$ (or $0.1 \leq q_{1-\alpha}(\mathcal{S})$) "B" $\in \hat{C}_{n,\alpha}(X_{n+1})$
 - ↪ $\mathbf{s}(X_{n+1}, \text{"D"}; \hat{p}) = 0.65$ (or 0.95) "D" $\notin \hat{C}_{n,\alpha}(X_{n+1})$
5. Form the predictive set associated to X_{n+1} : $\hat{C}_{n,\alpha}(X_{n+1}) = \{\text{"B"}\}$.

Other conformity score functions can be used to alleviate this issue and improve adaptiveness. One of them was proposed in Romano et al. (2020b) and is based on the intuitive idea that one may want to include classes by decreasing order of estimated probabilities until reaching a theoretically valid threshold, that might be different from $1 - \alpha$. Formally, given a predictor of estimated probabilities $\hat{p}(\cdot)$, for any $x \in \mathcal{X}$ define $\sigma_x : \{1, \dots, \#\mathcal{Y}\} \mapsto \mathcal{Y}$ such that $\hat{p}(x)_{\sigma_x(1)} \geq \dots \geq \hat{p}(x)_{\sigma_x(\#\mathcal{Y})}$. In other words, σ_x associates the descending ordering of the estimated probabilities on x . Then, for any given features $x \in \mathcal{X}$, and any label $y \in \mathcal{Y}$, the conformity score function is $\mathbf{s}(x, y; \hat{p}) := \sum_{l=1}^{\sigma_x^{-1}(y)} \hat{p}(x)_{\sigma_x(l)}$, that is, the sum of the estimated probabilities associated to classes at least as large as that of the true class y . Finally, on a test point X_{n+1} , it returns the set of classes $\{\sigma_{X_{n+1}}(1), \dots, \sigma_{X_{n+1}}(r^*)\}$, where $r^* := \arg\max_{1 \leq r \leq C} \left\{ \sum_{k=1}^r \hat{p}(X_{n+1})_{\sigma_{X_{n+1}}(k)} < q_{1-\alpha}(\mathcal{S}) \right\} + 1$. An illustration of the scores and predictive set construction is provided in Figure 3.7, along with a detailed toy use case example in Example 3.2.6 which highlights that this time the predictive sets adapts to the complexity of the test point.

Example 3.2.6 (toy use case of classification SCP with adaptive score).

Let consider again the demand-side management toy use case where $\mathcal{Y} = \{\text{"N"}, \text{"B"}, \text{"D"}\}$, and we wish to build predictive sets at the level $\alpha = 0.1$. Assume we have access to a calibration data set with $\#\text{Cal} = 10$ points.

1. Compute the scores on the calibration set using $\mathbf{s}(x, y; \hat{p}) := \sum_{l=1}^{\sigma_x^{-1}(y)} \hat{p}(x)_{\sigma_x(l)}$.

$Y_i, i \in \text{Cal}$	"N"	"N"	"N"	"B"	"B"	"B"	"B"	"D"	"D"	"D"
$\hat{p}_N(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_B(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_D(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

Define $\mathcal{S} = \{S_i, i \in \text{Cal}\} \cup \{+\infty\}$.

2. Compute their empirical quantile: $q_{1-\alpha}(\mathcal{S}) = 0.95$.

► Unconfident prediction on the test point:

3. Predict on a new point X_{n+1} , evaluate r^* to reach $q_{1-\alpha}(\mathcal{S})$ and obtain the associated predictive set:

$$\hat{p}(X_{n+1}) = (0.05, 0.45, 0.5), r^* = 2 \implies \hat{C}_{n,\alpha}(X_{n+1}) = \{\text{"B"}, \text{"D"}\}$$

► Confident prediction on the test point:

- 3bis. Predict on a new point X_{n+1} , evaluate r^* to reach $q_{1-\alpha}(\mathcal{S})$ and obtain the associated predictive set:

$$\hat{p}(X_{n+1}) = (0.03, 0.95, 0.02), r^* = 1 \implies \hat{C}_{n,\alpha}(X_{n+1}) = \{\text{"B"}\}$$

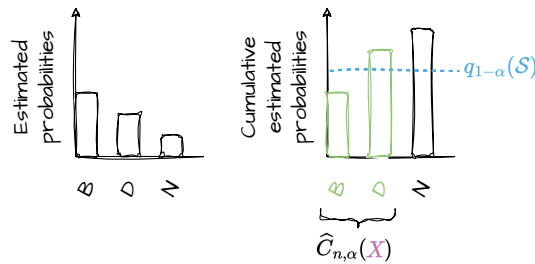


Figure 3.7: Illustration of Romano et al. (2020b) predictive sets construction. Figure highly inspired by Angelopoulos and Bates (2023).

Wrapping up

We have described a simple procedure, coined Split Conformal Prediction—a special case of the more generic framework of CP described in Section 3.4—, which quantifies the uncertainty of **any predictive algorithm** \mathcal{A} by returning predictive sets that enjoy **finite sample distribution-free** coverage guarantees, as long as the data set is *exchangeable*. In the reminder of this introductory chapter, our goal is to discuss *inherent bottlenecks* of (split) CP and provide an overview of the current research’ state in addressing them. Namely, Section 3.3 develops on the conditional guarantees, both empirically and theoretically; in Section 3.4 we present CP approaches that alleviate the statistical cost of data splitting; and lastly, in Section 3.5 we discuss extensions of CP when the unique assumption—data exchangeability—is not met. The research community on conformal methods has been growing quickly in the recent years. Therefore, these research directions are not exhaustive, the current research effort including also many branches that develop CP in specific domains.

3.3 On the design choices of CP and (empirical) conditional guarantees

Intrinsically, CP guarantees hold marginally over the test point (its features and its label) as well as marginally over the calibration set. They are obtained thanks to the fact that the *conformity scores* built by SCP are exchangeable. This is a fundamental point: the key step, and in a sense the definition, of CP (beyond SCP) is the **construction of exchangeable conformity scores**. In this section, we precisely propose to analyse the impact of the conformity scores definition, and then to study what conditional guarantees can be obtained by CP (beyond SCP).

3.3.1 What choices for the conformity scores?

The conformity scores are the cornerstone of CP, and their definition is crucial as they are the random variables that incorporate all the underlying information: the data distribution along with the fitted model behavior. A badly designed conformity function leads to predictive sets that are uninformative: taking an extreme case, an uninformed but legit possibility is to draw the scores i.i.d. from any exogenous distribution, e.g., $\mathcal{N}(0, 1)$. While the resulting predictive sets do not convey any useful information, this procedure benefits from the theoretical framework of CP and is valid. A more down-to-earth analysis is to remember the insights of the previous Section 3.2: while for any score function, the guarantees are marginal over nearly all the problem’s randomness, yet some score functions are associated to predictive sets empirically closer to conditional validity (e.g., CQR is closer to conditional validity than SCP with absolute value mean-residuals, adaptive classification (Example 3.2.6) is closer to conditional validity than the simplest classification case (Examples 3.2.4 and 3.2.5)).

Focusing temporarily on the regression setting, Table 3.1 illustrates the impact of the conformity score. All the methods presented in this table enjoy the exact same theoretical guarantees. However, their empirical performances and uses differ drastically. On the one hand, SCP with mean-regression and absolute value of the residuals is critically non-adaptive while CQR benefits from enhanced adaptivity. On the other hand, CQR can not be plugged in an operational pipeline predicting a mean value (i.e., using CQR, one can not say “The electricity prices tomorrow should be $90\text{€}/\text{MWh} \pm 5\text{€}/\text{MWh}$.” but only e.g., “The electricity prices tomorrow should be in between $87\text{€}/\text{MWh}$ and $97\text{€}/\text{MWh}$ ”²) unlike for SCP with mean-regression and absolute value of the residuals. In fact, they are in-betweens, and for example, in this figure, we add another conformity score function that we did not cover before and which is slightly less adaptive than CQR but is plugged on top of a mean-regression algorithm. Introduced in [Lei et al. \(2018\)](#), it consists in reweighting the absolute value of the residuals by an estimation of the dispersion of the exact same residuals $\hat{\rho}$.

Designing insightful conformity score function might appear intricate. In practice, it can be easier to think about the desired shape of the predictive sets. Interestingly, [Gupta et al. \(2022\)](#) shows that SCP’s output can be obtained equivalently through the design of the predictive sets themselves instead of defining the conformity function s . A model \hat{A} (chosen by the user) is fitted on the proper training set as in SCP. Then, a sequence of nested predictive sets taking their values in \mathcal{Y} is built, $\left(R_t(\cdot; \hat{A})\right)_{t \in \mathcal{T}}$ for some $\mathcal{T} \subseteq \mathbb{R}$, such that for any $t \leq t' \in \mathcal{T}^2$, for any $x \in \mathcal{X}$, $R_t(x; \hat{A}) \subseteq R_{t'}(x; \hat{A})$, and at the limits $R_{\inf \mathcal{T}} \equiv \emptyset$ and $R_{\sup \mathcal{T}} \equiv \mathcal{Y}$. For instance, with a mean regression, the parallel of the absolute values of the residuals conformity scores in terms of nested sets leads to $R_t(\cdot; \hat{\mu}) \equiv [\hat{\mu}(\cdot) \pm t]$ and $\mathcal{T} = \mathbb{R}_+$. Entry radius of y in the sets given by x are then computed on each of the calibration points as $\hat{r}(x, y) := \inf \left\{ t \in \mathcal{T} : y \in R_t(x; \hat{A}) \right\}$. Then, under exchangeability

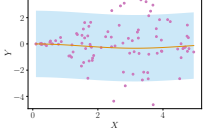
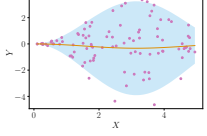
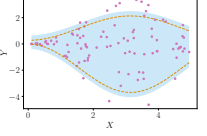
	Simplest SCP Vovk et al. (2005)	Locally weighted SCP Lei et al. (2018)	CQR Romano et al. (2019)
$s(\hat{A}(X), Y)$	$ \hat{\mu}(X) - Y $	$\frac{ \hat{\mu}(X) - Y }{\hat{\rho}(X)}$	$\max(\widehat{\text{QR}}_{\text{lower}}(X) - Y, Y - \widehat{\text{QR}}_{\text{upper}}(X))$
$\hat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})\hat{\rho}(x)]$	$[\widehat{\text{QR}}_{\text{lower}}(x) - q_{1-\alpha}(\mathcal{S}); \widehat{\text{QR}}_{\text{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$
Visu.			
✓	black-box around a “usable” prediction	black-box around a “usable” prediction	adaptive
✗	not adaptive	limited adaptiveness	no black-box around a “usable” prediction

Table 3.1: A comparison of some classical regression conformity scores.

²Note that to overcome this, an idea is to apply CQR directly on residuals of a mean-regression model.

of the data points, $(\hat{r}(X_i, Y_i))_{i=1}^n$ are exchangeable and play the role of the conformity scores. Denote the set of entry radii $\mathcal{R} = \{(\hat{r}(X_i, Y_i))_{i \in \text{Cal}}\} \cup \{+\infty\}$. We can finally define the predictive set as $\hat{C}_{n,\alpha}(x) := R_{q_{1-\alpha}(\mathcal{R})}(x; \hat{A}) = \{y \in \mathcal{Y} \text{ such that } \hat{r}(x, y) \leq q_{1-\alpha}(\mathcal{R})\}$. This formalism is appealing as it allows to first design the geometric shape of the predictive set, and only then deduce the algorithm to be deployed in order to output it. To illustrate this, we provide below some canonical examples of equivalences between the conformity score and the nested sets points of view.

Example 3.3.1 (Nested sets for the absolute value of the mean-regression residuals).

$$s(x, y; \hat{\mu}) = |y - \hat{\mu}(x)| \iff \begin{cases} R_t(\cdot; \hat{\mu}) \equiv [\hat{\mu}(\cdot) \pm t] \\ \mathcal{T} = \mathbb{R}_+ \end{cases}$$

Example 3.3.2 (Nested sets for CQR).

$$\begin{aligned} s(x, y; (\widehat{\text{QR}}_{\text{lower}}, \widehat{\text{QR}}_{\text{upper}})) \\ = \max(\widehat{\text{QR}}_{\text{lower}}(x) - y, y - \widehat{\text{QR}}_{\text{upper}}(x)) \end{aligned} \iff \begin{cases} R_t(\cdot; (\widehat{\text{QR}}_{\text{lower}}, \widehat{\text{QR}}_{\text{upper}})) \\ \equiv [\widehat{\text{QR}}_{\text{lower}}(\cdot) - t; \widehat{\text{QR}}_{\text{upper}}(\cdot) + t] \\ \mathcal{T} = \mathbb{R}_+ \end{cases}$$

Example 3.3.3 (Nested sets for the simplest classification).

$$s(x, y; \hat{p}) = 1 - \hat{p}(x)_y \iff \begin{cases} R_t(\cdot; \hat{p}) \equiv \{k \in \mathcal{Y} : \hat{p}(\cdot)_k \geq 1 - t\} \\ \mathcal{T} = [0, 1] \end{cases}$$

Example 3.3.4 (Nested sets for adaptive scores in classification).

Given a predictor of estimated probabilities $\hat{p}(\cdot)$, for any $x \in \mathcal{X}$, define $\sigma_x : \{1, \dots, \#\mathcal{Y}\} \mapsto \mathcal{Y}$ such that $\hat{p}(x)_{\sigma_x(1)} \geq \dots \geq \hat{p}(x)_{\sigma_x(\#\mathcal{Y})}$. In other words, σ_x associates the descending ordering.

Let $x \in \mathcal{X}$.

$$s(x, y; \hat{p}) = \sum_{l=1}^{\sigma_x^{-1}(y)} \hat{p}(x)_{\sigma_x(l)} \iff \begin{cases} R_t(x; \hat{p}) = \left\{ k \in \mathcal{Y} : \sum_{l=1}^{\sigma_x^{-1}(k)} \hat{p}(x)_{\sigma_x(l)} \leq t \right\} \\ \mathcal{T} = [0, 1] \end{cases}$$

3.3.2 On distribution-free X -conditional validity

Some scores allow to get “closer” to X -conditional coverage than others. However, unfortunately, as sketched at the end of Section 3.2.1, it is impossible to achieve informative distribution-free X -conditional validity. To state this negative result (that traces back to Vovk, 2012; Lei and Wasserman, 2014), let us first formally defined distribution X -conditional validity.

Definition 3.3.1 (distribution-free X -conditional validity).

An estimator $\hat{C}_{n,\alpha}$ achieves distribution-free X -conditional validity if for any distribution \mathcal{D} , for any associated exchangeable joint distribution $\mathcal{D}^{\otimes(n+1)} \in \mathcal{D}^{\text{exch}(n+1)}$, we have that:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}} \left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \mid X_{n+1} \right) \stackrel{a.s.}{\geq} 1 - \alpha.$$

Theorem 3.3.1 (impossibility of informative X -conditional validity).

Assume $\hat{C}_{n,\alpha}$ is distribution-free X -conditionally valid. Then, for any \mathcal{D} , for \mathcal{D}_X -almost all \mathcal{D}_X -non-atoms $x \in \mathcal{X}$:

- *Regression*: $\mathbb{P}_{\mathcal{D}^{\otimes(n)}} \left(\Lambda \left(\hat{C}_{n,\alpha}(x) \right) = \infty \right) \geq 1 - \alpha$, with Λ designing the Lebesgue measure,
- *Classification*: for any $y \in \mathcal{Y}$, $\mathbb{P}_{\mathcal{D}^{\otimes(n)}} \left(y \in \hat{C}_{n,\alpha}(x) \right) \geq 1 - \alpha$.

We provide below a proof which is highly inspired from the ones in [Vovk \(2012\)](#); [Lei and Wasserman \(2014\)](#), but the former is not constructive and the latter made the additional strong assumption that $\hat{C}_{n,\alpha}$ is also training-conditional. The remarks on Theorem 3.3.1 are deferred after this proof.

Proof. Assume $\hat{C}_{n,\alpha}$ be X -conditionally valid, as defined in Definition 3.3.1.

Let P a distribution on $\mathcal{X} \times \mathcal{Y}$, and let $x_0 \in \text{non-atom}(P_X)$.

Let $\varepsilon > 0$. Let $\varepsilon_n = \sqrt{2 \left(1 - \left(1 - \frac{\varepsilon^2}{2} \right)^{1/n} \right)}$.

Let $E \subseteq \mathcal{X}$ such that $x_0 \in E$ and $0 < P_X(E) \leq \varepsilon_n$ (this is possible as a non-atom of a distribution P_X belongs to its support).

Before diving in the details of the proof, let us define the total variation distance between two distributions P and Q on \mathcal{Z} , denoted $TV(P, Q)$:

$$TV(P, Q) := \sup_{Z \in \mathcal{Z}} |P(Z) - Q(Z)|.$$

► *Classification case.*

Let $y \in \mathcal{Y}$.

Define Q another distribution on $\mathcal{X} \times \mathcal{Y}$ such that for any $A \subseteq \mathcal{X}$ and for any $B \subseteq \mathcal{Y}$:

$$Q(A \times B) = P(A \cap E^c \times B) + P_X(A \cap E) S_y(B),$$

with S_y defined on \mathcal{Y} , which is a dirac on y .

On the one hand, exactly as in the regression case, by construction, $TV(P, Q) \leq P_X(E) \leq \varepsilon_n$. Hence, using Lemma 3.3.1, $TV(P^{\otimes(n)}, Q^{\otimes(n)}) \leq \varepsilon$. Therefore, for any $A \subseteq \mathcal{X}$ and for any $B \subseteq \mathcal{Y}$:

$$P^{\otimes(n)}(A \times B) \geq Q^{\otimes(n)}(A \times B) - \varepsilon. \quad (3.2)$$

On the other hand, let $x \in E$. As $\widehat{C}_{n,\alpha}$ is distribution-free X -conditionally valid, it satisfies:

$$\begin{aligned} 1 - \alpha &\leq \mathbb{P}_{Q^{n+1}} \left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}(x) | X^{(n+1)} = x \right) \\ &= \mathbb{E}_{Q^n} \left[\mathbb{E}_Q \left[\mathbb{1} \left\{ Y^{(n+1)} \in \widehat{C}_{n,\alpha}(x) \right\} | X^{(n+1)} = x \right] \right] \\ &= \mathbb{E}_{Q^n} \left[\mathbb{E}_Q \left[\mathbb{1} \left\{ y \in \widehat{C}_{n,\alpha}(x) \right\} | X^{(n+1)} = x \right] \right] \\ &= \mathbb{E}_{Q^n} \left[\mathbb{1} \left\{ y \in \widehat{C}_{n,\alpha}(x) \right\} \right] \\ &= \mathbb{P}_{Q^n} \left(y \in \widehat{C}_{n,\alpha}(x) \right). \end{aligned}$$

Combining with Equation (3.2), we finally get:

$$\mathbb{P}_{P^n} \left(y \in \widehat{C}_{n,\alpha}(x) \right) \geq 1 - \alpha - \varepsilon,$$

which concludes the proof for the classification case by letting $\varepsilon \rightarrow 0$.

► *Regression case.*

Let $D > 0$.

Define Q another distribution on $\mathcal{X} \times \mathcal{Y}$ such that for any $A \subseteq \mathcal{X}$ and for any $B \subseteq \mathcal{Y}$:

$$Q(A \times B) := P(A \cap E^c \times B) + P_X(A \cap E) R(B),$$

with R defined on \mathcal{Y} , uniform on $[-D; D]$.

On the one hand, by construction, $TV(P, Q) \leq P_X(E) \leq \varepsilon_n$. Hence, using Lemma 3.3.1, $TV(P^{\otimes(n)}, Q^{\otimes(n)}) \leq \varepsilon$. Therefore, for any $A \subseteq \mathcal{X}$ and for any $B \subseteq \mathcal{Y}$:

$$P^{\otimes(n)}(A \times B) \geq Q^{\otimes(n)}(A \times B) - \varepsilon. \quad (3.2)$$

On the other hand, let $x \in E$. As $\widehat{C}_{n,\alpha}$ is distribution-free X -conditionally valid, it satisfies:

$$\begin{aligned} 1 - \alpha &\leq \mathbb{P}_{Q^{n+1}} \left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}(x) | X^{(n+1)} = x \right) \\ &= \mathbb{E}_{Q^{\otimes(n)}} \left[\int_{\widehat{C}_{n,\alpha}(x)} q(y|x) dy \right] \\ &= \mathbb{E}_{Q^{\otimes(n)}} \left[\Lambda \left(\widehat{C}_{n,\alpha}(x) \cap [-D; D] \right) \times \frac{1}{2D} \right]. \end{aligned}$$

Note that $\Lambda \left(\widehat{C}_{n,\alpha}(x) \cap [-D; D] \right) \times \frac{1}{2D} \leq 1$. Therefore, using Lemma 3.3.2, for any $t > 0$:

$$\begin{aligned} \mathbb{P}_{Q^{\otimes(n)}} \left(\Lambda \left(\widehat{C}_{n,\alpha}(x) \cap [-D; D] \right) \times \frac{1}{2D} \geq 1 - t \right) &\geq 1 - \frac{\alpha}{t} \\ \mathbb{P}_{Q^{\otimes(n)}} \left(\Lambda \left(\widehat{C}_{n,\alpha}(x) \cap [-D; D] \right) \geq (1 - t)2D \right) &\geq 1 - \frac{\alpha}{t} \\ \Rightarrow \mathbb{P}_{Q^{\otimes(n)}} \left(\Lambda \left(\widehat{C}_{n,\alpha}(x) \right) \geq (1 - t)2D \right) &\geq 1 - \frac{\alpha}{t}. \end{aligned}$$

Let $t = 1 - \frac{1}{\sqrt{D}}$ and obtain $\mathbb{P}_{Q^{\otimes(n)}} \left(\Lambda \left(\widehat{C}_{n,\alpha}(x) \right) \geq 2\sqrt{D} \right) \geq 1 - \frac{\alpha}{1 - \frac{1}{\sqrt{D}}}$.

Combining with Equation (3.2), we finally get:

$$\mathbb{P}_{P^{\otimes(n)}} \left(\Lambda \left(\hat{C}_{n,\alpha}(x) \right) \geq 2\sqrt{D} \right) \geq 1 - \frac{\alpha}{1 - \frac{1}{\sqrt{D}}} - \varepsilon.$$

Letting $\varepsilon \rightarrow 0$ and $D \rightarrow +\infty$, the result is proven for the regression case. \square

This proof relies on the following Lemmas 3.3.1 and 3.3.2, whose proofs are available in Section 8.A.

Lemma 3.3.1 (total variation distance between i.i.d. distributions).

For P and Q two probability distributions, and $n \in \mathbb{N}^*$, it holds:

$$TV \left(P^{\otimes(n)}, Q^{\otimes(n)} \right) \leq \sqrt{2 \left(1 - \left(1 - \frac{TV(P, Q)^2}{2} \right)^n \right)}.$$

Lemma 3.3.2 (concentration for bounded random variable with high expectation).

Let Z be a random variable such that $0 \leq Z \leq 1$ and $\mathbb{E}[Z] \geq \beta$ with $\beta \in [0, 1]$. Then, for any $t > 0$, it holds $\mathbb{P}(Z \geq 1 - t) \geq 1 - \frac{1-\beta}{t}$.

Remark 3.3.1 (distribution-free X -conditional hardness result apply beyond CP).

Theorem 3.3.1 proves that if an estimator is X -conditionally valid on all distributions $\mathcal{D}^{\text{exch}(n+1)}$, then its predictive sets will necessarily be critically large and thus uninformative. To put it differently, this result holds for any estimator that is X -conditionally valid on all distributions $\mathcal{D}^{\text{exch}(n+1)}$, regardless on its underlying construction, which implies that the impossibility result holds beyond CP approaches.

Remark 3.3.2 (X -conditional estimators are overly large even on easy cases).

Theorem 3.3.1 proves that if an estimator is distribution-free X -conditionally valid, then under any given \mathcal{D} , its predictive sets will necessarily be critically large and thus uninformative. Crucially, it implies that on *any* distribution \mathcal{D} including the “nicest” ones (e.g., say Y is constant), the predictive set is useless: this is because in order to be X -conditionally valid on all distributions $\mathcal{D}^{\text{exch}(n+1)}$ it has to be overly conservative in any situation to ensure X -conditional coverage on more complex distributions.

Remark 3.3.3 (the lower bounds in Theorem 3.3.1 are tight).

Notice that, again, the naive predictor presented in Example 3.1.9—outputting \mathcal{Y} with probability $1 - \alpha$ and the empty set otherwise—is perfectly distribution-free conditionally valid (on X , on the calibration set, and on Y). However, the probability that its regression sets have infinite measure, or that its classification sets include any given label y , is exactly $1 - \alpha$ as both events only occur when it outputs \mathcal{Y} . Therefore, the lower bound in Theorem 3.3.1 is tight.

Remark 3.3.4 (interpretation of the classification case).

For classification, the result of Theorem 3.3.1 implies that *every label* is likely to be included in any distribution-free X -conditionally valid predictive set. Henceforth, the predictive set is likely to be large: especially, for any \mathcal{D} , for \mathcal{D}_X -almost all \mathcal{D}_X -non-atoms $x \in \mathcal{X}$, $\mathbb{E}_{\mathcal{D}^{\otimes(n)}} [\#\hat{C}_{n,\alpha}(x)] \geq (1 - \alpha)\#\mathcal{Y}$.

A natural question now is: can we relax the notion of X -conditional validity to make it a less lofty goal? Some elements of answer are provided in Barber et al. (2021a) in the regression setting. Their main result studies the following relaxation.

Definition 3.3.2 (distribution-free $(1 - \alpha, \delta)$ - X -conditional validity).

Let $\delta > 0$ be a tolerance level.

An estimator $\hat{C}_{n,\alpha}$ achieves distribution-free $(1 - \alpha, \delta)$ - X -conditional validity if for any distribution \mathcal{D} , for any $\mathcal{X} \subseteq \mathcal{X}$ such that $\mathbb{P}_{\mathcal{D}_X}(X \in \mathcal{X}) \geq \delta$, and for any associated exchangeable joint distribution $\mathcal{D}^{\otimes(n+1)} \in \mathcal{D}^{\text{exch}(n+1)}$, we have:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) | X_{n+1} \in \mathcal{X}) \geq 1 - \alpha.$$

The idea behind Definition 3.3.2 is that for any region of the features space that is large enough (in probability), then validity should be achieved on this region.

Theorem 3.3.2 (hardness of informative $(1 - \alpha, \delta)$ - X -conditional validity).

Let $\delta > 0$ be a tolerance level.

Assume $\hat{C}_{n,\alpha}$ is distribution-free $(1 - \alpha, \delta)$ - X -conditionally valid. Then, for any \mathcal{D} such that \mathcal{D}_X does not have atoms, it holds:

$$\mathbb{E}_{\mathcal{D}^{\otimes(n+1)}} [\Lambda(\hat{C}_{n,\alpha}(X_{n+1}))] \geq \inf_{c \in [0,1]} \left\{ \frac{1 - \alpha}{1 - c\alpha} \Delta_{1 - c\alpha\delta}^{\min} \right\},$$

where $\Delta_{1 - c\alpha\delta}^{\min} := \inf_{(1 - c\alpha\delta)\text{-MV estimators}} \left\{ \mathbb{E}_{\mathcal{D}^{\otimes(n+1)}} [\Lambda(\hat{C}_{n,\alpha}(X_{n+1}))] \right\}$ represents the smallest possible average measure of any predictive set achieving $1 - c\alpha\delta$ marginal validity on the distribution $\mathcal{D}^{\otimes(n+1)}$.

In other words and simplifying, Theorem 3.3.2 shows that an estimator achieving $(1 - \alpha, \delta)$ - X -conditional validity can not be more efficient than an estimator achieving distribution-free marginal validity at the level $1 - \alpha\delta$. However, in practice we are interested by the case where δ is small, leading to marginally valid estimators at the level $1 - \alpha\delta$ that are particularly inefficient, therefore the same would be true for $(1 - \alpha, \delta)$ - X -conditionally valid ones. This calls for further relaxation of X -conditional validity.

3.3.3 Y -conditional validity

Another form of conditional validity that might be desired in practice is to be valid conditional on Y . Indeed, one might want to cover at the same level whether the electricity

price is low or high for example. In classification, this is achievable for SCP (Vovk, 2012) by comparing the score on a given $y \in \mathcal{Y}$ only with calibration scores obtained by data points with the same label. This is described more formally in Algorithm 4. While this approach achieve Y -conditional validity, observe that it comes at the cost of smaller calibration sets. We have not touched upon this point until now, and will do so in the following Section 3.3.4, but we can already state that the smaller the calibration set, the higher the variance of our empirical quantile of the scores. For instance, this is all the most true in Algorithm 4 if there is important class imbalance in our data set and a class is unfrequent. To overcome this limitation, a very recent work (Ding et al., 2023) proposed to instead obtain cluster-conditional coverage, after having clustered the calibration data (therefore, an additional split is required to learn a mapping between the labels and the clusters).

3.3.4 Impact of the calibration set on the coverage

Let us now focus on the effect of the calibration set randomness in the coverage of the SCP predictive sets. As mentioned, SCP guarantee is conditional on the proper training set but marginalized over the calibration random variables. Vovk (2012) show that we can obtain a coverage guarantee after conditioning on the calibration set. It relies on deriving instead a probability approximately correct bound. We state one of the results in Theorem 3.3.3.

Theorem 3.3.3 (calibration conditional validity of SCP).

SCP outputs $\hat{C}_{n,\alpha}$ such that for any distribution \mathcal{D} and any $0 < \delta \leq 0.5$:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}} \left(\mathbb{P}_{\mathcal{D}} \left(Y_{n+1} \notin \hat{C}_{n,\alpha}(X_{n+1}) \mid (X_i, Y_i)_{i=1}^n \right) \leq \alpha + \sqrt{\frac{\log(1/\delta)}{2\#\text{Cal}}} \right) \geq 1 - \delta.$$

To state it differently, the bound of Theorem 3.3.3 controls the deviation of miscoverage with respect to the nominal level α of a predictive set built on a given calibration set. In particular, this deviation vanished with high probability when $\#\text{Cal}$ increases. We refer the interested reader to Vovk (2012) for a complete proof and a tighter

Algorithm 4 SCP in classification with Y -conditional coverage

Input: Learning algorithm \mathcal{A} , conformity score function s , miscoverage rate α , training set $(X_i, Y_i)_{i=1}^n$

Output: Prediction set $\hat{C}_{n,\alpha}$

- 1: Randomly split the training data $(X_i, Y_i)_{i=1}^n$ into a **proper training set** (size $\#\text{Tr}$) and a **calibration set** (size $\#\text{Cal}$)
- 2: Get \hat{A} (by training \mathcal{A} on the **proper training set** $(X_i, Y_i)_{i \in \text{Tr}}$)
- 3: **for** any candidate $y \in \mathcal{Y}$ **do**
- 4: On the **calibration set**, obtain a set of $\#\text{Cal}_y + 1$ **conformity scores** :

$$\mathcal{S}_y = \{S_i = \mathbf{s}(X_i, y; \hat{A}), i \in \text{Cal} \text{ such that } Y_i = y\} \cup \{+\infty\}$$

5: **end for**

- 6: For a new point X_{n+1} , return $\hat{C}_{n,\alpha}(X_{n+1}) \left\{ y \text{ such that } \mathbf{s}(X_{n+1}, y; \hat{A}) \leq q_{1-\alpha}(\mathcal{S}_y) \right\}$.
-

bound. The proof relies on the observation that $\mathbb{P}_{\mathcal{D}} \left(Y_{n+1} \notin \hat{C}_{n,\alpha}(X_{n+1}) \mid (X_i, Y_i)_{i=1}^n \right) \sim \text{Beta}(\lceil (1-\alpha)(\#\text{Cal} + 1) \rceil, \#\text{Cal} + 1 - \lceil (1-\alpha)(\#\text{Cal} + 1) \rceil)$ whose variance is approximately $\frac{\alpha(1-\alpha)}{\#\text{Cal}+2}$. Overall, these results give precise tools to analyse the influence of (the size of) the calibration set on the predictive coverage. If fitting a regression or classification model requires more data point than estimating a univariate quantity such as the $1-\alpha$ quantile of the scores' distribution, the variance induced by a small calibration should still be kept at a small enough level in order to output reliable predictive sets. Indeed, we do not want our predictions to greatly vary if we re-run the procedure on other i.i.d. data. Hence, there is a trade-off between proper training set (higher model accuracy induces efficient predictive sets) and calibration set (variability of the predictive sets), which depends on the target miscoverage level α . This is critically data and machine learning model dependent, but as an educated rule of thumb, in non-pathological scenarii, keeping between 30% and 10% of the training data for calibration has demonstrated to be a good compromise (Sesia and Candès, 2020, which studies extensively CQR and other related methods).

3.4 Avoiding data splitting: full CP and out-of-bags approaches

Therefore, splitting the training set might not be possible or desirable in practice. Again, to rephrase, when n is significantly small, one can not afford to throw away some observations and reduce the actual training size supplied to the learning algorithm \mathcal{A} . Generally, keeping some fresh training data apart for calibration lowers the statistical efficiency (i.e. \hat{A} gets poorer accuracy, leading to larger predictive sets) and increases the statistical variability. However, having access to calibration pointd that are exchangeable with the test point was key to SCP theory as it allowed the method to treat the test point as if part of calibration data. The goal of this section is to see if, and how, we can avoid data splitting or at least alleviate the impact of splitting.

3.4.1 Full Conformal Prediction

Failure of naive approach. A naive idea to avoid data splitting would be to keep all of training point to fit \mathcal{A} . Then, we could evaluate conformity scores on the exact same point and obtain a $1-\alpha$ empirical quantile of these score. Finally, a predictive set could be the set of all the y achieving a smaller score on the test features than this $1-\alpha$ empirical quantile. More formally:

1. Get \hat{A} by training the algorithm \mathcal{A} on $(X_i, Y_i)_{i=1}^n$.
2. Get the empirical quantile $q_{1-\alpha}(\mathcal{S})$ of the set of scores $\mathcal{S} = \left\{ \mathbf{s} \left(X_i, Y_i; \hat{A} \right) \right\}_{i=1}^n \cup \{\infty\}$.
3. Output the set $\left\{ y \text{ such that } \mathbf{s} \left(X_{n+1}, y; \hat{A} \right) \leq q_{1-\alpha}(\mathcal{S}) \right\}$.

However, \hat{A} has been obtained using the training set $(X_i, Y_i)_{i=1}^n$ but did not use X_{n+1} . Therefore we are comparing a *test* score to *train* scores. Thus $\mathbf{s} \left(X_{n+1}, y; \hat{A} \right)$ typically

stochastically dominates any element of $\left\{ \left(\mathbf{s} \left(X_i, Y_i; \hat{A} \right) \right)_{i=1}^n \right\}$. This in turn implies that such a set will typically under cover in practice, and can not enjoy any form of theoretical validity: they lost the backbone of SCP, as the scores are not exchangeable anymore.

In order to recover validity, we have to compare a score on y that is comparable to train scores. Full CP (Vovk et al., 2005) achieves this by retraining \mathcal{A} for any possible y as the value of Y_{n+1} . By doing so, the score on each test y is a train score, and when checking whether it is smaller to the empirical quantile of other training scores, we should be able to invoke the quantile Lemma 3.2.1 as the training data and the test data have (supposedly) been treated equally. A rigorous description of Full CP is given in Algorithm 5.

To state the theoretical validity of Full CP, we have to consider an additional assumption on the learning algorithm \mathcal{A} . Indeed, when describing with words Full CP, we justified the procedure by explaining that the scores are now exchangeable as all the data points have been treated equally. However, this is not always true: if the algorithm \mathcal{A} ignores the last element of its input data set, then having re-trained by including the candidate y has no influence and the score on this candidate still stochastically dominate the true training score. To ensure that exchangeability is preserved, we consider only algorithms \mathcal{A} that are invariant to permutation of their input. This is formally described in Definition 3.4.1, for both deterministic and stochastic \mathcal{A} .

Definition 3.4.1 (symmetrical algorithm).

► A *deterministic* learning algorithm \mathcal{A} is symmetric if for any data set $(X_i, Y_i)_{i=1}^n$, for any permutation σ on $\llbracket 1, n \rrbracket$:

$$\mathcal{A}((X_i, Y_i)_{i=1}^n) \stackrel{\text{a.s.}}{=} \mathcal{A}((X_{\sigma(i)}, Y_{\sigma(i)})_{i=1}^n).$$

► A *stochastic* learning algorithm \mathcal{A} is symmetric if for any data set $(X_i, Y_i)_{i=1}^n$, for any permutation σ on $\llbracket 1, n \rrbracket$, there exists a coupling that maps $\xi \sim \mathcal{U}([0, 1])$ to $\xi'_\sigma \sim \mathcal{U}([0, 1])$, which depends only on σ , such that^a, for a.s. $(X_i, Y_i)_{i=1}^n$:

$$\mathcal{A}((X_i, Y_i)_{i=1}^n; \xi) = \mathcal{A}((X_{\sigma(i)}, Y_{\sigma(i)})_{i=1}^n; \xi'_\sigma).$$

^aThis is the definition provided in Kim and Barber (2023).

Algorithm 5 Full CP

Input: Learning algorithm \mathcal{A} , conformity score function s , miscoverage rate α , training set $(X_i, Y_i)_{i=1}^n$, test point X_{n+1}

Output: Prediction set $\hat{C}_{n,\alpha}$

- 1: **for** any candidate $y \in \mathcal{Y}$ **do**
- 2: Get \hat{A}_y by training \mathcal{A} on $\{(X_i, Y_i)_{i=1}^n\} \cup \{(X_{n+1}, y)\}$
- 3: Obtain a set of training scores

$$\mathcal{S}_y^{(\text{train})} = \left\{ \left(\mathbf{s} \left(X_i, Y_i; \hat{A}_y \right) \right)_{i=1}^n \right\} \cup \left\{ \mathbf{s} \left(X_{n+1}, y; \hat{A}_y \right) \right\}$$

4: **end for**

5: Output the set $\hat{C}_{n,\alpha} = (X_{n+1}) \left\{ y \text{ such that } \mathbf{s} \left(X_{n+1}, y; \hat{A}_y \right) \leq q_{1-\alpha} \left(\mathcal{S}_y^{(\text{train})} \right) \right\}$.

Theorem 3.4.1 (marginal validity of Full CP).

FCP (Algorithm 5) with a symmetric algorithm \mathcal{A} outputs $\hat{C}_{n,\alpha}$ such that for any distribution \mathcal{D} , for any associated exchangeable joint distribution $\mathcal{D}^{\text{exch}(n+1)} \in \mathcal{D}^{\text{exch}(n+1)}$:

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}} \left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \right) \geq 1 - \alpha.$$

Additionally, if the scores $\{S_i\}_{i=1}^{n+1}$ are almost surely (a.s.) distinct:

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}} \left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \right) \leq 1 - \alpha + \frac{1}{n+1}.$$

Proof. Assume $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, and that \mathcal{A} is symmetric (possibly stochastic). Let σ be a permutation on $\llbracket 1, n+1 \rrbracket$.

$$\begin{aligned} \left(s \left(X_{\sigma(i)}, Y_{\sigma(i)}; \hat{A}_{Y_{n+1}} \right) \right)_{i=1}^{n+1} &:= \left(s \left(X_{\sigma(i)}, Y_{\sigma(i)}; \mathcal{A} \left((X_k, Y_k)_{k=1}^{n+1}; \xi \right) \right) \right)_{i=1}^{n+1} \\ &\stackrel{\text{by symmetry of } \mathcal{A}}{\rightarrow} \left(s \left(X_{\sigma(i)}, Y_{\sigma(i)}; \mathcal{A} \left((X_{\sigma(k)}, Y_{\sigma(k)})_{k=1}^{n+1}; \xi' \right) \right) \right)_{i=1}^{n+1} \\ &\stackrel{\text{by exchangeability and } \xi'_\sigma \perp (X_i, Y_i)_{i=1}^{n+1}}{\rightarrow} \stackrel{d}{=} \left(s \left(X_i, Y_i; \mathcal{A} \left((X_k, Y_k)_{k=1}^{n+1}; \xi'_\sigma \right) \right) \right)_{i=1}^{n+1} \\ &\stackrel{\text{as } \xi'_\sigma \stackrel{d}{=} \xi}{\rightarrow} \left(s \left(X_i, Y_i; \hat{A}_{Y_{n+1}} \right) \right)_{i=1}^{n+1}. \end{aligned}$$

Therefore, the scores are exchangeable and it only remains to apply the quantile Lemma 3.2.1. \square

Remark 3.4.1 (SCP is a particular case of Full CP).

SCP can be seen as a special case of Full CP where Full CP is only applied on the calibration data set, and the learning algorithm \mathcal{A} is independent of its input and always output some function \mathcal{A} that has in fact been trained only on the proper training set (this algorithm is indeed symmetric as it is independent of any component of its input).

Theorem 3.4.1 shows that this cautious treatment of the test point allows to retrieve validity without having to split the training data set. However, this comes with the need to fit numerous models. When \mathcal{Y} is not discrete, this is even impossible to perform exactly and it is usually approximated by binning \mathcal{Y} (Chen et al., 2016, 2018), but even while doing so, or when \mathcal{Y} is discrete, it can be computationally costly if there are many bins or classes, or if the learning algorithm \mathcal{A} has heavy computational load.

Exact computation is feasible in ridge or lasso regression (Noureddinov et al., 2001; Burnaev and Vovk, 2014; Lei, 2019), nearest neighbors or kernel smoothing algorithms (Cherubin et al., 2021), and approximations can be achieved under smooth and “regular” (such as convex) regression estimators (Ndiaye and Takeuchi, 2019) or algorithms satisfying (prediction) stability assumptions (Ndiaye, 2022), or when the predictive set of Full CP is in fact an interval (Ndiaye and Takeuchi, 2022).

Example 3.4.1 (standard FCP sets with an interpolating algorithm).

Assume \mathcal{A} interpolates. Then, for any candidate $y \in \mathcal{Y}$, \hat{A}_y is such that:

- *Regression*: $\hat{A}_y(X_i) = Y_i$ for $i \in \llbracket 1, n \rrbracket$ and $\hat{A}_y(X_{n+1}) = y$;
- *Classification*: $\hat{A}_y(X_i)_{Y_i} = 1$ for $i \in \llbracket 1, n \rrbracket$ and $\hat{A}_y(X_{n+1})_y = 1$

Henceforth, Full CP (with standard score functions) with an interpolating algorithm outputs \mathcal{Y} for any new test point.

Note that in this case all the scores are almost surely equals. As such this example does not contradict the upper bound of Theorem 3.4.1.

3.4.2 Jackknife+ and leave-one-out CP

A natural question that arises now is whether there exist theoretically valid intermediate methods between SCP and Full CP. An idea is to leverage leave-one-out strategies, in order to use all of the training data (unlike SCP) but only have n model fits (which often is smaller than for FCP). The first natural idea based on leave-one-out is to fit n model, leaving out a different training point to fit each model, and obtain a conformity score on the left out point. Then, the $1 - \alpha$ empirical quantile of these scores is computed and used to correct the prediction made on the test point by a model fitted this time on the n training points. This is formalized below.



1. For any $j \in \llbracket 1, n \rrbracket$: get \hat{A}_{-j} by training \mathcal{A} on $(X_i, Y_i)_{\substack{i=1 \\ i \neq j}}^n$.
2. Get the empirical quantile $q_{1-\alpha}(\mathcal{S})$ of the set of scores

$$\mathcal{S} = \{S_i = \mathbf{s}(X_i, Y_i; \hat{A}_{-i}), i \in \llbracket 1, n \rrbracket\} \cup \{+\infty\}.$$

3. Get \hat{A} by training \mathcal{A} on $(X_i, Y_i)_{i=1}^n$.
4. Output the set $\hat{C}_{n,\alpha}(X_{n+1}) = \{y \in \mathcal{Y} \text{ such that } \mathbf{s}(X_{n+1}, y; \hat{A}) \leq q_{1-\alpha}(\mathcal{S})\}.$

However, without stability assumptions on \mathcal{A} , there is absolutely no guarantee on the prediction of \hat{A} with scores based on $(\hat{A}_{-i})_i$ (Barber et al., 2021b). Indeed, this naive algorithm is comparing a score on the test point obtained through an algorithm that has seen n points, while the reference “calibration” scores rely on learning on $n - 1$ data points. To circumvent this issue, Barber et al. (2021b) introduce the Jackknife+ algorithm that treats the training points and the test point similarly: the idea is, that for each $i \in \llbracket 1, n \rrbracket$, the algorithm learns a model leaving the i -th point out to evaluate conformity on it, while also assessing the conformity of potential test points with this fitted model. Jackknife+ is written only for mean-regression and scores that are the absolute value of the residuals, but

Gupta et al. (2022) have shown that a tighter leave-one-out set can be built in a general setting. The core idea is exactly the same, hence we present here only the generalized and tighter version formalized in Gupta et al. (2022) but in terms of conformity scores, in Algorithm 6 (we recall that Gupta et al., 2022, work in their novel nested sets framework).

Again, the predictive set is built by looping over all possible $y \in \mathcal{Y}$ which can be tricky in practice. We refer the reader to Gupta et al. (2022) for an efficient implementation (linear time in n) of this algorithm, when each of the $\mathbb{1}\left\{\mathbf{s}\left(X_i, Y_i; \hat{A}_{-i}\right) < \mathbf{s}\left(X_{n+1}, y; \hat{A}_{-i}\right)\right\}$ takes value 1 only on an interval. In this case, it is possible to derive a Jackknife+ version of the algorithm, whose predictive sets include the ones of leave-one-out CP. This is a generalization of Jackknife+, which was written only for mean-regression and absolute value of the residuals scores, suggested again in Gupta et al. (2022). We rephrase it in terms of conformity scores in Algorithm 7 (recall from Definition 3.1.7 that $q_{\beta, \inf}(U_1, \dots, U_n) := \lfloor \beta \times n \rfloor$ smallest value of (U_1, \dots, U_n)). Theorem 3.4.2 specifies the theoretical guarantees that this algorithm obtain.

Theorem 3.4.2 (marginal validity of leave-one-out-CP and JK+).

Algorithms 6 and 7 with a symmetric algorithm \mathcal{A} output $\hat{C}_{n, \alpha}$ such that for any distribution \mathcal{D} , for any associated exchangeable joint distribution $\mathcal{D}^{\varepsilon(n+1)} \in \mathcal{D}^{\text{exch}(n+1)}$:

$$\mathbb{P}_{\mathcal{D}^{\varepsilon(n+1)}}\left(Y_{n+1} \in \hat{C}_{n, \alpha}(X_{n+1})\right) \geq 1 - 2\alpha.$$

Proof. We prove the result for Algorithm 6, as its predictive sets are included in the ones of Algorithm 7 (when those are well-defined).

Algorithm 6 Leave-one-out CP

Input: Learning algorithm \mathcal{A} , conformity score function s , miscoverage rate α , training set $(X_i, Y_i)_{i=1}^n$, test point X_{n+1}

Output: Prediction set $\hat{C}_{n, \alpha}$



- 1: **for** $j \in \llbracket 1, n \rrbracket$ **do**
- 2: Get \hat{A}_{-j} by training \mathcal{A} on $(X_i, Y_i)_{\substack{i=1 \\ i \neq j}}^n$
- 3: **end for**
- 4: For a new point X_{n+1} , return

$$\hat{C}_{n, \alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \sum_{i=1}^n \mathbb{1}\left\{\mathbf{s}\left(X_i, Y_i; \hat{A}_{-i}\right) < \mathbf{s}\left(X_{n+1}, y; \hat{A}_{-i}\right)\right\} < (1 - \alpha)(n + 1) \right\}$$

Algorithm 7 Generalized Jackknife+

Input: Learning algorithm \mathcal{A} , conformity score function s , miscoverage rate α , training set $(X_i, Y_i)_{i=1}^n$, test point X_{n+1}

Output: Prediction set $\hat{C}_{n,\alpha}(X_{n+1})$



- 1: **for** $j \in \llbracket 1, n \rrbracket$ **do**
- 2: Get \hat{A}_{-j} by training \mathcal{A} on $(X_i, Y_i)_{i=1, i \neq j}^n$
- 3: For a new point X_{n+1} , build

$$[\ell_{-j,\alpha}(X_{n+1}); u_{-j,\alpha}(X_{n+1})] := \left\{ y \in \mathcal{Y} : \mathbf{s} \left(X_{n+1}, y; \hat{A}_{-j} \right) \leq \mathbf{s} \left(X_j, Y_j; \hat{A}_{-j} \right) \right\}$$

- 4: **end for**
- 5: **Return**

$$\hat{C}_{n,\alpha}(X_{n+1}) = [q_{\alpha, \inf}((\ell_{-i,\alpha}(X_{n+1}))_{i=1}^n \cup \{-\infty\}); q_{1-\alpha}((u_{-i,\alpha}(X_{n+1}))_{i=1}^n \cup \{+\infty\})]$$

Step 1. Remark that:

$$\begin{aligned} & \left\{ Y_{n+1} \notin \hat{C}_{n,\alpha}(X_{n+1}) \right\} \\ &= \left\{ \sum_{j=1}^n \mathbb{1} \left\{ s(X_j, Y_j; \hat{A}_{-j}) < s(X_{n+1}, Y_{n+1}; \hat{A}_{-j}) \right\} \geq (1-\alpha)(n+1) \right\} \\ &:= \left\{ \sum_{j=1}^n \mathbb{1} \left\{ S^{(j),n+1} < S^{(n+1),j} \right\} \geq (1-\alpha)(n+1) \right\} \\ &:= \left\{ \sum_{j=1}^n \mathcal{C}_{n+1,j} \geq (1-\alpha)(n+1) \right\}. \end{aligned}$$

with $S^{(i),j} := s(X^{(i)}, Y^{(i)}; \hat{A}_{-(i,j)})$ is the score on data point i of the predictor that has been fitted without seeing nor data point i nor data point j , for $(i, j) \in \llbracket 1, n+1 \rrbracket^2$ and extending \hat{A}_{-i} to $\hat{A}_{-(i,j)} := \mathcal{A} \left((X_k, Y_k)_{k=1, k \notin \{i,j\}}^{n+1} \right)$, where the $n+1$ data point is added.

Denote by $\mathcal{C}_{\mathcal{A}}$ the function building the comparison matrix $\mathcal{C} \in \{0, 1\}^{(n+1) \times (n+1)}$:

$$\mathcal{C}_{\mathcal{A}} \left((X_k, Y_k)_{k=1}^{n+1} \right)_{i,j} = \mathbb{1} \left\{ S^{(i),j} > S^{(j),i} \right\} = \mathcal{C}_{i,j}.$$

Step 2. Deterministically, [Barber et al. \(2021b\)](#) shows that $\#\{i \in \llbracket 1, n+1 \rrbracket : \sum_{j=1}^{n+1} \mathcal{C}_{i,j} \geq (1-\alpha)(n+1)\} \leq 2\alpha(n+1)$. This is shown for *any* comparison matrix.

Step 3. The last (and crucial) step of leave-one-out conformal predictors is to show that thanks to the algorithm symmetry and data exchangeability, for any permutation σ on $\llbracket 1, n+1 \rrbracket$ it holds: $(\mathcal{C}_{\sigma(i),\sigma(j)})_{i,j} \stackrel{d}{=} (\mathcal{C}_{i,j})_{i,j}$.

Consider the general case where \mathcal{A} is a randomized algorithm and let σ a permutation on $\llbracket 1, n+1 \rrbracket$, and $(i, j) \in \llbracket 1, n+1 \rrbracket^2$.

$$\begin{aligned}
\mathcal{C}_{\sigma(i),\sigma(j)} &= \mathcal{C}_{\mathcal{A}} \left((X_j, Y_j)_{k=1}^{n+1} \right)_{\sigma(i),\sigma(j)} \\
&= \mathbb{1} \left\{ s \left(Y_{\sigma(i)}, X_{\sigma(i)}, \mathcal{A} \left((X_j, Y_j)_{k=1, k \notin \{\sigma(i), \sigma(j)\}}^{n+1} ; \xi \right) \right) \right. \\
&\quad \left. > s \left(Y_{\sigma(j)}, X_{\sigma(j)}, \mathcal{A} \left((X_j, Y_j)_{k=1, k \notin \{\sigma(i), \sigma(j)\}}^{n+1} ; \xi \right) \right) \right\} \\
\mathcal{A} \text{ is symmetric} \rightarrow &= \mathbb{1} \left\{ s \left(Y_{\sigma(i)}, X_{\sigma(i)}, \mathcal{A} \left((X_{\sigma(k)}, Y_{\sigma(k)})_{k=1, k \notin \{i, j\}}^{n+1} ; \xi'_{\sigma} \right) \right) \right. \\
&\quad \left. > s \left(Y_{\sigma(j)}, X_{\sigma(j)}, \mathcal{A} \left((X_{\sigma(k)}, Y_{\sigma(k)})_{k=1, k \notin \{i, j\}}^{n+1} ; \xi'_{\sigma} \right) \right) \right\} \\
\mathcal{C}_{\sigma(i),\sigma(j)} &= \mathcal{C}_{\mathcal{A}} \left((X_{\sigma(k)}, Y_{\sigma(k)})_{k=1}^{n+1} \right)_{i,j}
\end{aligned}$$

This holds for any $(i, j) \in \llbracket 1, n+1 \rrbracket^2$, hence, denoting Π_{σ} the matrix permutation associated with σ (i.e. $(\Pi_{\sigma}^T \mathcal{C} \Pi_{\sigma})_{i,j} = \mathcal{C}_{\sigma(i),\sigma(j)}$ for any $(i, j) \in \llbracket 1, n+1 \rrbracket^2$):

$$\begin{aligned}
\Pi_{\sigma}^T \mathcal{C} \Pi_{\sigma} &= \mathcal{C}_{\mathcal{A}} \left((X_{\sigma(k)}, Y_{\sigma(k)})_{k=1}^{n+1} \right) \\
\xi'_{\sigma} \perp (X_j, Y_j)_{k=1}^{n+1} &\rightarrow \stackrel{d}{=} \mathcal{C}_{\mathcal{A}} \left((X_k, Y_k)_{k=1}^{n+1} \right) = \mathcal{C} \\
&\text{and exchangeability}
\end{aligned}$$

This concludes the proof as therefore each element of $\llbracket 1, n+1 \rrbracket$ is equally likely to belong to $\{i \in \llbracket 1, n+1 \rrbracket : \sum_{j=1}^{n+1} \mathcal{C}_{i,j} \geq (1-\alpha)(n+1)\}$. \square

Remark 3.4.2 (on the lost factor 2).

The theoretical guarantee of leave-one-out-CP and JK+ presents a loss of coverage: the lower bound on the coverage is now in $1 - 2\alpha$. Empirically, it achieves approximately $1 - \alpha$ coverage, a bound also obtained in theory under algorithmic stability assumptions. However, this factor 2 is not an artefact of the proof, and [Barber et al. \(2021b\)](#) derive an example in which the lower bound is attained (asymptotically with n). This example relies on a highly non-stable learning algorithm due to its intrinsic design as well as due to the data distribution on which it is applied. In other words, to suffer from this loss of coverage, the combination of data distribution and algorithm should provoke important prediction instability. In particular, it should be the case that some of the $(\hat{A}_{-i})_{i=1}^n$, say for $i \in \text{bad}$, would be associated to higher scores than the rest of the models (i.e. for $i \notin \text{bad}$), but that between all the $i \in \text{bad}$ there is no clear ranking between the i .

3.4.3 CV+

For cases where n is already too large, an analogous of the corrected leave-one-out predictive sets can be defined for k -fold cross-validated scheme. The idea traces back to [Vovk \(2015\)](#), but we present here the version generalized by [Gupta et al. \(2022\)](#) from the suggested CV+ algorithm of [Barber et al. \(2021b\)](#). As in the previous subsection, we rephrase it in terms of conformity scores in Algorithm 8. We provide its theoretical guarantees in Theorem 3.4.3.

Algorithm 8 K -fold CP

Input: Learning algorithm \mathcal{A} , conformity score function s , miscoverage rate α , number of fold $K \in \mathbb{N}^*$, training set $(X_i, Y_i)_{i=1}^n$, test point X_{n+1}

Output: Prediction set $\hat{C}_{n,\alpha}$



- 1: Randomly split $(X_i, Y_i)_{i=1}^n$ into K folds F_1, \dots, F_K (we denote $k(i)$ the subset that includes i)
- 2: **for** $k \in \llbracket 1, K \rrbracket$ **do**
- 3: Get \hat{A}_{-k} by training \mathcal{A} on $(X_i, Y_i)_{k(i) \neq k}$
- 4: **end for**
- 5: For a new point X_{n+1} , return

$$\hat{C}_{n,\alpha}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \sum_{i=1}^n \mathbb{1} \left\{ s \left(X_i, Y_i; \hat{A}_{-k(i)} \right) < s \left(X_{n+1}, y; \hat{A}_{-k(i)} \right) \right\} < (1 - \alpha)(n + 1) \right\}$$

Theorem 3.4.3 (marginal validity of K -fold CP CV+).

Algorithm 8 with $K \in \mathbb{N}^*$ folds and with a symmetric algorithm \mathcal{A} outputs $\hat{C}_{n,\alpha}$ such that for any distribution \mathcal{D} , for any associated exchangeable joint distribution $\mathcal{D}^{\mathcal{E}(n+1)} \in \mathcal{D}^{\text{exch}(n+1)}$:

$$\begin{aligned} \mathbb{P}_{\mathcal{D}^{\mathcal{E}(n+1)}} \left(Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1}) \right) &\geq 1 - 2\alpha - \min \left(\frac{2(1 - 1/K)}{n/K + 1}, \frac{1 - K/n}{K + 1} \right) \\ &\geq 1 - 2\alpha - \sqrt{2/n}. \end{aligned}$$

In summary (Figure 3.8), there is a vast range of methods going from no splitting to a single split, passing through K -fold/CV+ approaches, enjoying finite sample distribution free marginal validity with any (symmetric) algorithm. While distribution-free X -conditional validity can not be attained by any of these methods, distribution-free Y -conditional

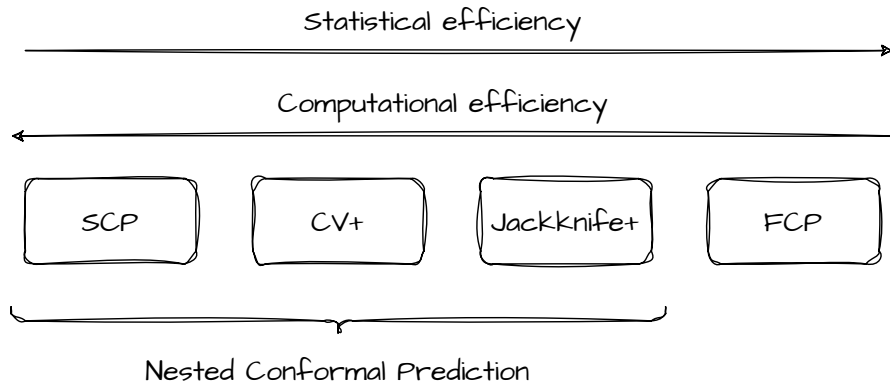


Figure 3.8: Range of CP frameworks in the spectrum of splitting strategies.

coverage is achievable at least in theory, and, finally, training-conditional coverage is obtained through PAC bounds for SCP (for which training-conditional coverage refers to calibration-conditional), but also K -fold CP/CV+ whose bound emphasizes that the controlling quantity is n/K which should be large, however Full CP and leave-one-out CP/Jackknife+ do not benefit from any training-conditional coverage unless stability assumptions are made on the learning algorithm \mathcal{A} (Vovk, 2012; Bian and Barber, 2023; Liang and Barber, 2023).

3.5 Beyond exchangeability

The last issue that we consider in this introductory chapter is how to extend CP to non-exchangeable settings? This is particularly challenging as it the one and only assumption required by conformal. Yet, it is an important direction to explore as exchangeability does not hold in many practical applications. Indeed, it can be broken by:

- Shifts between the training data and the test data, and in particular:
 - i) Covariate shift, i.e. $\mathcal{D}_X^{(\text{train})} \neq \mathcal{D}_X^{(\text{test})}$ while $\mathcal{D}_{Y|X}^{(\text{train})} = \mathcal{D}_{Y|X}^{(\text{test})}$;
 - ii) Label shift, i.e. $\mathcal{D}_Y^{(\text{train})} \neq \mathcal{D}_Y^{(\text{test})}$ while $\mathcal{D}_{X|Y}^{(\text{train})} = \mathcal{D}_{X|Y}^{(\text{test})}$;
 - iii) Arbitrary distribution shift on both the label and the covariates;
- Possibly many shifts, not only one, not necessarily a finite number;
- Temporal dependence, distributional drifts and non-stationarity.

This line of research has been especially active in the recent years. In this section, we focus on the main common ideas giving only some reference points that should allow an interested reader to navigate the overall literature more easily afterwards.

Under additional assumptions on the data distribution, such as strongly mixing noise, or on the quality of the fitted model that should be close to the generative model, theoretical results can be obtained in the data dependent context (see, e.g., Chernozhukov et al., 2018). Otherwise, there are two main settings: one in which we can rely on weighting strategy with a priori knowledge or estimation, and one in which we use feedback on the fly to understand (with a delay) how to adapt the predictive set estimator. The underlying assumption in all these methods is that even though data is not exchangeable anymore, there is some information from the historical data that we can leverage cautiously to build enhanced (in comparison with subsampling the historical data set) yet robust and not corrupted predictive sets.

3.5.1 Weighting strategies

The idea of weighting approaches is to assign more importance to the data points that we trust more or are closer in distribution to the test point. Until now, we have formalized CP as evaluating an empirical quantile of scores $\{q_{1-\alpha}(\{(S_i)_{i=1}^n\} \cup \{+\infty\}) := \lceil (1-\alpha)(n+1) \rceil$ smallest value of $\{(S_i)_{i=1}^n\} \cup \{+\infty\}$. In order to introduce weighting strategies, it is useful

to note that in fact this is equivalent to considering $Q_{\mathcal{D}_S}(1 - \alpha)$, with \mathcal{D}_S the empirical distribution of the scores, i.e. $\mathcal{D}_S := \frac{1}{n+1} \sum_{i=1}^n \delta_{S_i} + \frac{1}{n+1} \delta_{+\infty}$.

Tibshirani et al. (2019) introduced first the concept of weighted exchangeability (we refer the interested reader to the original paper for details) justifying weighted CP. They consider a setting in which the training data is drawn i.i.d. from some distribution \mathcal{D} , $(X_i, Y_i)_{i=1}^n \sim (\mathcal{D}_X \times \mathcal{D}_{Y|X})^{\otimes(n)}$, and we aim at predicting Y_{n+1} observing X_{n+1} , with $(X_{n+1}, Y_{n+1}) \sim \tilde{\mathcal{D}}_X \times \mathcal{D}_{Y|X}$ for some distribution $\tilde{\mathcal{D}}_X \neq \mathcal{D}_X$. The key idea is that if we know the ratio $\frac{d\tilde{\mathcal{D}}_X(x)}{d\mathcal{D}_X(x)} := w(x)$, then the normalized/probability weights defined by:

$$\begin{cases} \text{for any } i \in \llbracket 1, n \rrbracket, \omega_i(x) := \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)} \\ \omega_{n+1}(x) := \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}, \end{cases}$$

ensure that the data points are weighted exchangeable. Therefore, outputting the set

$$\hat{C}_{n,\alpha}(X_{n+1}) := \{y \in \mathcal{Y} : s(X_{n+1}, Y_{n+1}; \mathcal{A}((X_i, Y_i)_{i=1}^n)) \leq Q_{\Omega_S}(1 - \alpha)\},$$

with $\Omega_S := \sum_{i=1}^n \omega_i(X_{n+1}) \delta_{S_i} + \omega_{n+1}(X_{n+1}) \delta_{+\infty}$, is a marginally valid procedure.

Similarly, Podkopaev and Ramdas (2021) suggest to use this idea in situation where there is a label shift. Precisely, suppose again that the training data is drawn i.i.d. from some distribution \mathcal{D} , $(X_i, Y_i)_{i=1}^n \sim (\mathcal{D}_{X|Y} \times \mathcal{D}_Y)^{\otimes(n)}$, and we aim at *classifying* Y_{n+1} observing X_{n+1} , with $(X_{n+1}, Y_{n+1}) \sim \mathcal{D}_{X|Y} \times \tilde{\mathcal{D}}_Y$ for some distribution $\tilde{\mathcal{D}}_Y \neq \mathcal{D}_Y$. The challenge here is that the actual test label is unknown, unlike the test features X_{n+1} . However, in classification we can loop over all possible classes. Therefore, based on the ratio $w(y) := \frac{d\tilde{\mathcal{D}}_Y(y)}{d\mathcal{D}_Y(y)}$, one can constructs normalized/probability weights for each possible class $y \in \mathcal{Y}$:

$$\begin{cases} \text{for any } i \in \llbracket 1, n \rrbracket, \omega_i(y) = \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)} \\ \omega_{n+1}(y) := \frac{w(y)}{\sum_{j=1}^n w(Y_j) + w(y)}, \end{cases}$$

for any $i \in \llbracket 1, n \rrbracket$. Then, the predictive set is

$$\hat{C}_{n,\alpha}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y; \mathcal{A}((X_i, Y_i)_{i=1}^n)) \leq Q_{\Omega_S^y}(1 - \alpha)\}$$

with $\Omega_S^y := \sum_{i=1}^n \omega_i(y) \delta_{S_i} + \omega_{n+1}(y) \delta_{+\infty}$, is a marginally valid procedure.

In practice, both these likelihood ratio (for covariate and for label shifts) have to be estimated and the guarantees do not go through directly, even though improved empirical performances are obtained. Jin and Candès (2023) provide some theory on the loss of coverage incurred by an estimation.

Similar reweighting approaches have been further developped in the context of causal inference (Lei and Candès, 2021; Gui et al., 2023b), survival analysis (Candès et al., 2023) and active learning (Fannjiang et al., 2022).

What if the rupture point was unknown, the estimation of the likelihood ratio is not possible, or the data distribution is slightly drifting?

If we still can assume an access to some i.i.d. data points, but do not want to suppose estimation of the likelihood ratio is possible (possibly because different shifts are in fact plausible), it is possible to leverage tools from the distributionally robust optimization literature. In particular, [Cauchois et al. \(2024\)](#) provide predictive sets that are valid for any distribution shift (both on Y and X) as long as the shift remains bounded in f -divergence (e.g., Kullback-Leibler divergence) with respect to the train distribution.

If instead we cannot assume i.i.d. data points even in the training set, [Barber et al. \(2023\)](#) proposes to apply weights $(w_i)_{i=1}^n$ pre-defined by the user to each data point, relying on the same weighted quantile function than in [Tibshirani et al. \(2019\)](#); [Podkopaev and Ramdas \(2021\)](#). For example, in a time series context, one could apply exponential weights decaying in time (oldest points would receive lower weights) at a speed depending on the memory we consider representative. Importantly though, these weights can not be chosen in a way that depends on $(X_i, Y_i)_{i=1}^{n+1}$. The main theoretical result provided in the paper bounds the coverage loss due to the violation of exchangeability in the data set. Particularly, denoting again $(\omega_i)_{i=1}^{n+1}$ the normalized weights associated to the chosen weights $(w_i)_{i=1}^n$, their main result proves the following control on the coverage loss, which we state informally.

Informal theorem

Running weighted-CP with data independent normalized weights $(\omega_i)_{i=1}^{n+1}$ achieves:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1})\right) \geq 1 - \alpha - \sum_{i=1}^n \omega_i TV\left(\mathcal{S}, \mathcal{S}^{(i)}\right),$$

where $\mathcal{S}^{(i)} := (S_1, \dots, S_{i-1}, S_{n+1}, S_{i+1}, \dots, S_i)$, i.e. the set of scores when the test score S_{n+1} and the i -th score S_i have been swapped.

This result highlights that if we can choose the weights adequately, then coverage can be recovered. Maybe most importantly, it also sheds lights on the standard CP framework under violation of exchangeability. Indeed, taking uniform weights we recover the standard CP setting, and the result provides a characterization on the coverage deterioration depending on the strength of violation of exchangeability.

3.5.2 Online setting

Generalizing the time series framework, let us consider now that we have access to an initial data stream, $(X_t, Y_t)_{t=1}^{T_0}$, and that we aim at building predictive sets $\widehat{C}_{t,\alpha}$ for some T_1 subsequent observations. Our goal is that the predictive sets sequence enjoy theoretical guarantees without making any assumption on the data stream. This is highly challenging as it includes adversarial sequences. However, in this setting, we assume that at any prediction step $t \in \llbracket T_0 + 1, T_0 + T_1 \rrbracket$, $Y_{t-T_0}, \dots, Y_{t-1}$ have been revealed³. For example, this typically represents electricity prices forecasting where we have access to historical data on which to fit a model, and when predicting sequentially the next prices, any previous outcomes have already been revealed.

³This setting can be generalized to encapsulate forecast horizons $h > 1$.

In this setting, our ideal goal remains to control the probability of coverage with respect to the data distribution, that is building the smallest predictive set such that:

$$\mathbb{P}\left(Y_t \in \widehat{C}_{t,\alpha}(X_t)\right) \geq 1 - \alpha, \text{ for } t \in \llbracket T_0 + 1, T_0 + T_1 \rrbracket.$$

However, when considering any data stream without restrictions, including adversarial ones, this goal appears to be lofty. Therefore, in general, we focus on achieving realized frequency type guarantees, averaged over all the sequence, which we write as:

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{Y_t \in \widehat{C}_{t,\alpha}(X_t)\right\} \approx 1 - \alpha,$$

or asymptotically:

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{Y_t \in \widehat{C}_{t,\alpha}(X_t)\right\} \xrightarrow{T_1 \rightarrow +\infty} 1 - \alpha.$$

The key difference here is that the guarantee we target is not in probability anymore, and it allows to use strategies whose theory rely on deterministic arguments. The pioneer work in this framework is that of [Gibbs and Candès \(2021\)](#). They propose Adaptive Conformal Inference (ACI) that adapts iteratively the quantile level of the scores' quantile, depending on the coverages of the previous steps. Precisely, let $\alpha_{T_0+1} = \alpha$ and fix some $\gamma > 0$, which controls the speed of reaction to previous iterates. It can also be understood as playing the role of learning rate in an online gradient descent algorithm with respect to the pinball loss. The update scheme can be written as follows:

$$\begin{cases} \widehat{C}_{t,\alpha_t}(X_t) &= \left\{y \in \mathcal{Y} : s\left(X_t, y; \mathcal{A}\left((X_k, Y_k)_{k=1}^{t-1}\right)\right) \leq q_{1-\alpha_t}(\mathcal{S}_t)\right\} \\ \alpha_{t+1} &= \alpha_t + \gamma\left(\alpha - \mathbb{1}\left\{Y_t \notin \widehat{C}_{t,\alpha_t}(X_t)\right\}\right), \end{cases}$$

where the set of scores is now indexed by the time that passes, \mathcal{S}_t , to incorporate any pipeline such as re-training on the current data stream.

In other words, if ACI does not cover at time t , then $\alpha_{t+1} \leq \alpha_t$, and the size of the predictive set increases; conversely when it covers. Importantly, nothing prevents $\alpha_t \leq 0$ or $\alpha_t \geq 1$. While the later is rare (as α is typically small), the former can happen frequently for some γ , producing by convention $\widehat{C}_{t,\alpha_t} \equiv \mathcal{Y}$.

ACI, and some later extensions of it, enjoy an asymptotic valid frequency *for any data sequence*.

Informal theorem

For any sequence of data, we have with probability one that:

$$\left| \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{Y_t \in \widehat{C}_{t,\alpha_t}(X_t)\right\} - (1 - \alpha) \right| \leq \frac{1 - \alpha + \gamma}{\gamma T}.$$

In particular, it follows that:

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{Y_t \in \widehat{C}_{t,\alpha_t}(X_t)\right\} \xrightarrow[T_1 \rightarrow +\infty]{a.s.} 1 - \alpha.$$

Crucially, this long-term frequency result does not provide guidelines on how to pick γ , or even the contrary as it favors large γ that are associated to more frequent uninformative sets (i.e. outputting \mathcal{Y}) as well as more instability. In Chapter 5, based on Zaffran et al. (2022), we propose deeper theoretical analysis on the influence of γ on the efficiency of the resulting predictive sets. This allows us to provide a practical algorithm, **AgACI**, based on online aggregation based on expert advice, which is parameter-free and does not require to choose γ .

More recent developments include: Gibbs and Candès (2023) improving on ACI by online aggregation on a grid of different γ , similarly to what we propose in Chapter 5 through **AgACI**, at the crucial difference that the aggregation is on the value of α_t and not on the lower and upper bounds independently, which permits to derive theoretical guarantees on the regret of the proposed method; Bastani et al. (2022) which achieves stronger coverage guarantees (conditional on the effective level, and conditional on specified subsets of the explanatory variables); Bhatnagar et al. (2023) enjoying anytime regret bound, by leveraging tools from the strongly adaptive regret minimization literature; Angelopoulos et al. (2023) which extends upon ACI ideas by relying on control theory to add more information on the temporal structure, notably on the scores; Angelopoulos et al. (2024) proposing to use adaptive learning rates γ_t in ACI, and even retrieving asymptotic control in probability when the data points are in fact i.i.d., i.e. $\lim_{T_1 \rightarrow +\infty} \mathbb{P} \left(Y_{T_1} \in \widehat{C}_{T_1, \alpha}(X_{T_1}) \right) \rightarrow 1 - \alpha$. A very recent work (Yang et al., 2024) takes the counterpoint of most of these works by explicitly optimizing for efficiency of the intervals, while preserving long-term coverage.

Chapter 4

Technical Summary of the Contributions

This chapter detail each contributions of this manuscript. While motivated by the task of forecasting electricity prices, the methods developed are generic: they can be applied in *any* sensitive fields.

4.1 Contributions' summary of Part II – Time Series

Chapter 5 detailed summary. Our approach is to illustrate the usefulness of ACI on time series with general dependency and non-stationarity, as it was initially developed for distribution shifts.

We start by studying theoretically, using Markov Chain theory, the impact of γ on the length of the predictive intervals, in order to describe not only the *validity* but also the *efficiency* of ACI. This is critical as the convergence rate of ACI favors large γ , which are associated to frequent uninformative predictive sets. Moreover, ACI is usually applied without knowing the type of data it will encounter. If the scores are actually exchangeable, ACI's *validity* would not improve upon SCP (known to be *valid*), thus assessing ACI's impact on *efficiency* is necessary. Thereby, we first provide an analysis in the exchangeable case, then in the auto-regressive one (time series). Define $L(\alpha_t)$ the length of the interval predicted by ACI at time t (dependence in γ is hidden), and L_0 the length of the interval predicted by the non-adaptive algorithm (or equivalently, $\gamma = 0$).

Theorem 4.1.1. *Assume that: (i) $\alpha \in \mathbb{Q}$; (ii) the scores are exchangeable with quantile function Q ; (iii) the quantile function Q is perfectly estimated at each time; (iv) the quantile function Q is bounded and $\mathcal{C}^4([0, 1])$. Then, for all $\gamma > 0$, $(\alpha_t)_{t \geq T_0}$ forms a Markov Chain, that admits a stationary distribution π_γ , and $\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} L(\alpha_t) \xrightarrow[T_1 \rightarrow +\infty]{a.s.} \mathbb{E}_{\pi_\gamma}[L] \stackrel{not.}{=}$ $\mathbb{E}_{\tilde{\alpha} \sim \pi_\gamma}[L(\tilde{\alpha})]$. Moreover, as $\gamma \rightarrow 0$, $\mathbb{E}_{\pi_\gamma}[L] = L_0 + Q''(1 - \alpha)\frac{\gamma}{2}\alpha(1 - \alpha) + O(\gamma^{3/2})$.*

For standard distributions, $Q''(1 - \alpha) > 0$, and Theorem 4.1.1 implies that ACI on exchangeable scores degrades the efficiency linearly with γ compared to CP: $\gamma = 0$ (standard SCP) is better.

A second theorem along with numerical analysis prove that, if the residuals are autoregressive of coefficient φ (the higher the more important the temporal dependence) and the calibration is perfect, then there exists an optimal $\gamma^* > 0$ minimizing the average length for high φ , and its value depends on the time dependence strength.

These results stress that choosing γ is crucial but its optimal value, with respect to efficiency, depends on the unknown data distribution. Therefore, we design **AgACI**, a parameter-free method using online expert aggregation (Cesa-Bianchi and Lugosi, 2006). Based on the pinball loss of level $1 - \frac{\alpha}{2}$ (resp. $\frac{\alpha}{2}$), **AgACI** assigns weights to each expert (an expert is a version of ACI with some γ) depending on their previous performances in order to output a unique upper bound (resp. lower bound) which is the weighted mean of the experts upper (resp. lower) bounds.

Finally, we compare ACI with various γ , **AgACI** and benchmark methods, on extensive synthetic experiments of increasing temporal dependence and on the task on forecasting French electricity prices in 2019. These experiments highlight that:

- Benchmark methods are not robust to the increase of the temporal dependence;
- ACI is robust to this increase, maintaining validity in all settings with a well-chosen γ ;
- **AgACI** is robust to this increase without choosing γ , at the cost of not being the smallest.

Chapter 6 detailed summary. To go further on the application to electricity prices forecasting, we conduct extensive experiments on a novel data set containing the French electricity spot prices during the turbulent 2020-2021 years. First, we build a new explanatory variable revealing high predictive power, namely the nuclear availability. Then, we benchmark state-of-the-art probabilistic electricity prices forecastings methods, showcasing that picking a model a priori is complex as *i*) they all behave very differently, and *ii*) none of them maintains coverage on the most hazardous period of late 2021. Therefore, we study the performance of operational fixed prediction models that can be made adaptive through a plugged-in layer, useful when facing non-stationarity without completely retraining the underlying model. We consider two post-hoc layers: *i*) online CP through a proposal of novel conformalization that respects the forecast horizon during calibration, coined **OSSCP-horizon**, as well as **AgACI**, and *ii*) online aggregation of individual forecasts. Both approaches enhance the coverage of the resulting predictive intervals, and combining them through the aggregation of various **AgACI** appears to be the best strategy, on this particular task at least. Moreover, analysing this specific aggregation sheds light on many domain phenomena thanks to the aggregation weights interpretability: we are able to observe ruptures on 2020 Easter's day (significantly lower prices due to Covid19 lockdown on top of Easter's day) or on early October 2021 (corresponding to the increase in gas and carbon emission prices), and to showcase the importance of aggregating the lower and upper bounds independently as they model very distinct phenomena.

4.2 Contributions' summary of Part III – Missing Values

To encode missing values, we define the mask, or missing pattern, $M \in \mathcal{M} \subseteq \{0, 1\}^d$ as the binary random vector such that, for any $i \in \llbracket 1, d \rrbracket$, $M_i = 1$ if and only if X_i is missing. Therefore, there exists at most 2^d masks: this number grows exponentially in the problem's dimension, posing statistical and computational challenges. One of the most popular strategies to deal with missing values in a supervised learning framework suggests imputing the missing entries with plausible values to get completed data, on which any analysis can be performed (Le Morvan et al., 2021). This is called *impute-then-predict*.

Chapter 7 detailed summary. We study CP with missing covariates, aiming to build predictive sets that now depend on the mask, i.e. $\hat{C}_{n,\alpha}(X, M)$. Specifically, we study downstream Quantile Regression (QR) based CP, like CQR (Romano et al., 2019), on impute-then-predict strategies. Still, the proposed approaches also encapsulate other regression algorithms, and even classification.

We show that CP on impute-then-predict is *marginally* valid regardless of the model, missingness distribution, and imputation function. We describe how different masks (i.e. the set of observed covariates) introduce additional heteroskedasticity: *the predictive uncertainty strongly depends on the set of covariates observed*. We therefore focus on achieving valid coverage *conditionally on the mask*, coined MCV – Mask-Conditional-Validity. MCV is desirable in practice, as occurrence of missing values are linked to important attributes.

Traditional approaches such as QR and CQR fail to achieve MCV because they do not account for this core connection between missing values and uncertainty. Figure 4.1 shows on a toy example with only 3 features – thus $2^3 - 1 = 7$ possible masks – how the coverage of QR and CQR varies depending on the mask. Both methods dramatically undercover when the most important variable (X_2) is missing, and the loss of coverage worsens when

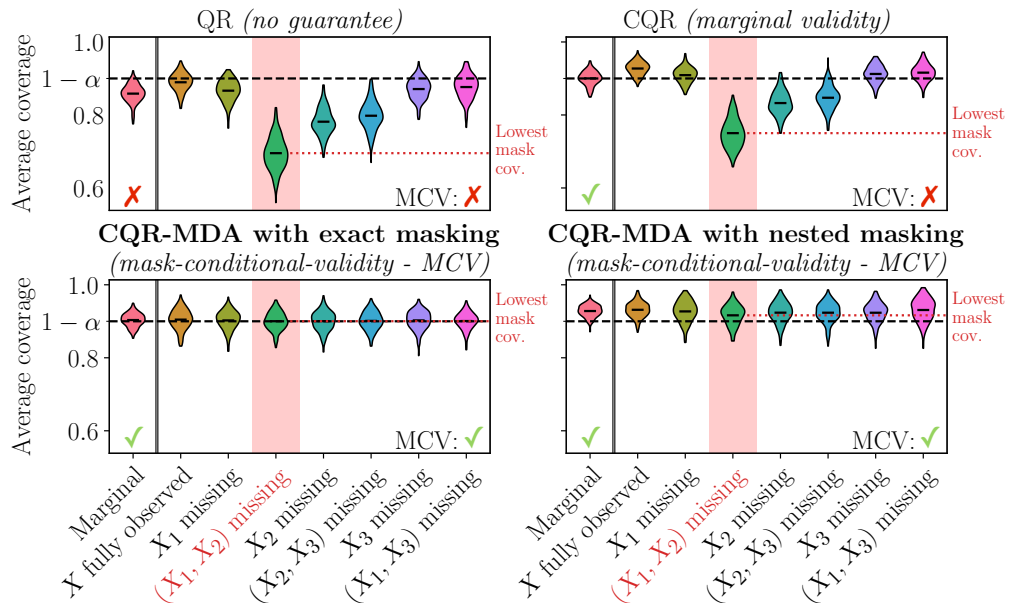


Figure 4.1: Coverage of the predictive intervals depending on which features are missing, among the 3 features. Evaluation over 200 runs.

additional features are missing.

We show how to form prediction intervals that are MCV, by suggesting two conformal methods sharing the same core idea of missing data augmentation (MDA): the calibration data is artificially masked to match the mask of the test point.

The first one, *CP-MDA with exact masking*, relies on building an ideal calibration set whose data points have the exact same mask as of the test point. We show its MCV under exchangeability and Missing Completely At Random.

The second one, *CP-MDA with nested masking*, does not require such an ideal calibration set. Instead, it builds a calibration set in which the data points have *at least* the same mask as the test point, i.e., this artificial masking results in calibration points having possibly more missing values than the test point. We show the latter method also achieves MCV, at the cost of an additional assumption: stochastic domination of the quantiles.

Figure 4.1 illustrates CP-MDA's MCV, as their **lowest mask coverage** is above $1 - \alpha$. We strengthen the empirical validation of our algorithms through more complex synthetic experiments than in Figure 4.1, along with semi-synthetic experiments where only the distribution of M given (X, Y) is controlled but not the distribution of (X, Y) . And finally, we conduct experiments on real critical care data. All of these experiments confirm that MDA achieves MCV while CQR fails to ensure MCV.

Chapter 8 detailed summary. Following the introduction of MCV criterion in Chapter 7, our objective in this chapter is to deepen the discussion on when and how it is possible to achieve MCV. Notably, we are interested in understanding *i*) what assumptions are necessary to ensure informative MCV is achievable, *ii*) how to design a MCV-tailored methodology with general probabilistic models, and *iii*) what happens when these assumptions are not satisfied.

First, we provide hardness results on (distribution-free) MCV.

Theorem 4.2.1. *If any $\hat{C}_{n,\alpha}$ is distribution-free MCV then for any distribution P , for any mask m such that $P_M(m) > 0$, it holds:*

- Regression: $\mathbb{P}_{P^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha}(X_{n+1}, m) \right) = \infty \right) \geq 1 - \alpha - \Delta_{m,n},$
- Classification: *for any $y \in \mathcal{Y}$, $\mathbb{P}_{P^{\otimes(n+1)}} \left(y \in \hat{C}_{n,\alpha}(X_{n+1}, m) \right) \geq 1 - \alpha - \Delta_{m,n},$*

where $\Delta_{m,n} \leq P_M(m)\sqrt{n+1}$.

In other words, any distribution-free MCV estimator outputs uninformative predictive sets on any mask that does not represent a high enough proportion of the training data. We deepen the analysis and show that this remain true *i*) if we suppose that the estimator is MCV only when M and X are independent, and *ii*) if we suppose that the estimator is MCV only when M is independent of Y given X . Therefore, to hopefully construct an estimator that provides meaningful MCV, it has to be MCV only on distribution such that the dependence between M and the pair (X, Y) is constrained. Importantly, this theoretical analysis brings new insights on the achievability of X -group-conditional validity (i.e. conditioning on the event $X \in \mathcal{V}(x)$), beyond MCV.

Second, we investigate the interplay between missing values and quantile regression. Characterizing it is hard in general, but becomes explicit under a multivariate Gaussian setting or linear model. We show that *i*) predictive uncertainty often increases with more missing values: we provide formal statements of this idea (in terms of conditional variance, inter-quantile distance and predictive interval length) and exhibit assumptions under which these properties are satisfied; *ii*) when one aims at estimating quantiles, it is crucial that the learner is able to retrieve the mask to construct intervals, in contrast to classic mean regression where the mask is not as essential; *iii*) especially, data dependent imputation might not be the best choice for predictive uncertainty quantification that is adaptive to the mask.

Third, we unify the algorithmic framework of Chapter 7 into a unique methodology, coined **CP-MDA-Nested***, that explicitly allows for the classification setting. It bridges the gap between the precision of strict subsampling to obtain the exact same mask (associated with high coverage variance), and the variance reduction of keeping all of the observations (associated with overly conservative predictive sets), by allowing any subsampling scheme, as long as it is independent of the calibration and test features and labels. Moreover, we draw an important analogy between **CP-MDA-Nested*** and leave-one-out or K -fold CP approaches. This enables us to provide theoretical guarantees on **CP-MDA-Nested*** in terms of MCV, under exchangeability and Missing Completely At Random assumptions.

Lastly, we conduct broader experiments than in Chapter 7 showcasing that **CP-MDA-Nested*** is empirically robust to strong dependence between M and X , as it maintains MCV under various Missing At Random and Missing Non At Random scenarios. However, when $Y \perp\!\!\!\perp M \mid X$ is not satisfied, **CP-MDA-Nested*** does not ensure MCV experimentally, unless the imputation is accurate enough. Overall, these numerical experiments emphasize the robustness of **CP-MDA-Nested*** beyond its theoretical scope of validity.

Time Series



Chapter 5

Adaptive Conformal Predictions for Time Series

Uncertainty quantification of predictive models is crucial in decision-making problems. Conformal prediction is a general and theoretically sound answer. However, it requires exchangeable data, excluding time series. While recent works tackled this issue, we argue that Adaptive Conformal Inference (ACI, [Gibbs and Candès, 2021](#)), developed for distribution-shift time series, is a good procedure for time series with general dependency. We theoretically analyse the impact of the learning rate on its efficiency in the exchangeable and auto-regressive case. We propose a parameter-free method, AgACI, that adaptively builds upon ACI based on online expert aggregation. We lead extensive fair simulations against competing methods that advocate for ACI's use in time series. We conduct a real case study: electricity price forecasting. The proposed aggregation algorithm provides efficient prediction intervals for day-ahead forecasting. All the code and data to reproduce the experiments are made available on [GitHub](#).

Contents

5.1	Introduction	62
5.2	Setting: ACI for time series	64
5.3	Impact of γ on ACI efficiency	65
5.3.1	Exchangeable case	66
5.3.2	AR(1) case	66
5.4	Adaptive strategies based on ACI	68
5.5	Numerical evaluation on synthetic data sets	69
5.5.1	Data generation process and settings	70
5.5.2	Impact of γ , performance of AgACI	71
5.5.3	Description of baseline methods	72
5.5.4	Experimental results: impact of φ, θ	73
5.6	Forecasting French electricity spot prices	74
5.6.1	Presentation of the price data	74
5.6.2	Price prediction with predictive intervals in 2019	75
5.7	Conclusion	76
5.A	Details on Split Conformal Prediction	77
5.B	Proof of the results presented in Section 5.3 and additional numerical experiments	81
5.C	Experimental details.	94
5.D	Additional experiments on synthetic data sets	97
5.E	Forecasting French electricity spot prices	100

5.1 Introduction

The increasing use of renewable intermittent energy leads to more dependent and volatile energy markets. Therefore, an accurate electricity price forecasting is required to stabilize energy production planning, gathering loads of research work as evidenced by recent substantial reviews (Weron, 2014; Lago et al., 2018, 2021). Furthermore, probabilistic forecasts are needed to develop risk-based strategies (Gaillard et al., 2016; Maciejowska et al., 2016; Nowotarski and Weron, 2018; Uniejewski and Weron, 2021). On the one hand, the lack of uncertainty quantification of predictive models is a major barrier to the adoption of powerful machine learning methods. On the other hand, probabilistic forecasts are only valid asymptotically or upon strong assumptions on the data.

Conformal prediction (CP, Vovk et al., 1999, 2005; Papadopoulos et al., 2002) is a promising framework to overcome both issues. It is a general procedure to build predictive intervals for any (black box) predictive model, such as neural networks, which are *valid* (i.e. achieve nominal marginal coverage) in finite sample and without any distributional assumptions except that the data are exchangeable.

Thereby, CP has received increasing attention lately, favored by the development of *split conformal prediction* (SCP, Lei et al., 2018, reformulated from *inductive* CP, Papadopoulos et al., 2002). More formally, suppose we have n training samples $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in \llbracket 1, n \rrbracket$, realizations of random variables $(X_1, Y_1), \dots, (X_n, Y_n)$, and that we aim at predicting a new observation y_{n+1} at x_{n+1} . Given a *miscoverage rate* $\alpha \in [0, 1]$ fixed by the user (typically 0.1 or 0.05) the aim is to build a predictive interval \mathcal{C}_α such that:

$$\mathbb{P} \{Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})\} \geq 1 - \alpha, \quad (5.1)$$

with \mathcal{C}_α as small as possible, in order to be informative. For the sequel, we call a *valid interval* an interval satisfying equation (5.1) and an *efficient interval* when it is as small as possible (Vovk et al., 2005; Shafer and Vovk, 2008).

To achieve this, SCP first splits the n points of the training set into two sets $\text{Tr}, \text{Cal} \subset \llbracket 1, n \rrbracket$, to create a *proper training set*, Tr , and a *calibration set*, Cal . On the proper training set a regression model $\hat{\mu}$ (chosen by the user) is fitted, and then used to predict on the calibration set. A *conformity score* is applied to assess the conformity between the calibration's response values and the predicted values, giving $S_{\text{Cal}} = \{(s_i)_{i \in \text{Cal}}\}$. In regression, usually the absolute value of the residuals is used, i.e. $s_i = |\hat{\mu}(x_i) - y_i|$. Finally, a corrected¹ $(1 - \hat{\alpha})$ -th quantile of these scores $\hat{Q}_{1-\hat{\alpha}}(S_{\text{Cal}})$ is computed to define the size of the interval, which, in its simplest form, is centered on the predicted value: $\mathcal{C}_\alpha(x_{n+1}) = \hat{C}_{\hat{\alpha}}(x_{n+1}) := [\hat{\mu}(x_{n+1}) \pm \hat{Q}_{1-\hat{\alpha}}(S_{\text{Cal}})]$. These steps are detailed in Section 5.A, and illustrated in Figure 5.9. More details on CP, including beyond regression, are given in Vovk et al. (2005); Angelopoulos and Bates (2023).

The cornerstone of SCP *validity* results is the exchangeability assumption of the data (see Lei et al., 2018, and Section 5.A.3). However, this assumption is not met in time series forecasting problems. Despite the lack of theoretical guarantees, several works have

¹The correction $\alpha \rightarrow \hat{\alpha}$ is needed because of the inflation of quantiles in finite sample (see Lemma 2 in Romano et al. (2019) or Section 2 in Lei et al. (2018)).

applied CP to time series. [Dashevskiy and Luo \(2008, 2011\)](#) apply original (*inductive*) CP ([Papadopoulos et al., 2002](#)) to both simulated (using Auto-Regressive Moving Average (ARMA) processes) and real network traffic data and obtain *valid* intervals. [Wisniewski et al. \(2020\)](#); [Kath and Ziel \(2021\)](#) apply SCP respectively to financial data (e.g. markets makers' net positions) and to electricity price forecasting on various markets. In order to account for the temporal aspect, they consider an online version of SCP. In both studies, the *validity* varied greatly depending on the markets and the underlying regression model, suggesting that further developments of CP and theoretical guarantees for time series are needed.

To this end, [Chernozhukov et al. \(2018\)](#) extend the CP theory to ergodic cases in order to include dependent data. [Xu and Xie \(2021\)](#) improve on that theory and propose a new algorithm, Ensemble Prediction Interval (EnbPI), adapted to time series by adding a sequential aspect.

Another case that breaks the exchangeability assumption is *distribution shift*, which allows for example to deal with cases where the test data is shifted with respect to the training data. [Tibshirani et al. \(2019\)](#) consider covariate shift while [Cauchois et al. \(2024\)](#) tackle a joint distributional shift setting (that is, of (X, Y)). In both studies, a single shift in the distribution is considered, a major limitation for applying these methods to time series. In an adversarial setting, [Gibbs and Candès \(2021\)](#) propose Adaptive Conformal Inference (ACI), accounting for an undefined number of shifts on the joint distribution. It is based on refitting the predictive model, as well as updating online the quantile level used by a recursive scheme depending on an hyper-parameter γ (a learning rate). Furthermore, they prove an asymptotic *validity* result for any data distribution.

We argue in this work that the design and guarantees of ACI can be beneficial for dependent data without distribution shifts.

Contributions. We propose to analyse ACI ([Gibbs and Candès, 2021](#)) in the context of time series with general dependency and make the following contributions:

- Relying on an asymptotic analysis of ACI's behaviour for simple time series distribution, we prove that ACI deteriorates *efficiency* in an exchangeable case (closed-form expression) while improving it in an AR setting (numerical approximation) with a well-chosen γ (Section 5.3).
- We introduce AgACI, a parameter-free method using online expert aggregation, to avoid choosing γ , achieving good performances in terms of *validity* and *efficiency* (Section 5.4).
- We compare ACI to EnbPI and online SCP on extensive synthetic experiments and we propose an easy-to-interpret visualisation combining *validity* and *efficiency* (Section 5.5).
- We forecast and give predictive intervals on French electricity prices, an area where accurate predictions, but also controlled predictive intervals, are required (Section 5.6).

To allow for better benchmarking of existing and new methods, we provide (re-)implementations in Python of (all) the described methods and a complete pipeline of analysis on [GitHub](#). As explained in Section 5.4, the code for AgACI is, for now, the only one available only in R.

Notations. In the sequel, the following notations are used: $\llbracket a, b \rrbracket := \{a, a + 1, \dots, b\}$;

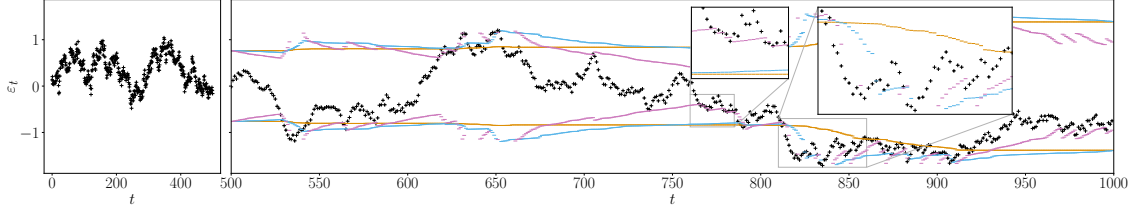


Figure 5.1: ACI on one simulated path ε_t , $t = 1, \dots, 1000$, from an AR(1) process (in black). The first 500 values form the initial calibration set (left subplot), and predicted interval bounds are computed on the last 500 points (right) for $\gamma = 0$, $\gamma = 0.01$ and $\gamma = 0.05$.

\mathbb{Q} refers to the set of rational numbers; $\mathcal{C}^4([0, 1])$ refers to the set of 4-times continuously differentiable functions on $[0, 1]$; $\stackrel{\text{not.}}{=}$ defines a notation; $\#A$ is the cardinal of the set A .

5.2 Setting: ACI for time series

In this section, we introduce ACI and our framework. We consider T_0 observations $(x_1, y_1), \dots, (x_{T_0}, y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$. The aim is to predict the response values and give predictive intervals for T_1 subsequent observations $x_{T_0+1}, \dots, x_{T_0+T_1}$ sequentially: at any prediction step $t \in \llbracket T_0 + 1, T_0 + T_1 \rrbracket$, $y_{t-T_0}, \dots, y_{t-1}$ have been revealed. Thereby, the data $((x_{t-T_0}, y_{t-T_0}), \dots, (x_{t-1}, y_{t-1}))$ are used for the construction of the predicted interval.

Adaptive Conformal Inference. Proposed by [Gibbs and Candès \(2021\)](#), ACI is designed to adapt CP to temporal distribution shifts. The idea of ACI is twofold. First, one considers an online procedure with a random split², i.e., Tr_t and Cal_t are random subsets of the last T_0 points. Second, to improve adaptation when the data is highly shifted, an *effective miscoverage level* α_t , updated recursively, is used instead of the target level α . Set $\alpha_1 = \alpha$, and for $t \geq 1$

$$\begin{cases} \hat{C}_{\alpha_t}(x_t) &= [\hat{\mu}(x_t) \pm \hat{Q}_{1-\alpha_t}(S_{\text{Cal}_t})] \\ \alpha_{t+1} &= \alpha_t + \gamma \left(\alpha - \mathbb{1}\{y_t \notin \hat{C}_{\alpha_t}(x_t)\} \right), \end{cases} \quad (5.2)$$

for $\gamma \geq 0$ ³. If ACI does not cover at time t , then $\alpha_{t+1} \leq \alpha_t$, and the size of the predictive interval increases; conversely when it covers. Nothing prevents $\alpha_t \leq 0$ or $\alpha_t \geq 1$. While the later is rare (as α is small) and produces by convention $\hat{C}_{\alpha_t}(\cdot) = \{\hat{\mu}(\cdot)\}$ (i.e. $\hat{Q}_{1-\alpha_t} = 0$), the former can happen frequently for some γ , giving $\hat{C}_{\alpha_t} \equiv \mathbb{R}$ ($\hat{Q}_{1-\alpha_t} = +\infty$).

How to deal with infinite intervals. A specificity of ACI's algorithm is thus to often produce infinite intervals. Defining the *average* length of an interval is then impossible. In order to assess the *efficiency* in the following, we consider two solutions: (i) imputing the length of infinite intervals by (twice) the overall maximum of the residuals, or $Q(1)$ if the residual's quantile function is known and bounded⁴; (ii) focusing on the median instead.

ACI on time series with general dependency. As highlighted by [Wisniewski et al. \(2020\)](#); [Kath and Ziel \(2021\)](#), the first step to adapt a method for dependent time series

²Figure 5.5(a) with [training](#) and [calibration](#) part shuffled randomly.

³ACI actually wraps around *any* CP procedure, here the definition is given using mean regression SCP.

⁴This happens in practice when the response and prediction are bounded, e.g., thanks to physical/real constraints as for the spot prices presented in Section 5.6.1, that are bounded by market rules.

is to work online which is the case for ACI. Moreover, the update of the quantile level according to the previous error implies that ACI could cope with a fitted model that has not correctly caught the temporal evolution, such as a trend, a seasonality pattern or a dependence on the past. Therefore, ACI is a perfect candidate for CP for time series with general dependency. To account for the temporal structure, we change the random split to a sequential split.⁵

To gain understanding on ACI in the context of dependent temporal data, we analyse a situation where a fitted regression model $\hat{\mu}$ produces AR(1) residuals, thus $y_t - \hat{\mu}(x_t) = \varepsilon_t$, where ε_t is an AR(1) process: $\varepsilon_{t+1} = 0.99\varepsilon_t + \xi_{t+1}$, with $\xi_t \sim \mathcal{N}(0, 0.01)$. We plot this toy example in Figure 5.1, for $T_0 = T_1 = 500$. Three versions of ACI are compared: $\gamma = 0$, the quantile level is not updated but the calibration set Cal_t is; $\gamma = 0.01$ and $\gamma = 0.05$. To obtain an insightful visualisation⁶, we represent the interval $[\pm \hat{Q}_{1-\alpha_t}(S_{\text{Cal}_t})]$ instead of $\hat{C}_{\alpha_t}(x_t)$. When no intervals are displayed, ACI is predicting \mathbb{R} . Here and in the sequel, we use $\alpha = 0.1$.

In this toy example, the coverage rate among many observations is *valid* for $\gamma \in \{0.01, 0.05\}$ (90% and 92% of points included) but not for $\gamma = 0$ (72.6%). Moreover, Figure 5.1 shows that the type of errors depends on γ . For $\gamma = 0$, ACI excludes consecutive observations (e.g. for $t \in [810, 860]$, zoomed-in plot). For $\gamma \in \{0.01, 0.05\}$, ACI manages to adapt to these observations, and the higher the γ , the less the adaptation is delayed. Furthermore, when the residuals are small and far from both interval bounds, ACI quickly reduces the interval's length and produces more *efficient* intervals. Consequently, ACI may also not cover on points for which the residuals have a relatively small values compared to the calibration's values (e.g. for $t \in [760, 785]$).

5.3 Impact of γ on ACI efficiency

The choice of the parameter γ strongly impacts the behaviour of ACI: while the method always satisfies the *asymptotic validity* property, i.e. $\frac{1}{T} \sum_{t=1}^T \mathbb{1}\{y_t \notin \hat{C}_{\alpha_t}(x_t)\} \xrightarrow[T \rightarrow \infty]{a.s.} \alpha$ (Proposition 4.1 in Gibbs and Candès, 2021), this property does not give any insight on the length of resulting intervals. Besides, this guarantee directly stems from the fact that $\frac{1}{T} \sum_{t=1}^T \mathbb{1}\{y_t \notin \hat{C}_{\alpha_t}(x_t)\} - \alpha \leq 2/(\gamma T)$. This tends to suggest the use of larger γ values, that unfortunately generate frequent infinite intervals. Here, we thus analyse the impact of γ on ACI's *efficiency* in simple yet insightful cases: in Section 5.3.1, focusing on the exchangeable case, then in Section 5.3.2, with a simple AR process on the residuals.

Approach. Our focus is on the impact of the key parameter γ . Analysing simple theoretical distributions allows to build intuition on the behaviour of the algorithm for more complex data structure. In order to derive theoretical results, we thus make supplementary modelling assumptions on the residuals, and do not consider the impact of the calibration set: we introduce Q the quantile function of the scores and assume, for all $\hat{\alpha}$ and t , $\hat{Q}_{1-\hat{\alpha}}(S_{\text{Cal}_t}) = Q(1 - \hat{\alpha})$. This corresponds to considering the limit as $\#\text{Cal} \rightarrow \infty$. This

⁵As in Figure 5.5(a). This is also consistent with OSSCP (Sec. 5.5.3).

⁶We suggest focusing the visualisation on the scores to analyse the behaviour of CP methods, as they are at the core of the *validity* proof. A detailed discussion on this is given in App. 5.A.5

allows to focus on the impact of recursive updates in (5.2) and describe their behaviour by relying on Markov Chain theory.

5.3.1 Exchangeable case

ACI is usually applied in an adversarial context. If the scores are actually exchangeable, ACI's *validity* would not improve upon SCP (known to be quasi-exactly *valid*), thus assessing ACI's impact on *efficiency* is necessary. Define $L(\alpha_t) = 2Q(1 - \alpha_t)$ the length of the interval predicted by the adaptive algorithm at time t , and $L_0 = 2Q(1 - \alpha)$ the length of the interval predicted by the non-adaptive algorithm (or equivalently, $\gamma = 0$).

Theorem 5.3.1. *Assume that: (i) $\alpha \in \mathbb{Q}$; (ii) the scores are exchangeable with quantile function Q ; (iii) the quantile function is perfectly estimated at each time (as defined above); (iv) the quantile function Q is bounded and $\mathcal{C}^4([0, 1])$. Then, for all $\gamma > 0$, $(\alpha_t)_{t>0}$ forms a Markov Chain, that admits a stationary distribution π_γ , and*

$$\frac{1}{T} \sum_{t=1}^T L(\alpha_t) \xrightarrow[T \rightarrow +\infty]{a.s.} \mathbb{E}_{\pi_\gamma}[L] \stackrel{not.}{=} \mathbb{E}_{\tilde{\alpha} \sim \pi_\gamma}[L(\tilde{\alpha})].$$

Moreover, as $\gamma \rightarrow 0$,

$$\mathbb{E}_{\pi_\gamma}[L] = L_0 + Q''(1 - \alpha) \frac{\gamma}{2} \alpha(1 - \alpha) + O(\gamma^{3/2}).$$

Interpretation of assumptions. Assumption (i) is weak since a practitioner will always select $\alpha \in \mathbb{Q}$ while assumption (ii) describes the classical exchangeable setting. The main assumptions are (iii) and (iv): (iii) can be interpreted as considering an infinite calibration set while (iv) is necessary⁷ in order to define $\mathbb{E}_{\pi_\gamma}[L]$: here, we extend $Q(1 - \hat{\alpha})$ by $Q(1)$ for $\hat{\alpha} < 0$. When $\hat{Q} \equiv \hat{Q}_t$ is the empirical quantile function on a calibration set Cal, the convergence in Theorem 5.3.1 holds conditionally to Cal. Finally, the regularity assumption on Q is purely technical.

Interpretation of the result. For standard distributions, $Q''(1 - \alpha) > 0$,⁸ and Theorem 5.3.1 implies that ACI on exchangeable scores *degrades* the *efficiency* linearly with γ compared to CP. This is an important takeaway from the analysis, that underlines that such adaptive algorithms may actually hinder the performance if the data does not have any temporal dependency, and a small γ is preferable. For example, if the residuals are standard gaussians, for $\alpha = 0.01$, setting $\gamma = 0.03$ (resp. $\gamma = 0.05$) will increase the length by 1.59% (resp. by 3.38%) with respect to $\gamma = 0$.

5.3.2 AR(1) case

We now consider the case of (highly) correlated residuals, which happens in many practical time series applications.

⁷ $\forall \gamma > 0$, $\mathbb{P}_{\pi_\gamma}(\tilde{\alpha} \leq 0) > 0$: we need $|Q(1)| < \infty$ to define $\mathbb{E}_{\pi_\gamma}[L]$.

⁸as $Q'(x) = \frac{1}{f(Q(x))}$ with f the scores' probability density function, $Q'(x)$ increases locally around x if and only if f decreases locally around $Q(x)$ (Q is increasing). Thus, $Q''(x) > 0$ if and only if f decreases locally around $Q(x)$. Thereby, for $x = 1 - \alpha$ high (usually the case), $Q''(1 - \alpha) > 0$ for standard distributions.

Definition 5.3.1 (AR(1) clipped). $\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1}$ with $(\xi_t)_t$ i.i.d. random variables admitting a continuous density with respect to Lebesgue measure, of support \mathcal{S} clipped at a large value R , and $[-R, R] \subset \mathcal{S}$

Theorem 5.3.2. Assume that: (i) $\alpha \in \mathbb{Q}$; (ii) the residuals follow an AR(1) process clipped at R of parameter φ (Definition 5.3.1); (iii) the quantile function Q of the stationary distribution of $(\varepsilon_t)_t$ is known. Then $(\alpha_t, \varepsilon_{t-1})$ is a homogeneous Markov Chain in \mathbb{R}^2 that admits a unique stationary distribution $\pi_{\gamma, \varphi}$. Moreover,

$$\frac{1}{T} \sum_{t=1}^T L(\alpha_t) \xrightarrow[T \rightarrow +\infty]{a.s.} \mathbb{E}_{\pi_{\gamma, \varphi}}[L].$$

We numerically estimate $\gamma_\varphi^* = \operatorname{argmin}_\gamma \mathbb{E}_{\pi_{\gamma, \varphi}}[L]$ in Figure 5.2. To do so, AR(1) processes of length $T = 10^6$ are simulated for various φ and asymptotic variance 1. ACI is applied on each of them, with 100 different $\gamma \in [0, 0.2]$. Figure 5.2 (left) represents the average length depending on γ for each φ , and (right) the values of γ minimizing this average length for each φ (for 25 repetitions of the experiment). The average length is computed after imputing all the infinite intervals' length by the maximum of the process, as explained in Section 5.2. A similar study using instead the median length is provided after the proofs in Section 5.B.

Interpretation. We make the following observations:

1. For high φ , ACI indeed improves for a strictly positive γ upon $\gamma = 0$. This proves that ACI can be used to produce smaller intervals for time series CP. The function $\gamma \mapsto \mathbb{E}_{\pi_{\gamma, \varphi}}[L]$ decreases until γ_φ^* , then increases again, as expected because very large γ cause the algorithm to be less stable and produce numerous infinite intervals.

2. In Figure 5.2 (left), zoomed-in plot, the black line represents asymptotic result of Theorem 5.3.1. We retrieve here that the expected length is minimal for $\gamma = 0$ and grows linearly with γ around 0. This behaviour is very similar for $\varphi = 0.6$.

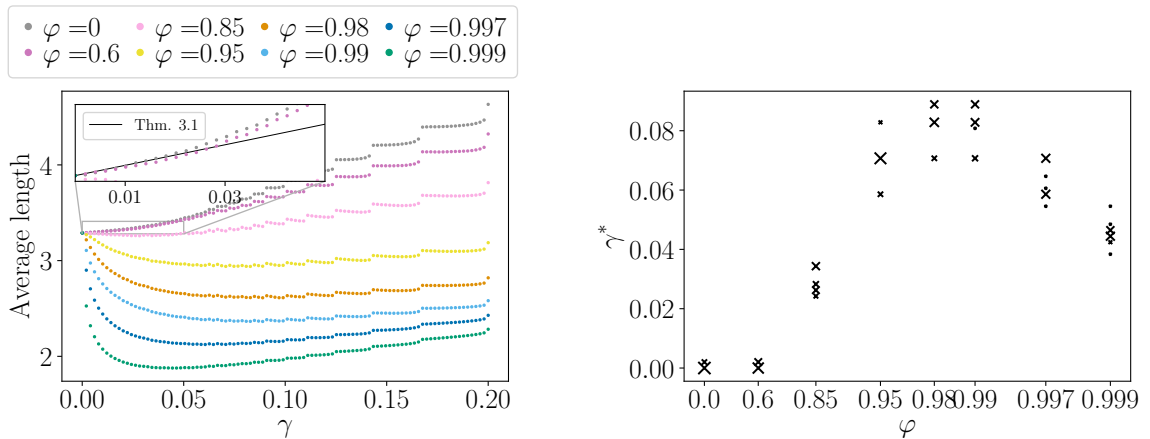


Figure 5.2: Left: evolution of the mean length depending on γ for various φ . Right: γ^* minimizing the average length for each φ (each cross has a size proportional to the number of runs for which γ^* was the minimizer).

3. For any γ , the function $\varphi \mapsto \mathbb{E}_{\pi_{\gamma, \varphi}}[L]$ is decreasing (Figure 5.2, left). Indeed, stronger correlation between residuals (i.e., a higher φ), allows to build smaller intervals. This confirms that ACI's impact strengthens with the strength of the temporal dependence.

4. Surprisingly, the function $\varphi \mapsto \gamma_\varphi^*$, that corresponds to the optimal learning rate for a given signal, is *non-monotonic*, (Figure 5.2, right). As $\gamma = 0$ is optimal for $\varphi = 0$, the function first increases. However, the optimal learning rate then diminishes as φ increases. This sheds light on the complex intrinsic tradeoffs of the method: for small values of φ , using $\gamma > 0$ simply degrades the *efficiency*; for “moderate” values of φ using a larger γ is necessary to quickly benefit from the short-term dependency between residuals; finally, for larger values of φ , the process exhibits a longer memory, thus it is crucial to find a smaller learning rate that produces more stable intervals, even if it means that the algorithm won't adapt as quickly.

What if $Q \neq \hat{Q}$? While our analysis provides a first step by comparing ACI to CP in the ideal case where the quantile distribution is known (for both methods), the impact of the finite-Cal is of interest. Indeed, if Cal is small, ACI can help to attain coverage **conditionally to a given** Cal even in the i.i.d. case. Yet intuitively, **marginally**, the randomness induced by ACI in the i.i.d. case would negatively impact efficiency w.r.t. $\gamma = 0$, even in the finite-Cal case. Finite sample trade-offs and general analysis of the case $Q \neq \hat{Q}$ is an important open direction.

Overall, these results highlight the importance of the choice of γ , as not choosing γ^* can lead to significantly larger intervals. In addition, they provide insights on the corresponding dynamics. Yet the choice of γ in more complex practical settings remains difficult: this calls for adaptive strategies.

5.4 Adaptive strategies based on ACI

To prevent the critical choice of γ an ideal solution is an adaptive strategy with a time dependent γ . We propose two strategies based on running ACI for $K \in \mathbb{N}$ values $\{(\gamma_k)_{k \leq K}\}$ of γ , chosen by the user. Overall, this does not increase the computational cost because Tr_t and Cal_t are shared between all ACI; thus the only additional cost is the computation of the K different quantiles. For any x_t , denote $\hat{C}_{\alpha_t, k}(x_t)$ the interval at time t built by ACI using γ_k .

Naive strategy. A simple strategy is to use at each step the γ that achieved in the past the best *efficiency* while ensuring *validity*. For stability purposes, consider a warm-up period $T_w \leq T_1 - 1$. For each $t \geq T_0 + T_w$, we select $k_{t+1}^* \in \operatorname{argmin}_{k \in \mathcal{A}_t} \left\{ t^{-1} \sum_{s=1}^t \text{length}(\hat{C}_{\alpha_s, k}(x_s)) \right\}$ with $\mathcal{A}_t = \{k \in \llbracket 1, K \rrbracket \mid t^{-1} \sum_{s=1}^t \mathbb{1}_{y_s \in \hat{C}_{\alpha_s, k}(x_s)} \geq 1 - \alpha\}$ or $k_{t+1}^* \in \operatorname{argmin}_{k \in \llbracket 1, K \rrbracket} \{ |1 - \alpha - t^{-1} \sum_{s=1}^t \mathbb{1}_{y_s \in \hat{C}_{\alpha_s, k}(x_s)}| \}$ if $\mathcal{A}_t = \emptyset$. For the first T_w steps, an arbitrary strategy is applied (in simulations, $\gamma = 0$ for $t \leq T_w = 50$).

Online Expert Aggregation on ACI (AgACI). Instead of picking one γ in the grid, we introduce an adaptive aggregation of *experts* (Cesa-Bianchi and Lugosi, 2006), with expert k being ACI with parameter γ_k . This strategy is detailed in Algorithm 9, and schematised in Figure 5.3. At each step t , it performs two independent aggregations of the K -ACI intervals

$\hat{C}_{\alpha_{t,k}}(\cdot) \stackrel{\text{not.}}{=} [\hat{b}_{t,k}^{(\ell)}(\cdot), \hat{b}_{t,k}^{(u)}(\cdot)]$, one for each bound, and outputs $\tilde{C}_t(\cdot) \stackrel{\text{not.}}{=} [\tilde{b}_t^{(\ell)}(\cdot), \tilde{b}_t^{(u)}(\cdot)]$. Aggregation computes an optimal weighted mean of the experts (Line 11), where the weights $\omega_{t,k}^{(\ell)}$, $\omega_{t,k}^{(u)}$ assigned to expert k depend on all experts performances (suffered *losses*) at time steps $1, \dots, t$ (Line 9). We use the pinball loss ρ_β , as it is frequent in quantile regression, where the pinball parameter β is chosen to $\alpha/2$ (resp. $1 - \alpha/2$) for the lower (resp. upper) bound. These losses are plugged in the *aggregation rule* Φ . Finally, the aggregation rule can include the computation of the gradients of the loss (*gradient trick*, see [Cesa-Bianchi and Lugosi, 2006](#), for more details). As aggregation rules require bounded experts, a thresholding step is added (Line 5).

Note that the pinball loss helps to avoid large intervals (e.g. it strongly penalizes infinite or very large intervals).

We chose Φ to be the Bernstein Online Aggregation (BOA, [Wintenberger, 2017](#), see Section 5.C.1 for a brief description), that was successfully applied for financial data ([Berrisch and Ziel, 2023](#); [Remlinger et al., 2023](#)). We rely on R package OPERA ([Gaillard and Goude, 2021](#)), which allows the user to easily select among many other aggregation rules (EWA ([Vovk, 1990](#)), ML-Poly ([Gaillard et al., 2014](#)) or FTRL ([Shalev-Shwartz and Singer, 2007](#); [Hazan, 2019](#)), etc.) that give similar results in our experiments. We use the gradient trick in the simulations. In the sequel, AgACI refers to AgACI using BOA and gradient trick.

Algorithm 9 Online Expert Aggregation on ACI (AgACI)

Input: Miscoverage rate α , grid $\{\gamma_k, k \in \llbracket 1, K \rrbracket\}$, aggregation rule Φ , threshold values $M^{(\ell)}, M^{(u)}$.

- 1: Let $\beta^{(\ell)} = \alpha/2$ and $\beta^{(u)} = 1 - \alpha/2$
- 2: **for** $t \in \llbracket T_0 + 1, T_0 + T_1 \rrbracket$ **do**
- 3: **for** $k \in \llbracket 1, K \rrbracket$ **do**
- 4: Compute $\hat{b}_{t,k}^{(\cdot)}(x_t)$ using ACI with γ_k .
- 5: **if** $\hat{b}_{t,k}^{(\cdot)}(x_t) \notin \mathbb{R}$ **then** set $\hat{b}_{t,k}^{(\cdot)}(x_t) = M^{(\cdot)}$
- 6: **end for**
- 7: Set $\tilde{C}_t(x_t) = [\tilde{b}_t^{(\ell)}(x_t), \tilde{b}_t^{(u)}(x_t)]$
- 8: **for** $k \in \llbracket 1, K \rrbracket$ **do**
- 9: $\omega_{t,k}^{(\cdot)} = \Phi(\{\rho_{\beta^{(\cdot)}}(y_s, \hat{b}_{s,l}^{(\cdot)}(x_s)), s \in \llbracket T_0 + 1, t \rrbracket, l \in \llbracket 1, K \rrbracket\})$
- 10: **end for**
- 11: Define $\tilde{b}_{t+1}^{(\cdot)}(x) = \frac{\sum_{k=1}^K \omega_{t,k}^{(\cdot)} \hat{b}_{t+1,k}^{(\cdot)}(x)}{\sum_{l=1}^K \omega_{t,l}^{(\cdot)}}$ for any $x \in \mathbb{R}^d$
- 12: **end for**

5.5 Numerical evaluation on synthetic data sets

In this section we conduct synthetic experiments on a wide range of data sets presented in Section 5.5.1. The goal of this section is twofold. First, in Section 5.5.2, comparing our proposed adaptive strategies to ACI with a wide range of γ values. Second, in Section 5.5.4,

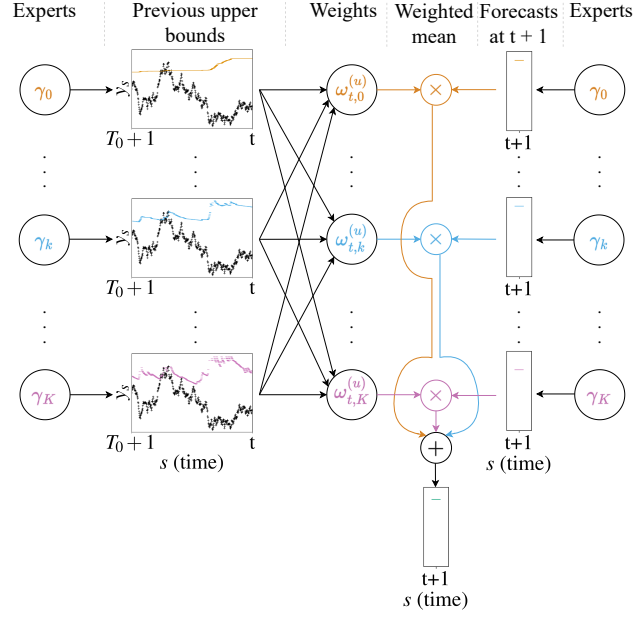


Figure 5.3: Scheme of AgACI algorithm, upper bound u only, for a forecast at time $t + 1$. A similar procedure is performed independently for the lower bound ℓ in parallel.

comparing performances of AgACI and ACI to that of competitors – namely EnbPI and online sequential SCP, described in Section 5.5.3.

5.5.1 Data generation process and settings

We generate data according to:

$$Y_t = 10 \sin(\pi X_{t,1} X_{t,2}) + 20 (X_{t,3} - 0.5)^2 + 10 X_{t,4} + 5 X_{t,5} + 0 X_{t,6} + \varepsilon_t, \quad (5.3)$$

where the X_t are multivariate uniformly distributed on $[0, 1]$, and $X_{t,6}$ represents an uninformative variable. The noise ε_t is generated from an ARMA(1,1) process of parameters φ and θ , i.e. $\varepsilon_{t+1} = \varphi \varepsilon_t + \xi_{t+1} + \theta \xi_t$, with ξ_t a white noise called the *innovation* (see Section 5.C.2 for details). When the noise is i.i.d., one retrieves the simulations from Friedman et al. (1983). The temporal dependence is present only in the noise in order to control its strength and its impact on the algorithms' performance.

Given the non-linear structure of the data generating process, we use a random forest (RF) as predictive model, with the same hyper-parameters through all the experiments (specified in Section 5.C.3).

To assess the impact of the temporal structure, we vary φ and θ in $\{0.1, 0.8, 0.9, 0.95, 0.99\}$. To focus on the impact of the dependence structure, the value of the innovation's variance is selected so that the asymptotic variance of ε_t is independent of φ, θ : here we choose $\lim_{t \rightarrow \infty} \text{Var}(\varepsilon_t) = 10$. For each set of parameters, we generate $n = 500$ samples $(\varepsilon_t)_{t \in [1, T_0+T_1]}$ with $T_0 = 200$. In the sequel we display the results on an ARMA(1,1) which are representative of all the results obtained. For the sake of simplicity, we consider $\varphi = \theta$. Complementary results (i) for an asymptotic variance of 1 (corresponding to a higher *signal to noise* ratio), (ii) for AR(1) and MA(1) models are available in Section 5.D.

Joint visualisation of validity & efficiency. In order to simultaneously assess *validity* and *efficiency*, in Figures 5.4, 5.6 and 5.8, we represent on the same graph the

empirical coverage and average median length (used for *efficiency* as imputing the infinite bounds by the maximum of the whole sequence is not always feasible in practice). In those three figures, the vertical dotted line represents the target miscoverage rate, $\alpha = 0.1$. Consequently, a method is *valid* when it lies at the right of this line, and the lower the better.

5.5.2 Impact of γ , performance of AgACI

Figure 5.4 illustrates the behaviour of ACI (with multiple values of γ), the naive strategy (empty triangles) and AgACI (black stars) for increasing (from left to right) values of φ , θ , with $T_1 = 200$. In particular, the top row shows the joint *validity* & *efficiency* and, for this figure only, we add in the bottom row the same graph using the average length after imputation (see details in Section 5.D) to assess *efficiency* in another way.

When γ is small, one observes an undercoverage, which increases when the temporal dependency of ε increases. Increasing γ enables ACI to increase the interval's size faster when we do not cover, and thus to improve *validity*, which is achieved for high values of γ ; however this also increases the frequency of uninformative (infinite) intervals, as deduced from the bottom row of Figure 5.4, where the average length after imputation grows with γ . Remark that these results do not contradict the validity result recalled at the beginning of Section 5.3, which is only asymptotic while we predict on 200 points. For φ, θ small, we observe that similarly to Theorem 5.3.1, the *efficiency* does not improve with γ . For moderate values of $\varphi, \theta \in \{0.8, 0.9, 0.95\}$, we observe that the average median length is decreasing with γ for $\gamma \geq 0.01$. This effect is observable on average but not present in all the 500 experiments. One possible explanation is that the shrinking effect of ACI on the predicted interval enables to significantly reduce the predicted interval when γ is large, and this effect is, on average, more important than the number of large intervals.

Moreover, the naive strategy is clearly not *valid*: indeed it results in greedily choosing a γ that achieved good results in the past, and is consequently slightly more likely to fail to

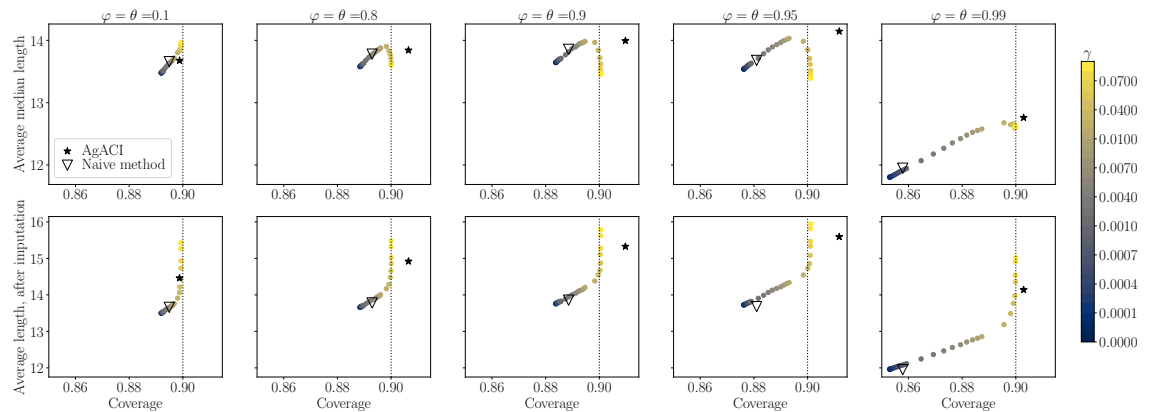


Figure 5.4: ACI performance with various θ , φ and γ on data simulated according to equation (5.3) with a Gaussian ARMA(1,1) noise of asymptotic variance 10 (see Section 5.C.2). Top row: average median length w.r.t. the coverage. Bottom row: average length after imputation w.r.t. the coverage. Stars correspond to AgACI, and empty triangles to the naive choice.

cover in future steps. Thereby, we do not consider it anymore. Finally, AgACI achieves *valid* coverage without increasing the median length with respect to each expert, and even improves the coverage. Overall, it appears to be a good candidate as a parameter-free method.

5.5.3 Description of baseline methods

We consider as baseline *online sequential split conformal prediction* (OSSCP), a generalisation of SCP⁹. The other competitor is EnbPI (Xu and Xie, 2021), specifically designed for time series. Pseudo-codes and details are given in Section 5.C.4. Offline SCP (for which $\text{Tr}_t \equiv \text{Tr}_0$ and $\text{Cal}_t \equiv \text{Cal}_0$) is not considered as a competitor because it is unfair to compare an *offline* algorithm to one that uses more recent data points. This corresponds to comparing a prediction at horizon T_{large} to one at horizon T_{small} . This is a limitation of the comparison in Xu and Xie (2021).

OSSCP. We consider an online version of SCP by refitting the underlying regression model and recalibrating using the newest points. Moreover, to appropriately account for the temporal structure of the data, we use a *sequential split* as in Wisniewski et al. (2020): at any t , the time indices in Tr_t are smaller than those of Cal_t . Not randomizing aims at excluding future observations from Tr_t , which may lead to an under-estimation of the errors on Cal_t , thus eventually to smaller intervals with under-coverage. We compare both splitting strategies on simulations in Section 5.D.4. OSSCP procedure is schematised in Figure 5.5(a).

Original EnbPI. EnbPI, Ensemble Prediction Interval (Xu and Xie, 2021), works by updating the list of *conformity scores* with the most recent ones so that the intervals adapt to latest performances, without refitting the underlying regression model. Thereby, the predicted intervals can adapt to seasonality and trend. In EnbPI, B bootstrap samples of the training set are generated and the regression algorithm is fitted on each bootstrap sample producing B predictors. Finally, the predictors are aggregated in two ways: first, for each training point of index $t \leq T_0$, EnbPI aggregates only the subset of predictors trained on bootstrap sample *excluding* (x_t, y_t) . This way, EnbPI constructs a set of hold-out calibration scores. Second, for test points of index $t > T_0$ EnbPI aggregates all the B predictors. A sketch of EnbPI is presented in Figure 5.5(b). Note that in Xu and Xie (2021) they use a classical bootstrap procedure, not dedicated to time series.

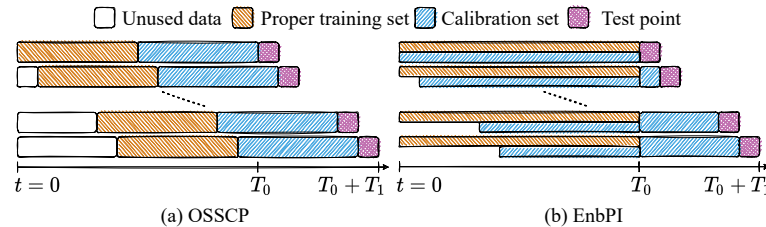


Figure 5.5: Scheme of the two baselines: OSSCP and EnbPI. In (a), Tr and Cal have equal size, but it can be changed.

⁹Recall here that inductive CP and SCP are equivalent methods.

They show empirically that it leads to *valid* coverage on real world time series, such as hourly wind power production and solar irradiation, while offline SCP fails to attain *valid* coverage.

EnbPI V2. Xu and Xie (2021) used the mean aggregation during the training phase and the $(1 - \alpha)$ -th quantile of the predictors for the prediction. We consider using the mean aggregation all along the procedure as mixing both aggregations may hurt the performance of the algorithm (as shown in the following simulations). Note that simultaneously to our work, authors released an updated version on ArXiv (Xu and Xie, 2021), incorporating a similar change.

5.5.4 Experimental results: impact of φ, θ

Figure 5.6 presents the results for data generated as in Section 5.5.1, for various (φ, θ) . Each sample contains 300 observations, with $T_0 = 200$ and $T_1 = 100$. We compare AgACI (with $K = 30$ experts), ACI (with $\gamma \in \{0.01, 0.05\}$), OSSCP, EnbPI and EnbPI V2 (with mean aggregation). To assess the impact and interest of an online procedure, we also add offline SCP. Finally, to ensure the robustness of our conclusions each experiment is repeated $n = 500$ times, and Figure 5.6 includes the standard errors (given by $\frac{\hat{\sigma}_n}{\sqrt{n}}$, where $\hat{\sigma}_n$ is the empirical standard deviation).

Each color is associated to a set (φ, θ) , each marker to an algorithm. To improve readability, we often link markers of the same method. There are thus two ways of analysing Figure 5.6: for a given method, the lines highlight the evolution of its performance with (φ, θ) ; for a given data distribution, the set of markers of its color allows to compare the methods. Figure 5.6, and results on AR(1) in Section 5.D.2.1, highlight that in an AR(1) or ARMA(1,1) process:

- Refitting the method (OSSCP vs Offline SCP) brings a significant improvement, that increases with higher dependence (higher values for φ and θ).
- All methods produce smaller intervals for $\varphi = \theta = 0.99$.
- EnbPI loses coverage while producing shorter intervals when the dependence increases. The performance of EnbPI depends significantly on the type and strength of dependence.
- EnbPI V2 is closer to the target coverage than EnbPI.
- OSSCP loses *validity* & coverage as φ and θ increase.
- While ACI with $\gamma = 0.01$ also struggles for high values of φ and θ such as 0.99, we observe that it still attains *valid* coverage with a well chosen γ . Most importantly, ACI performances are robust to the increase of the dependence strength: except for the $\varphi = \theta = 0.99$, its markers are really close to each other.
- AgACI always nearly attains *validity* (coverage is over 89.8% for all φ), and achieves the best *efficiency* performance among *valid* methods.

Note that ACI's *valid* coverage with some γ comes at the price of predicting more infinite intervals. A more detailed analysis on this phenomenon is conducted in Section 5.D.3. This can also be observed in graphs obtained with the average length after imputation, which are similar to Figure 5.6 and Section 5.D.2.1. In these graphs, the *validity* remains unchanged as expected, but the *efficiency* is more degraded for ACI with $\gamma = 0.05$ and for AgACI,

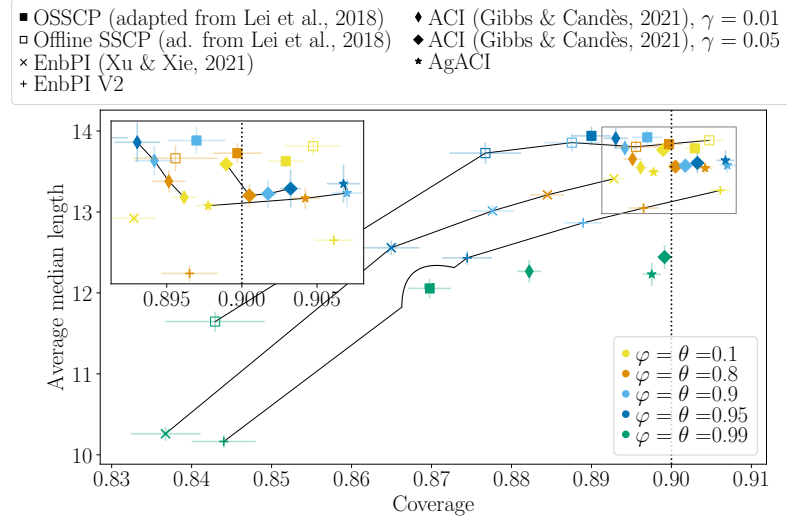


Figure 5.6: Performance of various CP methods on data simulated according to equation (5.3) with a Gaussian ARMA(1,1) noise of asymptotic variance 10 (see Section 5.C.2). Results aggregated from 500 independent runs. Empirical standard errors displayed.

since they produce more often uninformative intervals, as observed in Figure 5.4.

Summary. We highlight the following takeaways:

1. The temporal dependence impacts the *validity*.
2. Online is significantly better than offline.
3. **OSSCP**. Achieves *valid* coverage for φ and θ smaller than 0.9, but is not robust to the increasing dependence.
4. **EnbPI**. Its *validity* strongly depends on the data distribution (it is *valid* on a MA(1) noise, not in AR(1) and ARMA(1,1) noise). When the method is *valid*, it produces the smallest intervals. EnbPI V2 method should be preferred.
5. **ACI**. Achieves *valid* coverage for every simulation settings with a well chosen γ , or for dependence such that $\varphi < 0.95$. It is robust to the strength of the dependence.
6. **AgACI**. Achieves *valid* coverage for every simulation setting, with good *efficiency*.

5.6 Forecasting French electricity spot prices

In this last section, the task of forecasting French electricity spot prices with predictive intervals is considered in order to assess the methods on a real time series, and most importantly to show the relevance of ACI and AgACI in practice for time series without distribution shifts.

5.6.1 Presentation of the price data

The data set contains the French electricity spot prices, set by an auction market, from 2016 to 2019. Each day D before 12 AM, any producer (resp. supplier) submit their orders for the 24 hours of day $D + 1$. An order consists of an electricity volume in MWh offered for sale (resp. required to be purchased) and a price in €/MWh, at which they accept to sell (resp. buy) this volume. At 12 AM, the algorithm “Euphemia” ([EUPHEMIA](#)) fixes

the 24 hourly prices of day $D + 1$ according to these offers and additional constraints. Thereby, it is an hourly data set, containing $(3 \times 365 + 366) \times 24 = 35064$ observations. Our aim is to predict at day D (before 12 AM) the 24 prices of day $D + 1$. Given the prices' construction, we consider the following explanatory variables: day-ahead forecast consumption, day-of-the-week, 24 prices of the day $D - 1$ and 24 prices of the day $D - 7$. An extract of the considered data set is presented in Section 5.E.1.

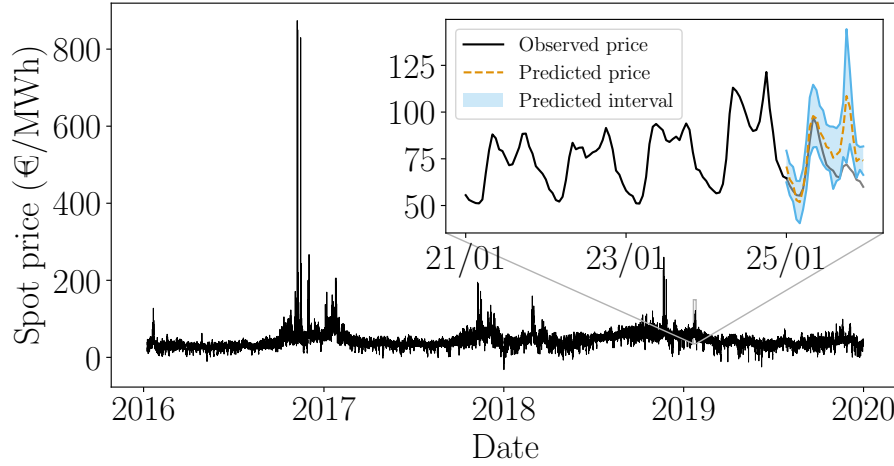


Figure 5.7: French electricity spot prices, from 2016 to 2019. Predicted intervals on the 25th of January 2019, using AgACI.

These prices exhibits medium to high peaks, as illustrated in Figure 5.7 where the French prices had reached 800 €/MWh in fall 2016, compared to an average price of approximately 40 €/MWh in 2019. These extreme events are mainly due to the non-storability of electricity and the inelasticity of the demand: when the demand is high compared to the available production, production units with expensive production costs must be called, leading to a huge market price.

5.6.2 Price prediction with predictive intervals in 2019

Since the 24 hours have very distinct patterns, we fit one model per hour, using again RF. We predict for the year 2019, using a sliding window of 3 years, as described in Figure 5.5(a), using one year and 6 months as proper training set and the most recent year and a half for calibration. The results are represented in Figure 5.8.

OSSCP over-covers but to a lesser extent than the offline version. This can be explained by a low presence of peaks during the test period. Indeed, by updating the whole procedure, the high peaks are “forgotten” which leads to small intervals while it is not the case for the offline version which leads to too large intervals. Thereby, online versions can help to improve *efficiency*, in addition to *validity*. **EnbPI** attains a *valid* coverage by over-covering. The under-coverage observed in the simulation study is not systematic, as in Xu and Xie (2021). **ACI** gives the smallest intervals with a correct coverage, for $\gamma = 0.01$ and $\gamma = 0.05$. The update of the quantile level enables to shrink the intervals. While the simulation in Section 5.5.4 study outlines that ACI improves *validity*, this application illustrates that it can provide *efficient* interval. **AgACI** is more *efficient* than $\gamma = 0$ while maintaining

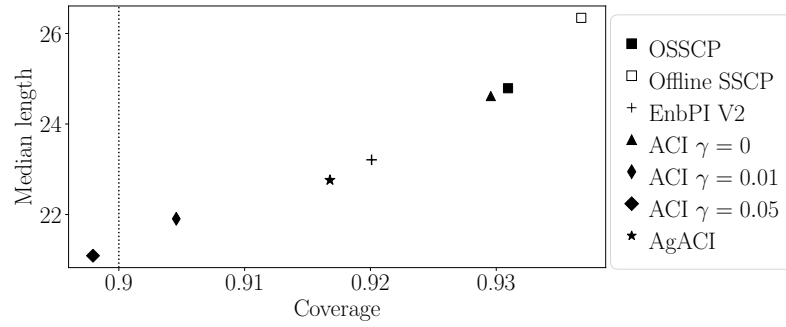


Figure 5.8: Performance of different CP methods on hourly spot electricity prices in France, trained from 2016 to 2018 and forecasted on 2019. Median length with respect to empirical coverage.

validity. Yet it slightly over-covers, and is slightly less *efficient* than ACI with a well chosen γ .

An illustration of the predicted intervals is given in the inset graphic of Figure 5.7, for AgACI, to highlight the practical relevance of such an approach on the spot prices.

However, as expected, these intervals only enjoy a *marginally* valid frequency. They do not have *conditional* guarantees. Especially, in this forecasting task, the predicted intervals cover the true prices around 88% of the time on week ends and Mondays, and 93% of the time on Tuesdays to Fridays (see Section 5.E.2). Further developments are needed to improve this unbalanced coverage.

5.7 Conclusion

This article shows why and how ACI can be used for interval prediction in the context of time series with general dependencies. We prove that ACI deteriorates *efficiency* compared to CP in the exchangeable case and analyse the dependency on γ in the AR case with the support of numerical simulations. We propose an algorithm, AgACI, based on online expert aggregation, that wraps around ACI to avoid the choice of γ . We conduct extensive experiments on synthetic time series for various strengths and structures of time dependence, demonstrating ACI's robustness and better performances than baselines, with well chosen γ or using AgACI. Finally we perform a detailed application study on the high-stakes electricity price forecasting problem in the energy transition era. Future work includes theoretical study of the proposed aggregation algorithm, including whether it preserves the asymptotic *validity* observed experimentally or to quantify its *efficiency* with respect to the performances of each expert.

Appendix to Adaptive Conformal Predictions for Time Series

The appendices are organized as follows. First, Section 5.A provides details about the Split Conformal Prediction procedure. Second, Section 5.B proves the results of Section 5.3 and conducts the numerical analysis of Section 5.3.2 in the case where the *efficiency* is computed using the median length. Then, Section 5.C contains details on the experimental setup (brief description of BOA, hyper-parameters, settings, pseudo-codes of competing algorithms). Finally, Sections 5.D and 5.E contain complementary numerical results, respectively on synthetic data sets and on the French electricity spot prices data set.

5.A Details on Split Conformal Prediction

In this section, we introduce and review the simplest theoretical properties of Split Conformal Prediction (SCP). More specifically, we present the whole algorithm, the theoretical guarantees and discuss the visualisation challenges arising when visualising a CP procedure.

5.A.1 Split Conformal Prediction Algorithm

Algorithm 10 Split Conformal Algorithm, with absolute value residuals scores

Input: Regression algorithm \mathcal{A} , significance level α , examples z_1, \dots, z_T with $z_t = (x_t, y_t)$.

Output: Prediction interval $\hat{\mathcal{C}}_\alpha(x)$ for any $x \in \mathbb{R}^d$.

- 1: Randomly split $\{1, \dots, T\}$ into two disjoint sets Tr and Cal.
 - 2: Fit a mean regression function: $\hat{\mu}(\cdot) \leftarrow \mathcal{A}(\{z_t, t \in \text{Tr}\})$
 - 3: **for** $j \in \text{Cal}$ **do**
 - 4: Set $s_j = |y_j - \hat{\mu}(x_j)|$, the *conformity scores*
 - 5: **end for**
 - 6: Set $S_{\text{Cal}} = \{s_j, j \in \text{Cal}\}$
 - 7: Compute $\hat{Q}_{1-\alpha^{\text{SCP}}}(S_{\text{Cal}})$, the $1 - \alpha^{\text{SCP}}$ -th empirical quantile of S_{Cal} , with $1 - \alpha^{\text{SCP}} := (1 - \alpha)(1 + 1/\#\text{Cal})$.
 - 8: Set $\hat{\mathcal{C}}_\alpha(x) = \left[\hat{\mu}(x) \pm \hat{Q}_{1-\alpha^{\text{SCP}}}(S_{\text{Cal}}) \right]$, for any $x \in \mathbb{R}^d$.
-

5.A.2 Illustration of the SCP procedure

Figure 5.9 provides a visualisation of the SCP procedure in a regression task. The conformity scores are taken to be the absolute value of the residuals.

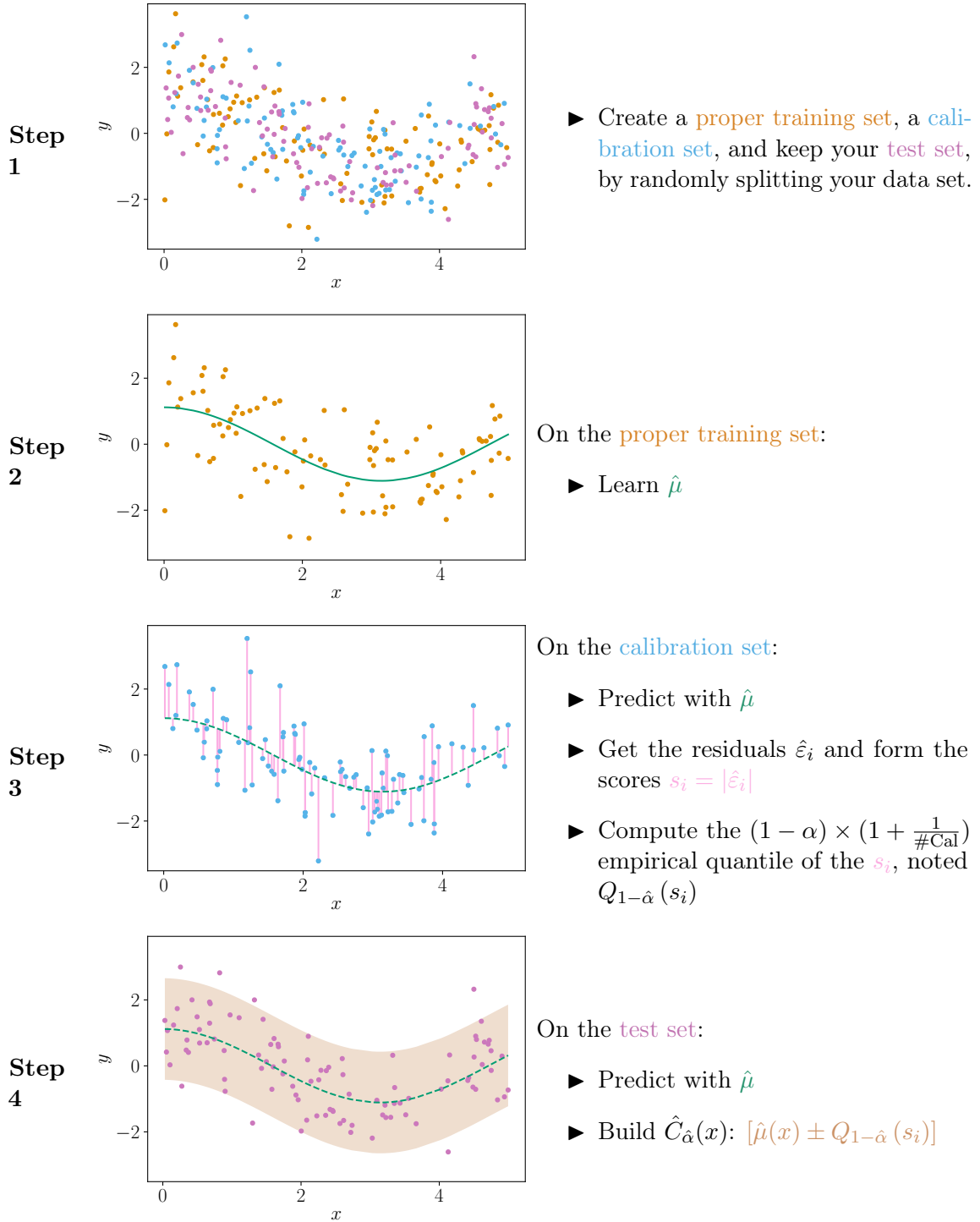


Figure 5.9: Schematic illustration of the Split Conformal Prediction procedure. Special case of a regression task, where the conformity scores are the absolute value of the residuals, as it is standard.

5.A.3 Theoretical guarantees of Split Conformal Prediction

Conformal prediction relies on the assumption that the data is exchangeable.

Definition 5.A.1 (Exchangeability). $(X_t, Y_t)_{t=1}^T$ are exchangeable if for any permutation σ of $\llbracket 1, T \rrbracket$ we have:

$$\mathcal{L}((X_1, Y_1), \dots, (X_T, Y_T)) = \mathcal{L}((X_{\sigma(1)}, Y_{\sigma(1)}), \dots, (X_{\sigma(T)}, Y_{\sigma(T)})),$$

where \mathcal{L} designates the joint distribution.

Lei et al. (2018) proves the following Theorem 5.A.1 about SCP quasi-exact *validity*.

Theorem 5.A.1. Suppose $(X_t, Y_t)_{t=1}^{T+1}$ are exchangeable, and we apply algorithm 10 on $(X_t, Y_t)_{t=1}^T$ to predict an interval on X_{T+1} , $\hat{\mathcal{C}}_\alpha(X_{T+1})$. Then we have:

$$\mathbb{P} \left\{ Y_{T+1} \in \hat{\mathcal{C}}_\alpha(X_{T+1}) \right\} \geq 1 - \alpha.$$

If, in addition, the scores S_{Cal} have a continuous joint distribution, we also have an upper bound:

$$\mathbb{P} \left\{ Y_{T+1} \in \hat{\mathcal{C}}_\alpha(X_{T+1}) \right\} \leq 1 - \alpha + \frac{1}{\#\text{Cal} + 1}.$$

5.A.4 Examples of dependent scores when data noise is exchangeable

In this subsection, we provide two examples that highlight the importance of adapting CP to time series. In these examples, the scores are non exchangeable while the true noise of the data is exchangeable.

Example 5.A.1 (Endogenous and not perfectly estimated). Assume $X_t = Y_{t-1} \in \mathbb{R}$ and that

$$Y_t = aY_{t-1} + \varepsilon_t,$$

where ε_t is a white noise. This corresponds to an order-1 Auto-Regressive (i.e. AR(1)).

Assume that the fitted model is $\hat{f}_t(x) = \hat{a}x$, with $\hat{a} \neq a$. Then, for any t , we have that:

$$\begin{aligned} \hat{\varepsilon}_t &= Y_t - \hat{Y}_t = (a - \hat{a}) Y_{t-1} + \varepsilon_t \\ \hat{\varepsilon}_t &= a\hat{\varepsilon}_{t-1} + \xi_t \end{aligned}$$

with $\xi_t = \varepsilon_t - \hat{a}\varepsilon_{t-1}$.

The residual process $(\hat{\varepsilon}_t)_{t \geq 0}$ is an ARMA(1,1) (Auto-Regressive Moving-Average, see section 5.C.2) of parameters $\varphi = a$ and $\theta = -\hat{a}$.

Thus, we have generated dependent residuals (ARMA residuals) even though the underlying model only had white noise. \square

Example 5.A.2 (Exogenous and misspecified). Assume $X_t \in \mathbb{R}^2$ and that:

$$Y_t = aX_{1,t} + bX_{2,t} + \varepsilon_t,$$

with $\varepsilon_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, $X_{2,t+1} = \varphi X_{2,t} + \xi_t$, $\xi_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $X_{1,t}$ can be any random variable.

Assume that we misspecify the model so that the fitted model is $\hat{f}_t(x) = ax_1$ for any $t \geq 0$. Then, for any $t \geq 0$, we have that

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_t = bX_{2,t} + \varepsilon_t.$$

Thus, we have generated dependent residuals (Auto-Regressive residuals) even if the underlying model only had i.i.d. Gaussian noise. \square

5.A.5 How should we visualise CP predicted intervals?

We propose to have a closer look at how are constructed the prediction of this method. In this aim, we introduce model 5.A.1.

Model 5.A.1.

$$\begin{aligned} x_t &= \cos\left(\frac{2\pi}{180}t\right) + \sin\left(\frac{2\pi}{180}t\right) + \frac{t}{100} \\ \varepsilon_{t+1} &= 0.99\varepsilon_t + \xi_{t+1}, \quad \xi_t \sim \mathcal{N}(0, 0.01) \\ Y_t &= f_t(x_t) + \varepsilon_t = x_t + \varepsilon_t \end{aligned}$$

In this model 5.A.1, the explanatory variables are deterministic. A generation from this model is represented in Figure 5.10. The first subplot, Figure 5.10a, represents x_t across time. The second subplot, Figure 5.10b, represents the noise ε_t across time. Finally, the last subplot, Figure 5.10c, represents the whole process Y_t across time.

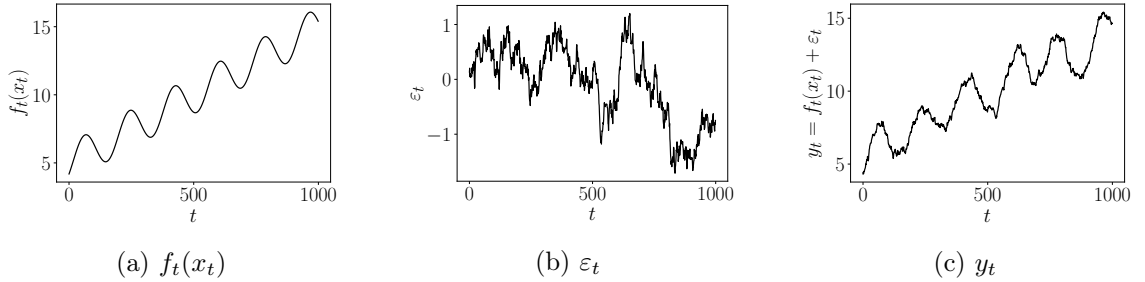


Figure 5.10: Representation of data simulated according to model 5.A.1.

The aim is to predict intervals of coverage 0.9 for values of Y_t , at $t > 500$, that is to say $T_0 = 500$ here. For simplicity, we assume $\hat{f}_t = f_t$ at each time step t and we do not represent the points used to obtain this perfect regression model. There are two ways of visualizing the predictions, that are represented in each row of Figure 5.11. If the focus of the analysis is on a specific application with the aim of analysing the whole prediction, it is relevant to represent the response y_t itself and the associated intervals. This is represented in the first row of Figure 5.11. Nevertheless, to better understand a CP method, it is relevant to represent the scores and the corresponding intervals, rescaled. This is represented in the second row of Figure 5.11 (even if the residuals are displayed and not their absolute value, i.e. the scores).

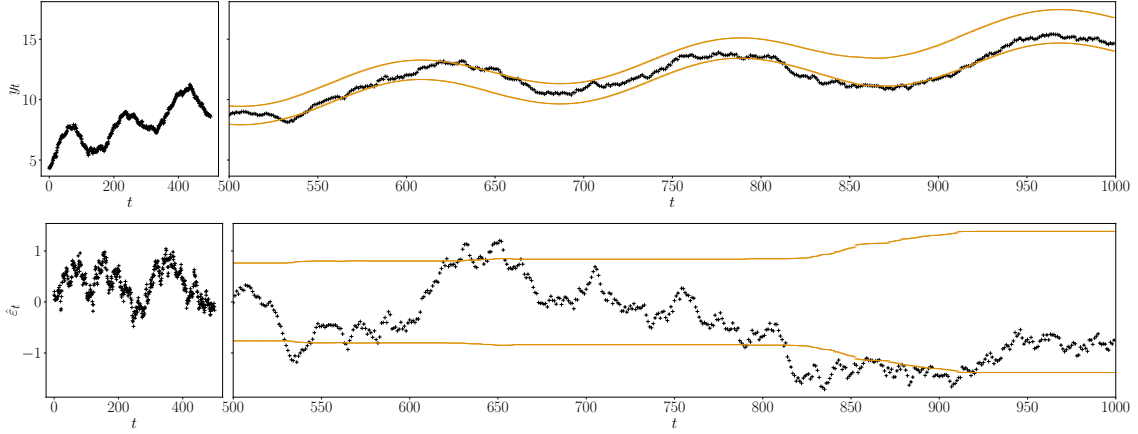


Figure 5.11: Visualisation of OSSCP on simulated data, from model model 5.A.1. 1000 data points are generated. The 500 first ones form the initial calibration set, displayed on the first subplot of each row. The 500 last ones are the ones the algorithm tries to predict. They are displayed on the right subplot of each row. Observed values are in black, predicted intervals bounds are displayed in orange

To better understand the difference between the two visualizations, let's look specifically at some observations. In the first line of the Figure 5.11, we can see that the intervals widen for $t \in [801; 900]$, while struggling to include the observations. Nevertheless, it is difficult to understand the underlying phenomenon on such a plot. Indeed, the points seem very similar to those for $t \in [660; 720]$. What considerably influences the CP are the scores and not the observed values. Thus, in the second line, at times $t \in [801; 900]$, we observe more clearly that the values go out of the previous range of values, being around 1.5 in absolute value. This explains why the intervals widen: the calibration set contains more and more high values, which increases the value of the quantile and, therefore, the length of the interval. To conclude, to analyse and assess the performances of CP procedures, we recommend representing the intervals around the *conformity scores* (or the residuals, depending on the score function) rather than the observed values. This is because the scores are what truly determine the conformal behaviour.

5.B Proof of the results presented in Section 5.3 and additional numerical experiments

5.B.1 Proof of Theorem 5.3.1

We recall here Theorem 5.3.1.

Theorem 5.3.1. *Assume that: (i) $\alpha \in \mathbb{Q}$; (ii) the scores are exchangeable with quantile function Q ; (iii) the quantile function is perfectly estimated at each time (as defined above); (iv) the quantile function Q is bounded and $\mathcal{C}^4([0, 1])$. Then, for all $\gamma > 0$, $(\alpha_t)_{t>0}$ forms a Markov Chain, that admits a stationary distribution π_γ , and*

$$\frac{1}{T} \sum_{t=1}^T L(\alpha_t) \xrightarrow[T \rightarrow +\infty]{a.s.} \mathbb{E}_{\pi_\gamma}[L] \stackrel{not.}{=} \mathbb{E}_{\tilde{\alpha} \sim \pi_\gamma}[L(\tilde{\alpha})].$$

Moreover, as $\gamma \rightarrow 0$,

$$\mathbb{E}_{\pi_\gamma}[L] = L_0 + Q''(1 - \alpha) \frac{\gamma}{2} \alpha(1 - \alpha) + O(\gamma^{3/2}).$$

To prove Theorem 5.3.1, we rely on the following lemmas, that will be proved after the theorem. We denote B_β a Bernoulli random variable of parameter β and $P(x)$ designates the projection of x onto $[0, 1]$. Finally, for $\gamma > 0$, define the following Markov Chain:

$$\alpha_{t+1} = \alpha_t + \gamma (\alpha - B_{P(\alpha_t)}) \text{ for } t > 0, \quad (5.4)$$

We introduce $(p, q) \in \mathbb{N} \times \mathbb{N}^*$, $p < q$, s.t. $\alpha = \frac{p}{q}$, and:

$$\mathcal{A} = \left\{ \alpha + \gamma \frac{\gcd(q-p, p)}{q} \mathbb{Z} \right\} \cap]\gamma(\alpha - 1), 1 + \gamma\alpha[. \quad (5.5)$$

Lemma 5.B.1 (Finite state space). *Assume that $\alpha \in \mathbb{Q}$. Then, for any $\gamma > 0$, the Markov Chain defined by $\alpha_1 \in \mathcal{A}$ and $\alpha_{t+1} = \alpha_t + \gamma (\alpha - B_{P(\alpha_t)})$, for $t > 0$ has a finite state space \mathcal{A} .*

Lemma 5.B.2 (Irreducibility). *Assume that $\alpha \in \mathbb{Q}$. Then, for any $\gamma > 0$, the Markov Chain defined by Equation (5.4), for $t > 0$ and $\alpha_1 \in \mathcal{A}$, is irreducible.*

Thereby we will prove that the chain admits a unique stationary distribution π_γ , we now compute the first four moments of the stationary distribution in Lemmas 5.B.3 to 5.B.6. The final proof relies on a Taylor expansion, that requires to control these four moments.

Lemma 5.B.3 (Expectation). *Let $\gamma > 0$ and consider again the Markov Chain defined in equation (5.4). We have:*

$$\mathbb{E}_{\pi_\gamma} [(P(\tilde{\alpha}) - \alpha)] = 0.$$

Lemma 5.B.4 (Second order moment). *Let $\gamma > 0$ and consider again the Markov Chain defined in equation (5.4). As $\gamma \rightarrow 0$, we have:*

$$\mathbb{E}_{\pi_\gamma} [(P(\tilde{\alpha}) - \alpha)^2] = \frac{\gamma}{2} \alpha(1 - \alpha) + O(\gamma^2).$$

Lemma 5.B.5 (Third order moment). *Let $\gamma > 0$ and consider again the Markov Chain defined in equation (5.4). As $\gamma \rightarrow 0$, we have:*

$$\mathbb{E}_{\pi_\gamma} [(P(\tilde{\alpha}) - \alpha)^3] = O(\gamma^{3/2}).$$

Lemma 5.B.6 (Fourth order moment). *Let $\gamma > 0$ and consider again the Markov Chain defined in equation (5.4). As $\gamma \rightarrow 0$, we have:*

$$\mathbb{E}_{\pi_\gamma} [(P(\tilde{\alpha}) - \alpha)^4] = O(\gamma^2).$$

The proofs of these Lemmas are postponed to Sections 5.B.2 and 5.B.3. Here, we first give the proof of the main theorem.

Proof of Theorem 5.3.1. Let $\gamma > 0$. For any $t > 0$ we have, for the recursion introduced in Equation (5.2), that

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1}_{y_t \notin \hat{C}_{\alpha_t}(x_t)} \right) = \alpha_t + \gamma \left(\alpha - \mathbb{1}_{S_t > \hat{Q}_{1-P(\alpha_t)}} \right),$$

where S_t is the conformity score at time t . Noting that $\mathbb{1}_{S_t > \hat{Q}_t(1-P(\alpha_t))} \stackrel{d}{=} B_{\mathbb{P}(S_t > \hat{Q}_t(1-P(\alpha_t)))}$, we obtain:

$$\begin{aligned} \alpha_{t+1} &\stackrel{d}{=} \alpha_t + \gamma \left(\alpha - B_{\mathbb{P}(S_t > \hat{Q}_t(1-P(\alpha_t)))} \right) \\ &\stackrel{d}{=} \alpha_t + \gamma \left(\alpha - B_{\mathbb{P}(S_t > Q(1-P(\alpha_t)))} \right) \\ &\stackrel{d}{=} \alpha_t + \gamma \left(\alpha - B_{P(\alpha_t)} \right), \end{aligned}$$

where the second line results from assumption (ii) and (iii), and the last equation from assumption (iii) only. Consequently, by induction, the chain defined by Equation (5.2) and

$$\alpha_{t+1} = \alpha_t + \gamma \left(\alpha - B_{P(\alpha_t)} \right), \quad (5.6)$$

with $\alpha_1 = \alpha$, have the same distribution.

Using assumption (i), Lemma 5.B.1 ensures that the state space \mathcal{A} of the Markov Chain defined in equation (5.6) is finite. Furthermore, Lemma 5.B.2 also ensures that the chain is irreducible. Therefore, the chain is irreducible on a finite state space, thus it admits a unique stationary distribution, noted π_γ and for any positive function f such that $\int f d\pi_\gamma < \infty$, we have (Meyn and Tweedie, 2012, Theorem 17.1.7):

$$\frac{1}{T} \sum_{t=1}^T f(\alpha_t) \xrightarrow[T \rightarrow \infty]{a.s.} \int f d\pi_\gamma.$$

Remark that $L(\beta) = 2Q(1 - P(\beta))$ for any β . Therefore, combined with previous result we get the first result of Theorem 5.3.1:

$$\frac{1}{T} \sum_{t=1}^T L(\alpha_t) \xrightarrow[T \rightarrow +\infty]{a.s.} \mathbb{E}_{\tilde{\alpha} \sim \pi_\gamma} [L(\tilde{\alpha})].$$

We now need to characterize $\mathbb{E}_{\tilde{\alpha} \sim \pi_\gamma} [L(\tilde{\alpha})] = 2\mathbb{E}_{\tilde{\alpha} \sim \pi_\gamma} [Q(1 - P(\tilde{\alpha}))]$ as $\gamma \rightarrow 0$. Assume that $Q \in \mathcal{C}^4([0, 1])$. Using Taylor series expansion, for any $\tilde{\alpha} \in \mathcal{A}$, there exists $\beta(\tilde{\alpha}) \in [0, 1]$:

$$\begin{aligned} Q(1 - P(\tilde{\alpha})) &= Q(1 - \alpha) + Q'(1 - \alpha)(\alpha - P(\tilde{\alpha})) + \frac{Q''(1 - \alpha)}{2}(\alpha - P(\tilde{\alpha}))^2 \\ &\quad + \frac{Q'''(1 - \alpha)}{6}(\alpha - P(\tilde{\alpha}))^3 + \frac{Q''''(1 - \beta(\tilde{\alpha}))}{24}(\alpha - P(\tilde{\alpha}))^4. \end{aligned} \quad (5.7)$$

To conclude, we take the expectation under π_γ of equation (5.7), which gives:

$$\begin{aligned} \mathbb{E}_{\pi_\gamma} [Q(1 - P(\tilde{\alpha}))] &= Q(1 - \alpha) + Q'(1 - \alpha)\mathbb{E}_{\pi_\gamma} [(\alpha - P(\tilde{\alpha}))] \\ &\quad + \frac{Q''(1 - \alpha)}{2}\mathbb{E}_{\pi_\gamma} [(\alpha - P(\tilde{\alpha}))^2] + \frac{Q'''(1 - \alpha)}{6}\mathbb{E}_{\pi_\gamma} [(\alpha - P(\tilde{\alpha}))^3] \\ &\quad + \mathbb{E}_{\pi_\gamma} \left[\frac{Q''''(1 - \beta(\tilde{\alpha}))}{24}(\alpha - P(\tilde{\alpha}))^4 \right]. \end{aligned} \quad (5.8)$$

Injecting results of Lemmas 5.B.3 to 5.B.5 in equation (5.8), we obtain:

$$\begin{aligned} \mathbb{E}_{\pi_\gamma} [Q(1 - P(\tilde{\alpha}))] &= Q(1 - \alpha) + \frac{Q''(1 - \alpha)}{4} \gamma \alpha (1 - \alpha) + O(\gamma^{3/2}) \\ &\quad + \mathbb{E}_{\pi_\gamma} \left[\frac{Q'''(1 - \beta(\tilde{\alpha}))}{24} (\alpha - P(\tilde{\alpha}))^4 \right]. \end{aligned} \quad (5.9)$$

Finally, we can control the last term since $Q \in \mathcal{C}^4([0, 1])$ by assumption, thus there exists $M > 0$ such that for any $x \in [0, 1]$, $|Q'''(1 - x)| < M$. Hence, using Lemma 5.B.6 we obtain:

$$\begin{aligned} |\mathbb{E}_{\pi_\gamma} [Q'''(1 - \beta(\tilde{\alpha}))(\alpha - P(\tilde{\alpha}))^4]| &\leq \mathbb{E}_{\pi_\gamma} [|Q'''(1 - \beta(\tilde{\alpha}))| (\alpha - P(\tilde{\alpha}))^4] \\ &\leq M \mathbb{E}_{\pi_\gamma} [(\alpha - P(\tilde{\alpha}))^4] \\ &\leq MO(\gamma^{3/2}) \\ \mathbb{E}_{\pi_\gamma} [Q'''(1 - \beta(\tilde{\alpha}))(\alpha - P(\tilde{\alpha}))^4] &= O(\gamma^{3/2}). \end{aligned} \quad (5.10)$$

Finally, combining equations (5.9) and (5.10) to conclude the proof by obtaining:

$$\mathbb{E}_{\pi_\gamma} [Q(1 - P(\tilde{\alpha}))] = Q(1 - \alpha) + \frac{Q''(1 - \alpha)}{4} \gamma \alpha (1 - \alpha) + O(\gamma^{3/2}). \quad (5.11)$$

□

This concludes the proof of Theorem 5.3.1.

Remark: is it possible to use only 3 moments? The proof here relies on the control of the first four moments. It is not clear that the same result could be obtained using only a third order Taylor expansion, as we would then require a bound on $\mathbb{E}[|P(\tilde{\alpha}) - \alpha|^3]$, which is *not* guaranteed to be $O(\gamma^{3/2})$, contrary to $\mathbb{E}[(P(\tilde{\alpha}) - \alpha)^3]$.

5.B.2 Proof of Lemmas 5.B.1 and 5.B.2

Proof of Lemma 5.B.1. Let $\gamma > 0$ and denote $\alpha = \frac{p}{q}$ with $0 < p < q$ and $(p, q) \in \mathbb{N}^2$. We denote E the state space of the Markov Chain defined by equation (5.6), starting from $a \in \mathcal{A}$. We show that $E = \mathcal{A}$.

First, (α_t) is strictly bounded by $\gamma(\alpha - 1)$ and $1 + \gamma\alpha$. Thus $E \subset]\gamma(\alpha - 1), \gamma\alpha[$. Secondly, for any starting point $\alpha_1 \in \mathcal{A}$, we can observe that:

$$\begin{aligned} \{\alpha_t, t \geq 1\} &\stackrel{a.s.}{\subset} \alpha_1 + \{k\gamma(\alpha - 1) + n\gamma\alpha, (k, n) \in \mathbb{N}^2\} \\ &\subset \alpha_1 + \{k\gamma(\alpha - 1) + n\gamma\alpha, (k, n) \in \mathbb{Z}^2\} \\ &= \alpha_1 + \{k\gamma \frac{p-q}{q} + n\gamma \frac{p}{q}, (k, n) \in \mathbb{Z}^2\} \\ &= \alpha_1 + \frac{\gamma}{q} \{(q-p)\mathbb{Z} + p\mathbb{Z}\} \\ &= \alpha_1 + \frac{\gamma}{q} \gcd(q-p, p)\mathbb{Z} \\ &= \alpha + \frac{\gamma}{q} \gcd(q-p, p)\mathbb{Z} \end{aligned}$$

where $\gcd(a, b)$ is the greatest common divisor of a and b . We have used at the last line that $\alpha_1 \in \mathcal{A}$ writes as $\alpha + \frac{\gamma}{q} \gcd(q-p, p)k$, for some $k \in \mathbb{Z}$. Combining both results, we get that:

$$E \subset \left\{ \alpha + \frac{\gamma}{q} \gcd(q-p, p)\mathbb{Z} \right\} \cap]\gamma(\alpha - 1), \gamma\alpha[.$$

This shows that the state space is finite and a subset of \mathcal{A} . The reciprocal implication is proved in the following Lemma, together with irreducibility. \square

Proof of Lemma 5.B.2. Our objective is to show that there is a path of positive probability going from any point of the state space \mathcal{A} to any point of the same state space \mathcal{A} . Note that the chain always has at most two options when on a state x : make a step $\gamma\alpha$, with probability $1 - P(x)$, or a step $\gamma(\alpha - 1)$, with probability $P(x)$.

Let $(x, y) \in \mathcal{A}^2$. Thereby, there exist $(k, n), (l, m) \in \mathbb{N}^2$ such that:

$$\begin{aligned} x &= \alpha + k\gamma\alpha + n\gamma(\alpha - 1) \\ y &= \alpha + l\gamma\alpha + m\gamma(\alpha - 1). \end{aligned}$$

Thus, starting from x , to attain y , the chain has to make the path $y - x = (l - k)\gamma\alpha + (m - n)\gamma(\alpha - 1)$.

Noting that for any $h \in \mathbb{N}$ we have $\gamma\alpha(q - p)h + \gamma(\alpha - 1)hp = 0$, we can equivalently write that:

$$y - x = u\gamma\alpha + v\gamma(\alpha - 1), \quad (5.12)$$

with $(u, v) \in \mathbb{N}^2 \setminus \{(0, 0)\}$.

Thus, for any $(x, y) \in \mathcal{A}^2$ there exists $(u, v) \in \mathbb{N}^2 \setminus \{(0, 0)\}$ such that $y - x = u\gamma\alpha + v\gamma(\alpha - 1)$.

Let's show by induction on $u + v$ that for any $(u, v) \in \mathbb{N}^2$, and $(x, y) \in \mathcal{A}^2$ satisfying Equation (5.12) there exists a path of strictly positive probability between x and y .

Initialization. Suppose first that $u + v = 1$. Then, there are two options: $u = 1$ and $v = 0$ or the reverse. Assume the former: Equation (5.12) gives $y = x + \gamma\alpha$ and necessarily $x < 1$ since $y < 1 + \gamma\alpha$ because $y \in \mathcal{A}$. Thereby the step $\gamma\alpha$ has a probability $1 - P(x) > 0$ to occur. Thus the chain can attain y starting from x , i.e., $\mathbb{P}(\alpha_2 = y | \alpha_1 = x) > 0$. The second case works similarly, by observing that necessarily $x > 0$.

Heredity. Let $m \in \mathbb{N}$. We assume that for any $(u, v) \in \mathbb{N}^2$ such that $u + v = m$, and $(x, y) \in \mathcal{A}^2$ satisfying Equation (5.12) there exists a path of strictly positive probability between x and y , or formally there exists $t \in \mathbb{N}$ such that $\mathbb{P}(\alpha_t = y | \alpha_1 = x) > 0$.

Suppose now that $u + v = m + 1$ with $m \in \mathbb{N}^*$. If $v = 0$, then $y = x + u\gamma\alpha$ and similarly than for $v = 0$ and $u = 1$, the step $\gamma\alpha$ is probable. Let $z = x + \gamma\alpha$. We have:

- $\mathbb{P}(\alpha_2 = z | \alpha_1 = x) = 1 - P(x) > 0$.
- By our induction hypothesis, (y, z) satisfy Eq. 5.12 with $u + v = m$, thus there exists t such that $\mathbb{P}(\alpha_t = y | \alpha_2 = z) > 0$.

Overall, $\mathbb{P}(\alpha_t = y | \alpha_1 = x) > 0$.

If instead $u = 0$, then $y = x + v\gamma(\alpha - 1)$ and as for $u = 0$ and $v = 1$, the step $\gamma\alpha$ is of strictly positive probability and we conclude similarly.

Finally, if both u and v are non-null, then we can make the step $\gamma(\alpha - 1)$ if $x > 0$ and the step $\gamma\alpha$ otherwise, before using our induction hypothesis.

This shows that we can build a path of strictly positive probability for any $(x, y) \in \mathcal{A}^2$, and thereby that the chain is irreducible. \square

5.B.3 Control of the first four moments: Lemmas 5.B.3 to 5.B.6

In the following Lemmas, to compute the first order moments of π_γ , we consider the chain $\alpha_{t+1} = \alpha_t + \gamma(\alpha - B_{P(\alpha_t)})$ for $t > 0$, launched from the stationary distribution $\alpha_1 \sim \pi_\gamma$. Thanks to the stationarity property, for all $t \geq 1$, $\alpha_t \sim \pi_\gamma$.

Proof of Lemma 5.B.3. Let $\gamma > 0$. To derive $\mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)]$ we start by equation (5.6) with $t = 1$:

$$\begin{aligned} \alpha_2 &= \alpha_1 + \gamma(\alpha - B_{P(\alpha_1)}) \\ \mathbb{E}[\alpha_2] &= \mathbb{E}[\alpha_1] + \gamma(\alpha - \mathbb{E}[B_{P(\alpha_1)}]) \quad \text{taking expectation} \\ 0 &= \gamma(\alpha - \mathbb{E}_{\pi_\gamma}[B_{P(\alpha_1)}]) \quad \text{using } \mathbb{E}[\alpha_1] = \mathbb{E}[\alpha_2] = \mathbb{E}_{\pi_\gamma}[\alpha] \\ \mathbb{E}_{\pi_\gamma}[\mathbb{E}[B_{P(\alpha_1)}|\alpha_1]] &= \alpha \\ \mathbb{E}_{\pi_\gamma}[P(\alpha_1)] &= \alpha. \end{aligned}$$

□

Proof of Lemma 5.B.4. Let $\gamma > 0$. To derive $\mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)^2]$ we start by equation (5.6) with $t = 1$:

$$\begin{aligned} (\alpha_2 - \alpha)^2 &= (\alpha_1 - \alpha)^2 + \gamma^2(\alpha - B_{P(\alpha_1)})^2 + 2\gamma(\alpha - B_{P(\alpha_1)})(\alpha_1 - \alpha) \\ \mathbb{E}_{\pi_\gamma}[(\alpha_2 - \alpha)^2] &= \mathbb{E}_{\pi_\gamma}[(\alpha_1 - \alpha)^2] + \gamma^2\mathbb{E}_{\pi_\gamma}[(\alpha - B_{P(\alpha_1)})^2] \\ &\quad + 2\gamma\mathbb{E}_{\pi_\gamma}[(\alpha - B_{P(\alpha_1)})(\alpha_1 - \alpha)] \\ 0 &= \gamma^2\mathbb{E}_{\pi_\gamma}[(\alpha - B_{P(\alpha_1)})^2] + 2\gamma\mathbb{E}_{\pi_\gamma}[(\alpha - P(\alpha_1))(\alpha_1 - \alpha)] \end{aligned}$$

Consequently,

$$\begin{aligned} 2\gamma\mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)(\alpha_1 - P(\alpha_1) + P(\alpha_1) - \alpha)] &= \\ \gamma^2\mathbb{E}_{\pi_\gamma}[(\alpha - B_{P(\alpha_1)} + P(\alpha_1) - P(\alpha_1))^2] &= \\ \implies 2\gamma\mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)^2] - 2\gamma\mathbb{E}_{\pi_\gamma}[(\alpha - P(\alpha_1))(\alpha_1 - P(\alpha_1))] &= \\ \gamma^2\mathbb{E}_{\pi_\gamma}[(\alpha - B_{P(\alpha_1)} + P(\alpha_1) - P(\alpha_1))^2] &= \\ \implies (2 - \gamma)\mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)^2] = \gamma\mathbb{E}_{\pi_\gamma}[P(\alpha_1)(1 - P(\alpha_1))] & \quad (5.13) \\ + 2\mathbb{E}_{\pi_\gamma}[(\alpha - P(\alpha_1))(\alpha_1 - P(\alpha_1))] & \end{aligned}$$

We can compute $\mathbb{E}_{\pi_\gamma}[P(\alpha_1)(1 - P(\alpha_1))]$:

$$\begin{aligned} \mathbb{E}_{\pi_\gamma}[P(\alpha_1)(1 - P(\alpha_1)) - \alpha(1 - \alpha)] &= \\ = \mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)(1 - P(\alpha_1)) + \alpha(1 - P(\alpha_1)) - \alpha(1 - \alpha)] &= \\ = \mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)(1 - P(\alpha_1)) + \alpha(\alpha - P(\alpha_1))] &= \\ = \mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)(1 - P(\alpha_1) - \alpha)] &= \\ = \mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)(\alpha - P(\alpha_1) + 1 - 2\alpha)] &= \\ = -\mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)^2] + \mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)(1 - 2\alpha)] &= \\ = -\mathbb{E}_{\pi_\gamma}[(P(\alpha_1) - \alpha)^2] & \end{aligned}$$

$$\Rightarrow \mathbb{E}_{\pi_\gamma} [P(\alpha_1)(1 - P(\alpha_1))] = \alpha(1 - \alpha) - \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] \quad (5.14)$$

Reinjecting equation (5.14) in equation (5.13):

$$\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] = \frac{\gamma}{2}\alpha(1 - \alpha) + \mathbb{E}_{\pi_\gamma} [(\alpha - P(\alpha_1))(\alpha_1 - P(\alpha_1))] \quad (5.15)$$

We are now going to derive an upper and lower bound of $\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2]$. Note that $\text{sign}(\alpha - P(\alpha_1)) = -\text{sign}(\alpha_1 - P(\alpha_1))$, thus $\mathbb{E}_{\pi_\gamma} [(\alpha - P(\alpha_1))(\alpha_1 - P(\alpha_1))] \leq 0$. Hence we obtain the following upper bound:

$$\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] \leq \frac{\gamma}{2}\alpha(1 - \alpha). \quad (5.16)$$

Furthermore, using again this observation, and additionally that $|\alpha - P(\alpha_1)| \leq 1$ and $|\alpha_1 - P(\alpha_1)| \leq \gamma$ and from equation (5.15), we can obtain:

$$\begin{aligned} \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] &\geq \frac{\gamma}{2}\alpha(1 - \alpha) - \gamma \mathbb{P}_{\pi_\gamma}(\alpha_1 \notin [0, 1]) \\ &\geq \frac{\gamma}{2}\alpha(1 - \alpha) - \gamma C_\alpha^{-1} \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] \\ \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] &\geq \frac{1}{1 + \gamma C_\alpha^{-1}} \frac{\gamma}{2}\alpha(1 - \alpha), \end{aligned} \quad (5.17)$$

where the second inequality holds by observing that:

$$\begin{aligned} \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] &\geq (1 - \alpha)^2 \mathbb{P}_{\pi_\gamma}(\alpha_1 > 1) + \alpha^2 \mathbb{P}_{\pi_\gamma}(\alpha_1 < 0) \\ \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] &\geq C_\alpha \mathbb{P}_{\pi_\gamma}(\alpha_1 \notin [0, 1]) \\ \Rightarrow \mathbb{P}_{\pi_\gamma}(\alpha_1 \notin [0, 1]) &\leq C_\alpha^{-1} \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] \end{aligned}$$

with $C_\alpha = \min(\alpha^2, (1 - \alpha)^2)$.

Gathering equations (5.16) and (5.17), we obtain:

$$\begin{aligned} \frac{1}{(1 + \gamma C_\alpha^{-1})} \frac{\gamma}{2}\alpha(1 - \alpha) &\leq \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] \leq \frac{\gamma}{2}\alpha(1 - \alpha) \\ \left(\frac{1}{(1 + \gamma C_\alpha^{-1})} - 1 \right) \frac{\gamma}{2}\alpha(1 - \alpha) &\leq \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] - \frac{\gamma}{2}\alpha(1 - \alpha) \leq 0 \\ \Rightarrow \left| \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] - \frac{\gamma}{2}\alpha(1 - \alpha) \right| &\leq \frac{\gamma^2 C_\alpha^{-1}}{2(1 + \gamma C_\alpha^{-1})} \alpha(1 - \alpha) \\ \Rightarrow \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] - \frac{\gamma}{2}\alpha(1 - \alpha) &= O(\gamma^2). \end{aligned} \quad (5.18)$$

□

Proof of Lemma 5.B.5. Let $\gamma > 0$. We start again by using equation (5.6) and removing the first terms as $\mathbb{E}_{\pi_\gamma} [(\alpha_2 - \alpha)^3] = \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^3]$. Then we will isolate $\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^3]$ and finally we will dominate each term obtained.

$$\begin{aligned} 0 &= 3\gamma \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - B_{P(\alpha_1)})] + 3\gamma^2 \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - B_{P(\alpha_1)})^2] \\ &\quad + \gamma^3 \mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^3] \\ 0 &= 3\gamma \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - P(\alpha_1))] + 3\gamma^2 \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - P(\alpha_1))^2] \\ &\quad + 6\gamma^2 \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - P(\alpha_1))(P(\alpha_1) - B_{P(\alpha_1)})] \\ &\quad + 3\gamma^2 \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(P(\alpha_1) - B_{P(\alpha_1)})^2] + \gamma^3 \mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^3]. \end{aligned}$$

Hence,

$$\begin{aligned}
3\gamma\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^3] &= 3\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))] \\
&\quad + 6\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)(\alpha - P(\alpha_1))] \\
&\quad + 3\gamma^2\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - P(\alpha_1))^2] \\
&\quad + 3\gamma^2\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)P(\alpha_1)(1 - P(\alpha_1))] \\
&\quad + \gamma^3\mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^3] \\
3\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^3] &= 3\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))] \\
&\quad - 6\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2] \\
&\quad + 3\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - P(\alpha_1))^2] \\
&\quad + 3\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)P(\alpha_1)(1 - P(\alpha_1))] \\
&\quad + \gamma^2\mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^3] \\
3|\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^3]| &\leq 3|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))]| \\
&\quad + 6|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2]| \\
&\quad + 3\gamma|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - P(\alpha_1))^2]| \\
&\quad + 3\gamma|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)P(\alpha_1)(1 - P(\alpha_1))]| \\
&\quad + \gamma^2|\mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^3]|. \tag{5.19}
\end{aligned}$$

To conclude, we can bound each term of the right hand side of equation (5.19). In order of appearance we obtain:

$$\begin{aligned}
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))]| &\leq \mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2 |\alpha - P(\alpha_1)|] \\
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))]| &\leq \gamma^2. \tag{5.20}
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2]| &\leq \mathbb{E}_{\pi_\gamma} [|\alpha_1 - P(\alpha_1)| (P(\alpha_1) - \alpha)^2] \\
&\leq \gamma\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] \\
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2]| &\leq \frac{\gamma^2}{2}\alpha(1 - \alpha) + O(\gamma^3), \tag{5.21}
\end{aligned}$$

where the last equality is obtained by using Lemma 5.B.4.

$$\begin{aligned}
\gamma|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - P(\alpha_1))^2]| &\leq \gamma\mathbb{E}_{\pi_\gamma} [|\alpha_1 - \alpha| (\alpha - P(\alpha_1))^2] \\
&\leq \gamma D_{\gamma,\alpha}\mathbb{E}_{\pi_\gamma} [(\alpha - P(\alpha_1))^2] \\
\gamma|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - P(\alpha_1))^2]| &\leq D_{\gamma,\alpha}\frac{\gamma^2}{2}\alpha(1 - \alpha) + O(\gamma^3), \tag{5.22}
\end{aligned}$$

again using Lemma 5.B.4, and with $D_{\gamma,\alpha} = \max(1 + \gamma\alpha, \gamma(1 - \alpha)) - \alpha = O(1)$.

$$\begin{aligned}
 & \gamma \left| \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)P(\alpha_1)(1 - P(\alpha_1))] \right| \\
 & \leq \gamma \left| \mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))P(\alpha_1)(1 - P(\alpha_1))] \right| \\
 & \quad + \gamma \left| \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)P(\alpha_1)(1 - P(\alpha_1))] \right| \\
 & \leq \gamma \frac{1}{4} \mathbb{E}_{\pi_\gamma} [|\alpha_1 - P(\alpha_1)|] + \gamma \frac{1}{4} \mathbb{E}_{\pi_\gamma} [|P(\alpha_1) - \alpha|] \\
 & \leq \frac{\gamma^2}{4} + \frac{\gamma}{4} \sqrt{\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2]} \\
 & \leq \frac{\gamma^2}{4} + \frac{\gamma}{4} \sqrt{\frac{\gamma}{2} \alpha(1 - \alpha) + O(\gamma^2)} \\
 & \implies \gamma \left| \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)P(\alpha_1)(1 - P(\alpha_1))] \right| \leq O(\gamma^{3/2}), \tag{5.23}
 \end{aligned}$$

where the last inequality comes from Lemma 5.B.4 a third time.

$$\gamma^2 \left| \mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^3] \right| \leq \gamma^2 \max(\alpha^3, (1 - \alpha)^3). \tag{5.24}$$

Gathering equations (5.20) to (5.24) together with equation (5.19), we obtain the following upper bound:

$$\begin{aligned}
 3 \left| \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^3] \right| & \leq 3\gamma^2 + 3\gamma^2 \alpha(1 - \alpha) + O(\gamma^3) + 3D_{\gamma, \alpha} \frac{\gamma^2}{2} \alpha(1 - \alpha) \\
 & \quad + O(\gamma^3) + O(\gamma^{3/2}) + \gamma^2 \max(\alpha^3, (1 - \alpha)^3),
 \end{aligned}$$

which leads to:

$$\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^3] = O(\gamma^{3/2}). \tag{5.25}$$

□

Proof of Lemma 5.B.6. Let $\gamma > 0$. For the fourth order moment, the proof works in the same way for the third order moment, Lemma 5.B.5.

$$\begin{aligned}
 0 & = 4\gamma \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^3(\alpha - B_{P(\alpha_1)})] + 6\gamma^2 \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - B_{P(\alpha_1)})^2] \\
 & \quad + 4\gamma^3 \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - B_{P(\alpha_1)})^3] + \gamma^4 \mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^4] \\
 0 & = 4\gamma \mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1) + P(\alpha_1) - \alpha)^3(\alpha - P(\alpha_1))] \\
 & \quad + 6\gamma^2 \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - P(\alpha_1) + P(\alpha_1) - B_{P(\alpha_1)})^2] \\
 & \quad + 4\gamma^3 \mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - B_{P(\alpha_1)})^3] + \gamma^4 \mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^4]
 \end{aligned}$$

$$\begin{aligned}
4\gamma\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^4] &= 4\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))] \\
&\quad + 12\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(P(\alpha_1) - \alpha)(\alpha - P(\alpha_1))] \\
&\quad + 12\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2(\alpha - P(\alpha_1))] \\
&\quad + 6\gamma^2\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - P(\alpha_1))^2] \\
&\quad + 0 + 6\gamma^2\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(P(\alpha_1) - B_{P(\alpha_1)})^2] \\
&\quad + 4\gamma^3\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - B_{P(\alpha_1)})^3] + \gamma^4\mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^4] \\
4\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^4] &= 4\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))] \\
&\quad - 12\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(P(\alpha_1) - \alpha)^2] \\
&\quad - 12\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^3] \\
&\quad + 6\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - P(\alpha_1))^2] \\
&\quad + 6\gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2P(\alpha_1)(1 - P(\alpha_1))] \\
&\quad + 4\gamma^2\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - B_{P(\alpha_1)})^3] + \gamma^3\mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^4] \\
4|\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^4]| &\leq 4|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))]| \\
&\quad + 12|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(P(\alpha_1) - \alpha)^2]| \\
&\quad + 12|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^3]| \\
&\quad + 6\gamma|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - P(\alpha_1))^2]| \\
&\quad + 6\gamma|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2P(\alpha_1)(1 - P(\alpha_1))]| \\
&\quad + 4\gamma^2|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - B_{P(\alpha_1)})^3]| \\
&\quad + \gamma^3|\mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^4]|. \tag{5.26}
\end{aligned}$$

We are now going to dominate each term of the right hand side of equation (5.26) in order of appearance.

$$\begin{aligned}
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))]| &\leq \mathbb{E}_{\pi_\gamma} [|\alpha_1 - P(\alpha_1)|^3 |\alpha - P(\alpha_1)|] \\
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))]| &\leq \gamma^3 \tag{5.27}
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(P(\alpha_1) - \alpha)^2]| &= \mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(P(\alpha_1) - \alpha)^2] \\
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2(P(\alpha_1) - \alpha)^2]| &\leq \gamma^2. \tag{5.28}
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^3]| &\leq \mathbb{E}_{\pi_\gamma} [|\alpha_1 - P(\alpha_1)| |P(\alpha_1) - \alpha|^3] \\
&\leq \gamma\mathbb{E}_{\pi_\gamma} [|P(\alpha_1) - \alpha|^3] \\
|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^3]| &\leq O(\gamma^{5/2}). \tag{5.29}
\end{aligned}$$

where the last inequality holds using Lemma 5.B.5.

$$\begin{aligned}
\gamma|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - P(\alpha_1))^2]| &= \gamma\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - P(\alpha_1))^2] \\
&\leq \gamma D_{\gamma,\alpha}^2 \left(\frac{\gamma}{2} \alpha(1 - \alpha) + O(\gamma^2) \right) \\
\gamma|\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2(\alpha - P(\alpha_1))^2]| &\leq D_{\gamma,\alpha}^2 \frac{\gamma^2}{2} \alpha(1 - \alpha) + O(\gamma^3). \tag{5.30}
\end{aligned}$$

again where we've used Lemma 5.B.5, and re-used its notation

$$D_{\gamma,\alpha} = \max(1 + \gamma\alpha, \gamma(1 - \alpha)) - \alpha = O(1).$$

$$\begin{aligned} \gamma |\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2 P(\alpha_1)(1 - P(\alpha_1))]| &= \gamma |\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2 P(\alpha_1)(1 - P(\alpha_1))]| \\ &\quad + 2\mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)P(\alpha_1)(1 - P(\alpha_1))] \\ &\quad + \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2 P(\alpha_1)(1 - P(\alpha_1))] | \\ &\leq \frac{\gamma}{4} \mathbb{E}_{\pi_\gamma} [(\alpha_1 - P(\alpha_1))^2] + \frac{\gamma}{2} \mathbb{E}_{\pi_\gamma} [|\alpha_1 - P(\alpha_1)| |P(\alpha_1) - \alpha|] \\ &\quad + \frac{\gamma}{4} \mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^2] \end{aligned}$$

Hence we have

$$\gamma |\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)^2 P(\alpha_1)(1 - P(\alpha_1))]| \leq \frac{\gamma^3}{4} + \frac{\gamma^2}{2} + \frac{\gamma^2}{8} \alpha(1 - \alpha) + O(\gamma^3). \quad (5.31)$$

Again where we've used Lemma 5.B.5.

$$\begin{aligned} \gamma^2 |\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - B_{P(\alpha_1)})^3]| &\leq \gamma^2 \mathbb{E}_{\pi_\gamma} [|\alpha_1 - \alpha| |\alpha - B_{P(\alpha_1)}|^3] \\ &\leq \gamma^2 D_{\gamma,\alpha} \mathbb{E}_{\pi_\gamma} [|\alpha - B_{P(\alpha_1)}|^3] \\ \gamma^2 |\mathbb{E}_{\pi_\gamma} [(\alpha_1 - \alpha)(\alpha - B_{P(\alpha_1)})^3]| &\leq \gamma^2 D_{\gamma,\alpha} \max(\alpha^3, (1 - \alpha)^3). \end{aligned} \quad (5.32)$$

$$\gamma^3 |\mathbb{E}_{\pi_\gamma} [(\alpha - B_{P(\alpha_1)})^4]| \leq \gamma^3 \max(\alpha^4, (1 - \alpha)^4). \quad (5.33)$$

Gathering equations (5.27) to (5.33) together with equation (5.26), we obtain finally:

$$\mathbb{E}_{\pi_\gamma} [(P(\alpha_1) - \alpha)^4] = O(\gamma^2). \quad (5.34)$$

□

5.B.4 Proof of Theorem 5.3.2

In this section, we prove Theorem 5.3.2. Recall the theorem:

Theorem 5.3.2. *Assume that: (i) $\alpha \in \mathbb{Q}$; (ii) the residuals follow an AR(1) process (i.e., $\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1}$ with $(\xi_t)_t$ i.i.d. random variables admitting a continuous density with respect to Lebesgue measure, of support \mathcal{S}) clipped at a large value R , and $[-R, R] \subset \mathcal{S}$; (iii) the quantile function Q of the stationary distribution of $(\varepsilon_t)_t$ is known. Then $(\alpha_t, \varepsilon_{t-1})$ is a homogeneous Markov Chain in \mathbb{R}^2 that admits a unique stationary distribution $\pi_{\gamma,\varphi}$. Moreover,*

$$\frac{1}{T} \sum_{t=1}^T L(\alpha_t) \xrightarrow[T \rightarrow +\infty]{a.s.} \mathbb{E}_{\pi_{\gamma,\varphi}} [L].$$

We consider $Z_t = (\alpha_t, \varepsilon_{t-1})$ defined in the state-space $\mathcal{Z} = \mathcal{A} \times [-R, R]$ by

$$\begin{cases} \alpha_{t+1} &= \alpha_t + \gamma (\alpha - \mathbb{1}\{|\varepsilon_t| > Q_{1-P(\alpha_t)}\}) , \\ \varepsilon_t &= -R \vee (\varphi \varepsilon_{t-1} + \xi_t) \wedge R \end{cases}$$

That is, $(\alpha_t)_{t \geq 0}$ is the recurrence defined by Equation (5.2), and $(\varepsilon_t)_{t \geq 0}$ is an AR(1) process with parameters φ clipped at some large value R . Finally, $(\xi_t)_t$ is a sequence of i.i.d. r.v. admitting a continuous density with respect to the Lebesgue measure, of support $\mathcal{S} \supset [-R, R]$.

This chain is defined for parameters α, R considered as fixed, and we focus on the influence of γ, φ . The main difference w.r.t. the previous section is that the state space is not countable anymore. More precisely, the state space is a product of a finite discrete set and an interval of \mathbb{R} .

The state-space \mathcal{Z} is $\mathcal{A} \times [-R, R]$, where \mathcal{A} is defined in the previous Section 5.B.1, equation (5.5). We equip \mathcal{Z} with the σ -algebra $\mathcal{F} = \mathcal{P}(\mathcal{A}) \times \mathcal{B}(\mathbb{R})$, where $\mathcal{P}(\mathcal{A})$ is the power-set of the finite set \mathcal{A} and $\mathcal{B}(\mathbb{R})$ is the borel set of \mathbb{R} .

Lemma 5.B.7. *The sequence $(Z_t)_{t \geq 0}$ is a Markov chain. Moreover, the chain is Harris-recurrent, and admits a stationary distribution $\pi_{\gamma, \varphi}$.*

Proof. We observe that

$$Z_t = \begin{pmatrix} \alpha_{t+1} \\ \varepsilon_t \end{pmatrix} = \begin{pmatrix} \alpha_t + \gamma (\alpha - \mathbb{1}\{|\varphi \varepsilon_{t-1} + \xi_t| > Q_{1-P(\alpha_t)}\}) \\ -R \vee (\varphi \varepsilon_{t-1} + \xi_t) \wedge R \end{pmatrix} =: F_{\gamma, \varphi}(Z_{t-1}, \xi_t). \quad (5.35)$$

For a function $F_{\gamma, \varphi} : \mathbb{R}^2 \times \mathbb{R}$. Consequently, Z_t follows a *Non-Linear State Space* model (Meyn and Tweedie, 2012, Section 2.2.2 and Chapter 7). We denote $P_{\gamma, \varphi}$ the probability kernel or Markov transition function, that is, for any $z = (a, e) \in \mathcal{Z}$, and $F \in \mathcal{F}$:

$$P_{\gamma, \varphi}(z, F) = \mathbb{P}(Z_1 \in F | Z_0 = z).$$

Remark that relying on Equation (5.35), we have an explicit formula for $P_{\gamma, \varphi}$. Defining the sequence of functions $(F_t)_{t \geq 1}$ such that

$$F_{t+1}(z_0, \xi_1, \dots, \xi_{t+1}) = F_{\gamma, \varphi}(F_t(z_0, \xi_1, \dots, \xi_t), \xi_{t+1})$$

where z_0 and (ξ_i) are arbitrary real numbers. By induction we have that for any initial condition $Z_0 = z_0 \in \mathcal{Z}$ and any $t \in \mathbb{N}$,

$$Z_t = F_t(z_0, \xi_1, \dots, \xi_t),$$

which immediately implies that the t -step transition function may be expressed as

$$P_{\gamma, \varphi}^t(z, F) = \mathbb{P}(F_t(z, \xi_1, \dots, \xi_t) \in F) = \int \dots \int \mathbb{1}\{F_t(z, \xi_1, \dots, \xi_t) \in F\} p(d\xi_1) \dots p(d\xi_t)$$

where p is the distribution of ξ .

We first prove that the chain is ψ -irreducible, for $\psi = \mu \otimes \lambda_{\text{Leb}}$, with μ the uniform probability measure on \mathcal{A} and λ_{Leb} the Lebesgue measure on $[-R; R]$.¹⁰

For any $z_0 = (a_0, e_0) \in \mathcal{Z}$ and $F = \{a'\} \times \mathcal{O}$, with \mathcal{O} open set, such that $\psi(F) \neq 0$ we have that, for some t large enough

$$\mathbb{P}(Z_t \in F | Z_0 = z_0) > 0.$$

Indeed,

1. There exists a path $(a_0, \dots, a_t = a')$ in \mathcal{A} from a_0 to a' such that for all $s \in \{1, \dots, t-1\}$, $0 < a_s < 1$; and $a_{s+1} - a_s \in \{\gamma(\alpha - 1), \gamma\alpha\}$ similarly to the proof of Lemma 5.B.2 since $\alpha \in \mathbb{Q}$.
2. Let E_{s+1} be the event such that we obtain a_{s+1} from a_s . Technically, if
 - a. if $a_{s+1} - a_s = \gamma(\alpha - 1)$, $E_{s+1} = \{\xi_s \text{ such that } |\varphi\varepsilon_{s-1} + \xi_s| > Q_{1-a_s}\}$
 - b. conversely, if $a_{s+1} - a_s = \gamma\alpha$, $E_{s+1} = \{\xi_s \text{ such that } |\varphi\varepsilon_{s-1} + \xi_s| \leq Q_{1-a_s}\}$.
3. Then if $0 < a' < 1$, we can directly conclude, as we have that for all $s \in \{1, \dots, t\}$,

$$\mathbb{P}(Z_{s+1} \in \{a_{s+1}\} \times E_{s+1} | Z_s = (a_s, z_s)) = \mathbb{P}(E_{s+1}) > \delta > 0.$$

Indeed (for case a.):

$$\begin{aligned} \mathbb{P}(E_{s+1}) &= \mathbb{P}\{\xi_s \text{ such that } |\varphi\varepsilon_{s-1} + \xi_s| > Q_{1-a_s}\} \\ &> \min \left(\mathbb{P}\{\xi_s \text{ such that } \xi_s > Q_{1-\min \mathcal{A} \cup \mathbb{R}_+^*}\}, \right. \\ &\quad \left. \mathbb{P}\{\xi_s \text{ such that } \xi_s < -Q_{1-\min \mathcal{A} \cup \mathbb{R}_+^*}\} \right) =: \delta, \end{aligned}$$

with $\delta > 0$ by the assumption (ii) (esp. the fact that the support \mathcal{S} of ξ_s includes $[-R, R]$).

The proof is similar for case b.

Consequently, $\mathbb{P}(Z_t \in F | Z_0 = z_0) > \delta^t > 0$.

4. The argument extends to the case where $a' < (0, 1)$, relying on the fact that $\psi(F) > 0$.

Moreover, the argument can be extended to show that for any a', \mathcal{O} , there exists δ' such that for all a_0, e_0 , there exists $t \leq C_{\alpha, \gamma}$ (e.g., $C_{\alpha, \gamma} = \frac{2}{\alpha\gamma}$) such that

$$\mathbb{P}(Z_t \in F | Z_0 = z_0) > \delta'.$$

Which proves that the chain will visit infinitely many times any Borel set F with probability 1, and is consequently Harris-recurrent (Meyn and Tweedie, 2012, Chapter 9). Using Theorem 10.0.1 in Meyn and Tweedie (2012), we conclude that the chain admits a unique stationary distribution $\pi_{\gamma, \varphi}$.

Finally, applying (Theorem 17.1.7 Meyn and Tweedie, 2012) to the later result gives:

$$\frac{1}{T} \sum_{t=1}^T L(\alpha_t) \xrightarrow[T \rightarrow +\infty]{a.s.} \mathbb{E}_{\pi_{\gamma, \varphi}}[L].$$

□

¹⁰Moreover ψ is transformed to remove mass from the sets that cannot be reached by the chain $(Z_t)_t$, i.e., if B is such that $\mathbb{P}(Z_t \in B) = 0$ for all t . This only concerns extremely marginal points, possible only $\alpha > 1$ or $\alpha = \min \mathcal{A}$, for which we assign a zero mass to $\alpha \times \mathcal{U}$ for some \mathcal{U} .

5.B.5 Numerical study of ACI efficiency with AR(1) residuals, with respect to the median length

We here reproduce the same experiment as in Section 5.3.2, but focus on the *efficiency* as the median of the intervals' lengths instead of the average (after imputation). Results are given in Figure 5.12.

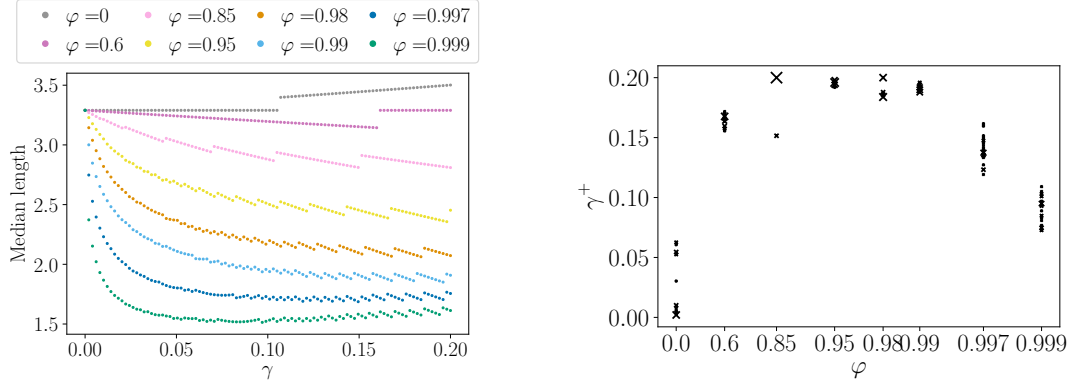


Figure 5.12: Left: evolution of the median length depending on γ for various φ . Right: γ^+ minimizing the median length for each φ .

Observations are very similar to the average length case, especially regarding (i) the monotonicity of the median interval length w.r.t. φ , (ii) the existence of a minimum γ_φ^+ to the function $\gamma \mapsto \text{Med}_{\mathcal{B}_n}[\tilde{\text{ff}}] := \arg\min_m \mathbb{E}_{\mathcal{B}_n}[|\tilde{\text{ff}} - m|]$ (iii) the non-monotonicity of $\varphi \mapsto \gamma_\varphi^+$.

5.C Experimental details.

5.C.1 Details on the BOA procedure

The Bernstein Online Aggregation (BOA) procedure (Wintenberger, 2017) is a type of aggregation rule Φ . The weights outputted by BOA have an exponential form. In the exponent is plugged the difference between the loss suffered by the last aggregated forecast and the current squared loss suffered by the expert, instead of plugging the losses suffered by the experts (this would be Exponential Weighted Aggregation, Vovk, 1990). As stated in Wintenberger (2017), “this procedure favors online learners that predicted accurately and which past predictions losses are close to the loss of the last aggregative online learner, ensuring the stability in time and a small quadratic variation”. For more details, we refer the reader to the original paper Wintenberger (2017).

5.C.2 Details ARMA(1,1) processes

Definition 5.C.1 (ARMA(1,1) process). We say that ε_t is an ARMA(1,1) process if for any t :

$$\varepsilon_{t+1} = \varphi \varepsilon_t + \xi_{t+1} + \theta \xi_t,$$

with:

- $\theta + \varphi \neq 0$, $|\varphi| < 1$ and $|\theta| < 1$;
- ξ_t is a white noise of variance σ^2 , called the *innovation*.

The asymptotic variance of this process is:

$$\text{Var}(\varepsilon_t) = \sigma^2 \frac{1 - 2\varphi\theta + \theta^2}{1 - \varphi^2}. \quad (5.36)$$

An ARMA(1,1) is thus characterised by three parameters: the coefficients φ and θ and the innovation's variance σ^2 . The larger the coefficients, in absolute value, the greater the time dependence and variance. Note that when $\varphi = 0$, the ARMA(0,1) process corresponds to a MA(1) and when $\theta = 0$, the ARMA(1,0) process corresponds to an AR(1).

To fix the asymptotic variance of an ARMA(1,1) of parameters φ and θ to v , we fix $\sigma^2 = v \frac{1 - \varphi^2}{1 - 2\varphi\theta + \theta^2}$.

5.C.3 Random forest parameters

All the random forests model have the same parameters, that are the following:

- Number of trees: 1000
- Minimum sample per leaf: 1 (default)
- Maximum number of features: d (default)

Furthermore, for EnbPI, as there is already an individual bootstrap in the algorithm, the random forest regressors do not bootstrap them again.

5.C.4 Details about the baselines and comparison

5.C.4.1 ENBPI FULL ALGORITHM

In order to be self-contained and precise the modifications done in EnbPI V2, the EnbPI algorithm from [Xu and Xie \(2021\)](#) is recalled in the following. In [purple](#) we precise the difference in EnbPI V2.

Remark on the bootstrap approach. The bootstrap scheme is not adapted to time series, even if such strategies have been developed ([Härdle et al., 2003](#); [Kreiss and Paparoditis, 2012](#); [Cai and Davies, 2012](#)), and could be used to improve the adequation of EnbPI with the time series framework. Furthermore, recent works have proposed modifications of RF in the dependent setting ([Goehry, 2020](#); [Goehry et al., 2021](#); [Saha et al., 2021](#)). Generalizing these improvements to any ensemble method and use it for EnbPI could also enhance its performance, but is out of the scope of this paper.

Algorithm 11 Sequential Distribution-free Ensemble Batch Prediction Intervals (EnbPI)

Input: Training data $\{(x_i, y_i)\}_{i=1}^T$, regression algorithm \mathcal{A} , decision threshold α , aggregation function φ , number of bootstrap models B , the batch size s , and test data $\{(x_t, y_t)\}_{t=T+1}^{T+T_1}$, with y_t revealed only after the batch of s prediction intervals with t in the batch are constructed.

Output: Ensemble prediction intervals $\{C_\alpha(x_t)\}_{t=T+1}^{T+T_1}$

```

1: for  $b = 1, \dots, B$  do
2:   Sample with replacement an index set  $S_b = (i_1, \dots, i_T)$  from indices  $(1, \dots, T)$ 
3:   Compute  $\hat{f}^b = \mathcal{A}(\{(x_i, y_i) \mid i \in S_b\})$ 
4: end for
5: Initialise  $\varepsilon = \{\}$ 
6: for  $i = 1, \dots, T$  do
7:    $\hat{f}_{-i}^\varphi(x_i) = \varphi\left(\left\{\hat{f}^b(x_i) \mid i \notin S_b\right\}\right)$ 
8:   Compute  $\hat{\varepsilon}_i^\varphi = |y_i - \hat{f}_{-i}^\varphi(x_i)|$ 
9:    $\varepsilon = \varepsilon \cup \{\hat{\varepsilon}_i^\varphi\}$ 
10: end for
11: for  $t = T + 1, \dots, T + T_1$  do
12:   Let  $\hat{f}_{-t}^\varphi(x_t) = (1 - \alpha)$  quantile of  $\left\{\hat{f}_{-i}^\varphi(x_t)\right\}_{i=1}^T$  EnbPI V2: this is replaced by
       
$$\hat{f}_{-t}^\varphi(x_t) = \varphi\left(\left\{\hat{f}_{-i}^\varphi(x_t)\right\}_{i=1}^T\right).$$

13:   Let  $w_t^\varphi = (1 - \alpha)$  quantile of  $\varepsilon$ 
14:   Return  $C_{T,t}^{\varphi,\alpha}(x_t) = \left[\hat{f}_{-t}^\varphi(x_t) \pm w_t^\varphi\right]$ 
15:   if  $t - T = 0 \pmod s$  then
16:     for  $j = t - 1, \dots, t - 1$  do
17:       Compute  $\hat{\varepsilon}_j^\varphi = |y_j - \hat{f}_{-j}^\varphi(x_t)|$ 
18:        $\varepsilon = (\varepsilon - \{\hat{\varepsilon}_1^\varphi\}) \cup \{\hat{\varepsilon}_j^\varphi\}$  and reset index of  $\varepsilon$ 
19:     end for
20:   end if
21: end for

```

5.C.4.2 DETAILS ON THE IMPLEMENTATION

We conclude this section by summarizing computational aspects of the methods. One of the contributions is to provide a unified experimental framework. Therefore, in Table 5.1, we display the current available code for these methods, and what is available in the proposed repository.

Table 5.1: Summary of available code online for each method and the proposed code in the repository. The programming language is specified, and, when relevant, the nature of the code.

Methods	Currently available		Contribution	
	Language	Details	Language	Options
CP	R		Python	
OSCP	not available		Python	randomised split
EnbPI	Python		Python	same aggregation function
ACI	R script	no general function	Python	randomised split

5.D Additional experiments on synthetic data sets

In this section, we provide supplemental results on the synthetic data sets presented in Section 5.5.1.

First, in Section 5.D.1 the sensitivity analysis of ACI γ as well as the comparison to the naive strategy and AgACI is extended to AR(1) and MA(1) processes of asymptotic variance 10.

Then, in Section 5.D.2, the comparison of all the CP methods for time series (initiated in Section 5.5.4) is also extended to these noises, that is AR(1) and MA(1) processes of asymptotic variance 10 (Section 5.D.2.1), and to ARMA(1,1), AR(1) and MA(1) processes of asymptotic variance 1 (Section 5.D.2.2).

Next, we discuss in Section 5.5.4 that the improved *validity* for $\gamma = 0.05$ in comparison to $\gamma = 0.01$ comes at the cost of more infinite intervals. This analysis is detailed in Section 5.D.3.

Finally, we compare randomized and sequential split in Section 5.D.4.

Imputation. The rationale to impute the infinite intervals is the following. We take the maximum of the absolute values of the residuals on the test set, noted $|\varepsilon|_{\max}$. Then, for any $t \in \llbracket T_0 + 1, T_0 + T_1 \rrbracket$, if the predicted upper (resp. lower) bound $\hat{b}_t^{(u)}(x_t)$ is such that $\hat{b}_t(x_t) > \hat{\mu}_t(x_t) + |\varepsilon|_{\max}$ (resp. $\hat{b}_t^{(\ell)}(x_t) < \hat{\mu}_t(x_t) - |\varepsilon|_{\max}$) we impute it by $\hat{\mu}_t(x_t) + |\varepsilon|_{\max}$ (resp. $\hat{\mu}_t(x_t) - |\varepsilon|_{\max}$).

5.D.1 Additional experimental results of ACI sensitivity to γ , presented in Section 5.5.2

In this subsection, we provide similar results to those of Section 5.5.2, for different models on the noise. Especially, we consider AR(1) and MA(1) processes.

Observations. The behaviour of the AR(1) process is very similar to the one of ARMA(1,1). On the other hand, for the MA case, the dependence structure is too weak to observe a significant effect of γ . All ACI methods produce nearly valid intervals, with coverage above 89.25%.

Results are given in Figures 5.13 and 5.14.

5.D.2 Comparison to baselines, extension of Section 5.5.4

5.D.2.1 ASYMPTOTIC VARIANCE FIXED TO 10.

Figure 5.15 displays the results on data generated according to Section 5.5.1, for an asymptotic variance of the noise of 10 (as in Figure 5.6), when this noise is an AR(1) or MA(1) process.

Observations. As in the previous section, the methods' performances are greatly impacted by the type and strength of dependence structure. Figure 5.15 shows that while ARMA(1,1) and AR(1) noises lead to similar patterns, it is not the case for an MA(1) noise. In the latter, θ has little influence: the five performances (one for each θ) are similar within each method. In addition, offline sequential SCP is very close to OSSCP. This is expected

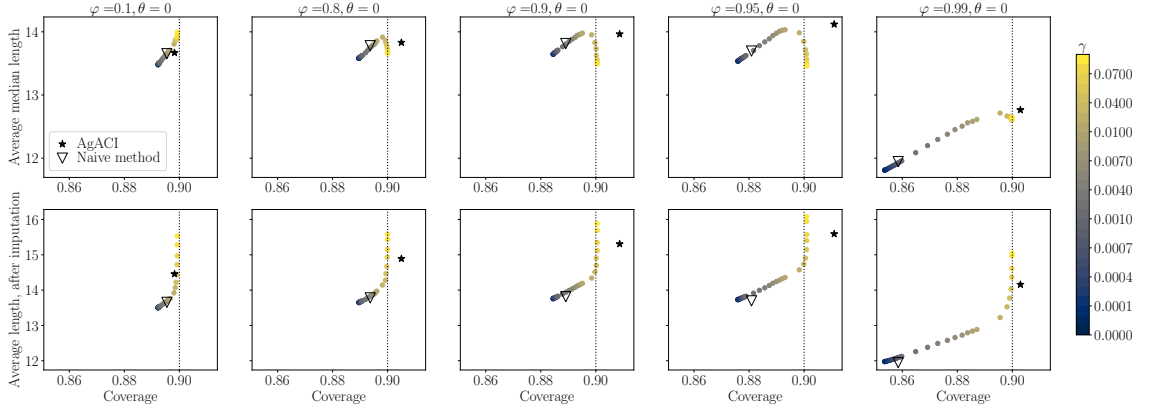


Figure 5.13: ACI performance with various θ , φ and γ on data simulated according to equation (5.3) with a Gaussian AR(1) noise of asymptotic variance 10 (see Section 5.C.2). Top row: average median length with respect to the coverage. Bottom row: percentage of infinite intervals. Stars correspond to the proposed online expert aggregation strategy, and empty triangles to the naive choice.

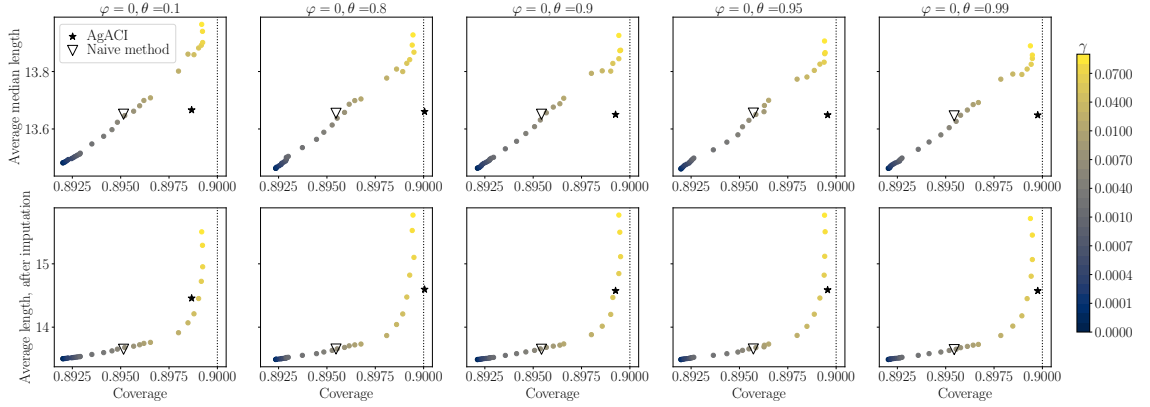


Figure 5.14: ACI performance with various θ , φ and γ on data simulated according to equation (5.3) with a Gaussian MA(1) noise of asymptotic variance 10 (see Section 5.C.2). Top row: average median length with respect to the coverage. Bottom row: percentage of infinite intervals. Stars correspond to the proposed online expert aggregation strategy, and empty triangles to the naive choice.

as a MA(1) process has very short memory, and the temporal dependence is thus small even for $\theta = 0.99$.

5.D.2.2 ASYMPTOTIC VARIANCE FIXED TO 1.

We now fix the asymptotic variance of the noise to 1. The results are plotted in Figure 5.16. Note that this is an easier setting than previously, as the signal to noise ratio is higher for this asymptotic variance.

Observations. Similarly to Figure 5.15, θ has little influence when the noise is a MA(1). On AR(1) and ARMA(1,1) noises (left and middle subplots), the patterns are similar. First, we observe again the improvement thanks to the online mode (empty squares versus solid ones), which increases when the dependence increases. Second, all the methods achieve *validity* or are significantly closer to achieving it than when the asymptotic variance is set to 10 (this is related to the high signal to noise ratio mentioned at the beginning of

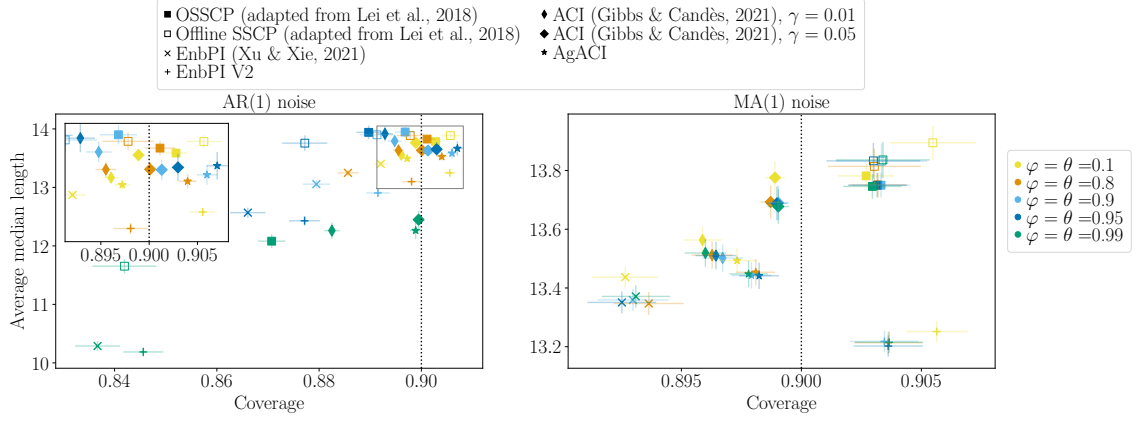


Figure 5.15: Performance of various interval prediction methods on data simulated according to equation (5.3) with a Gaussian AR(1) (left) and MA(1) (right) noise of asymptotic variance 10 (see Section 5.C.2). Results aggregated from 500 independent runs. Empirical standard errors are displayed.

this section). Third, EnbPI V2 is *valid* for $\varphi = \theta \leq 0.95$ and provides the most *efficient* intervals for these values. Nevertheless, its performances, as well as those of EnbPI, follow a clear trend (similar to that of Figure 5.6): when the dependence increases, the coverage decreases, as well as the length. EnbPI does not seem to be robust to the increasing temporal dependence in these experiments.

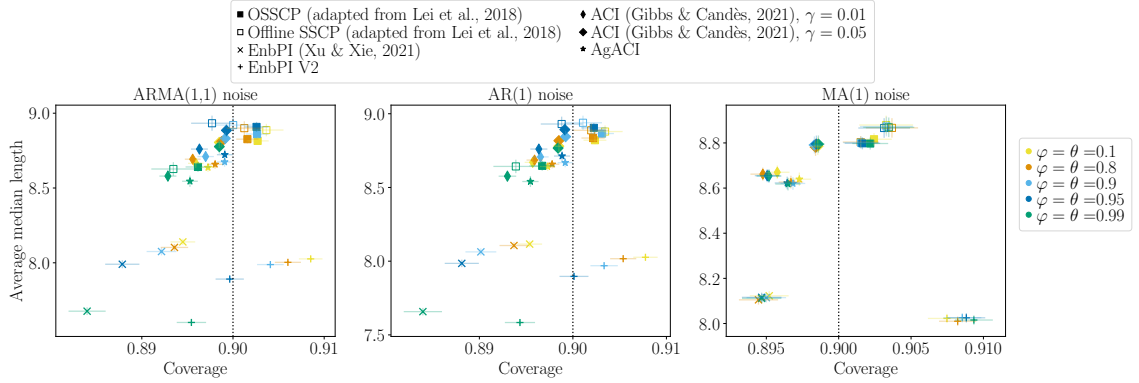


Figure 5.16: Performance of interval prediction methods on data simulated according to equation (5.3) with an ARMA(1,1) (left), AR(1) (center) and MA(1) (right) noise with a $\mathcal{N}(0, 1 - \frac{\varphi^2}{1 - 2\varphi\theta + \theta^2})$ innovation. Results aggregated from 500 independent runs. Empirical standard errors are displayed.

5.D.3 Closer look at infinite intervals

In this subsection, we investigate further the infinite intervals generated by ACI for ARMA(1,1), AR(1) and MA(1) noise models. We report the results in Table 5.2. The central two columns present the percentage of infinite intervals, for $\gamma = 0.01$ and $\gamma = 0.05$. A first obvious observation is that the number of infinite intervals is orders of magnitude smaller for $\gamma = 0.01$ than for $\gamma = 0.05$. The last column represents the proportion of points for which $\gamma = 0.05$ predicts \mathbb{R} and that are *not* covered for $\gamma = 0.01$. This suggests that for those intervals, predicting an infinite interval was somehow justified in the sense that the point

Table 5.2: Percentage of infinite intervals for ACI, on an ARMA(1,1) noise (first five rows), on an AR(1) noise ($\theta = 0$, next five rows) and a MA(1) noise ($\varphi = 0$, last five rows). The central two columns present the percentage of infinite intervals, for $\gamma = 0.01$ and $\gamma = 0.05$. The last column represents the proportion of points for which $\gamma = 0.05$ predicts \mathbb{R} and that are *not* covered for $\gamma = 0.01$.

Noise parameters	$\gamma = 0.01$	$\gamma = 0.05$	Intersection
$\varphi = \theta = 0.1$	0	1.12	53 out of 562 (9.43%)
$\varphi = \theta = 0.8$	0	2.76	263 out of 1381 (19.04%)
$\varphi = \theta = 0.9$	0	3.72	425 out of 1862 (22.83%)
$\varphi = \theta = 0.95$	0.03	4.45	514 out of 2224 (23.11%)
$\varphi = \theta = 0.99$	0.04	6.22	554 out of 3109 (17.82%)
$\varphi = 0.1$	0	1	37 out of 500 (7.40%)
$\varphi = 0.8$	0	2.75	212 out of 1373 (15.44%)
$\varphi = 0.9$	0	3.24	359 out of 1622 (22.13%)
$\varphi = 0.95$	0.03	4.32	488 out of 2160 (22.59%)
$\varphi = 0.99$	0.06	6.15	560 out of 3073 (18.22%)
$\theta = 0.1$	0	1.03	38 out of 516 (7.36%)
$\theta = 0.8$	0	1.42	49 out of 710 (6.90%)
$\theta = 0.9$	0	1.54	47 out of 772 (6.09%)
$\theta = 0.95$	0	1.54	45 out of 770 (5.84%)
$\theta = 0.99$	0	1.56	53 out of 781 (6.79%)

was seemingly challenging to cover (as $\gamma = 0.01$ failed to cover). For example, in the first line ($\varphi = \theta = 0.1$) we read that there are 562 points that result in infinite intervals for $\gamma = 0.05$, among which 53 lead to finite predictions for $\gamma = 0.01$ failing to cover on that point. This means only 9.43 % of 562 infinite intervals that can be considered as “somehow justified”. This analysis highlights that $\gamma = 0.05$ seem to predict more infinite intervals than necessary, to compensate for easy errors as explained in Section 5.2.

5.D.4 Randomised, sequential and other splits.

In Figure 5.17, we compare the sequential split strategy (dark markers) used in our experiments to the randomised version (clear markers), on online SCP. We observe that the intervals produced by the randomised version are significantly smaller than the sequential one, while covering slightly less.

Another splitting strategy would consist in calibrating on the first points and training on the last ones. Up to our knowledge, this has not been used in practice. This way, we could hope to obtain a better model for the point prediction task. Nevertheless, we would be calibrating on really different data than the test ones. Thereby, the impact of this scheme regarding the interval prediction task performance is not straightforward. This is why we focus here on the sequential split, which is the most intuitive approach. Analysing further all of these effects theoretically or with extensive numerical experiments would be beneficial to the time series conformal prediction domain.

5.E Forecasting French electricity spot prices

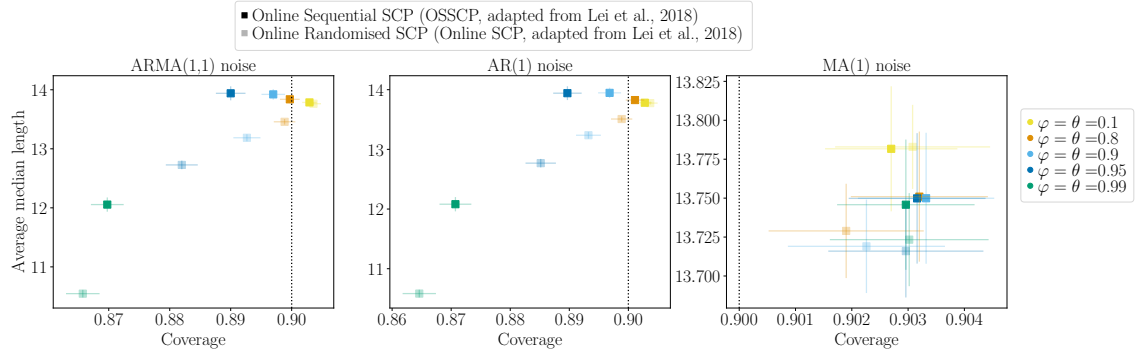


Figure 5.17: Performance of interval prediction methods on data simulated according to equation (5.3) with a Gaussian ARMA(1,1) (left), AR(1) (middle) and MA(1) (right) noise of asymptotic variance 10 (see Section 5.C.2). Randomised methods are displayed. Results aggregated from 500 independent runs. Empirical standard errors are displayed.

5.E.1 Details about the data set

Table 5.3 presents an extract of the French electricity spot prices data set used in Section 5.6. In this table, 2×23 columns are hidden for clarity and space: the 24 prices of $D - 7$ and the 24 prices of $D - 7$ are used as variables.

Table 5.3: Extract of the built data set, for French electricity spot price forecasting.

Date and time	Price	Price D-1	Price D-7	For. cons.	DOW
11/01/16 0PM	21.95	15.58	13.78	58800	Monday
11/01/16 1PM	20.04	19.05	13.44	57600	Monday
⋮	⋮	⋮	⋮	⋮	⋮
12/01/16 0PM	21.51	21.95	25.03	61600	Tuesday
12/01/16 1PM	19.81	20.04	24.42	59800	Tuesday
⋮	⋮	⋮	⋮	⋮	⋮
18/01/16 0PM	38.14	37.86	21.95	70400	Monday
18/01/16 1PM	35.66	34.60	20.04	69500	Monday
⋮	⋮	⋮	⋮	⋮	⋮

5.E.2 Forecasting year 2019

In Figure 5.18 we observe that on January 25, 2019, the forecasts are very different from the actual values. Nevertheless, the prediction intervals manage to include these observations for almost all hours (except after 5 pm) and almost all methods (EnbPI does not include points earlier, starting at 11 am).

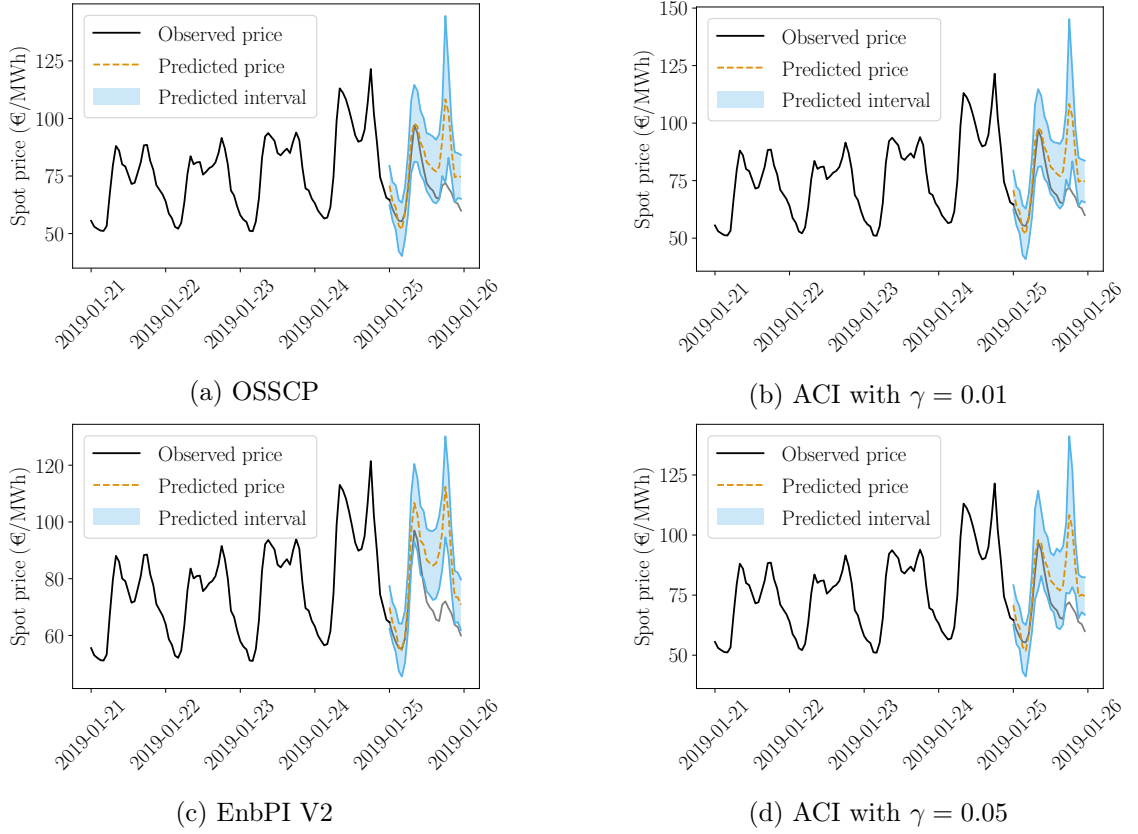


Figure 5.18: Representation of predicted intervals around point forecasts on the 25th of January of 2019.

In Figure 5.19 we observe that the four algorithms suffer from an unbalanced coverage depending on the day-of-the-week (each algorithm in a different extent). That is, they cover more than 90% of the observations on Tuesdays to Fridays, but less than 90% on Mondays and week-ends (Saturdays and Sundays).

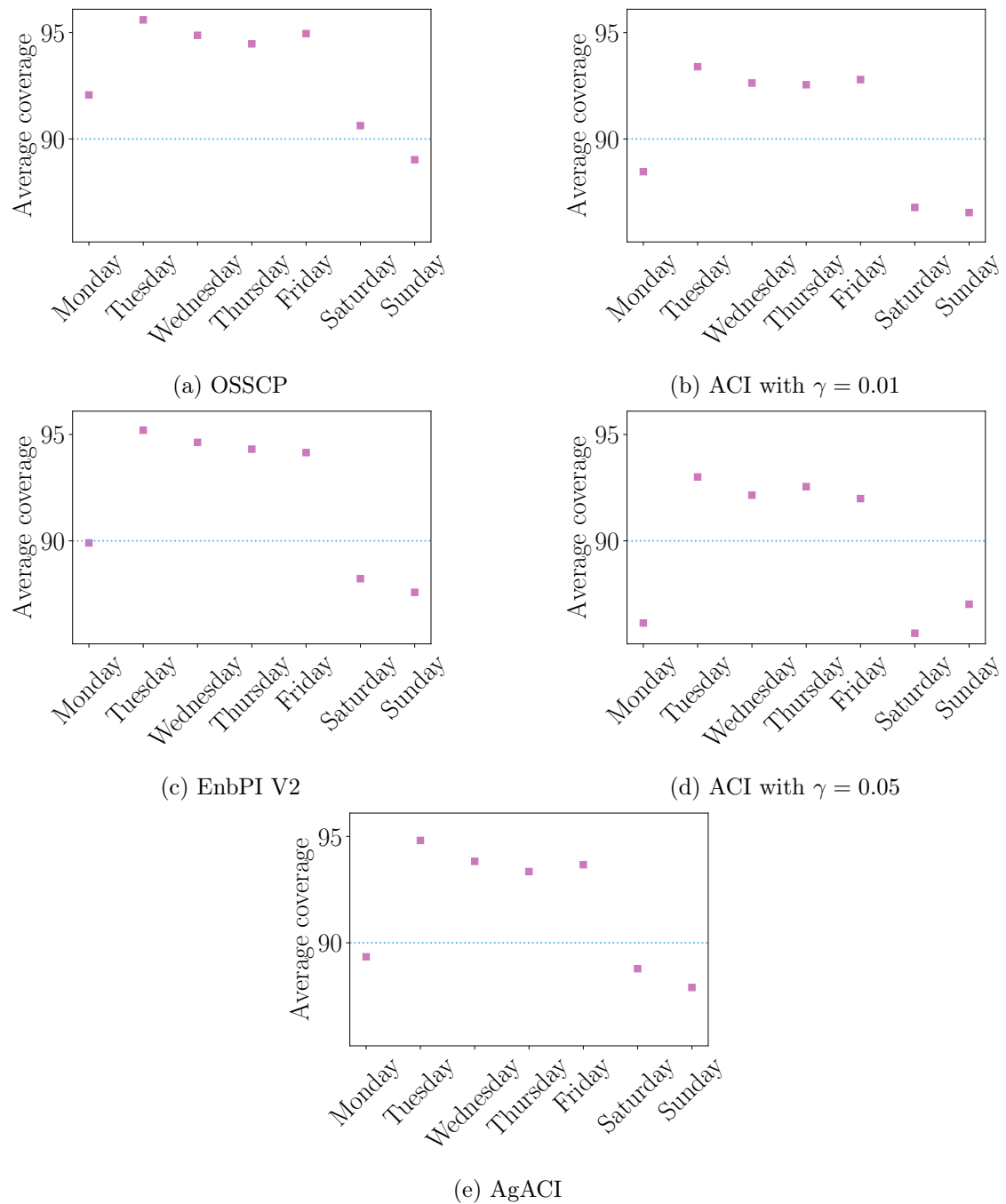


Figure 5.19: Coverage proportion during 2019 depending on the day-of-the-week.

Chapter 6

Adaptive Probabilistic Forecasting of French Electricity Spot Prices in 2020 and 2021

Electricity price forecasting (EPF) plays a major role for electricity companies as a fundamental entry for trading decisions or energy management operations. As electricity can not be stored, electricity prices are highly volatile which make EPF a particularly difficult task. This is all the more true when dramatic fortuitous events disrupt the markets. Trading and more generally energy management decisions require risk management tools which are based on probabilistic EPF (PEPF). In this challenging context, we argue in favor of the deployment of highly adaptive black-boxes strategies allowing to turn any forecasts into a robust adaptive predictive interval, such as conformal prediction and online aggregation, as a fundamental last layer of any operational pipeline.

We propose to investigate a novel data set containing the French electricity spot prices during the turbulent 2020-2021 years, and build a new explanatory feature revealing high predictive power, namely the nuclear availability. Benchmarking state-of-the-art PEPF on this data set highlights the difficulty of choosing a given model, as they all behave very differently in practice, and none of them is reliable. However, we propose an adequate conformalisation, `OSSCP-horizon`, that improves the performances of PEPF methods, even in the most hazardous period of late 2021. Finally, we emphasize that combining it with online aggregation significantly outperforms any other approaches, and should be the preferred pipeline, as it provides trustworthy probabilistic forecasts.

Contents

6.1	Introduction	106
6.2	Data presentation and insightful new explanatory variables	108
6.2.1	Dataset's description	108
6.2.2	First point forecast and feature importance	109
6.3	Probabilistic forecasting methods	110
6.3.1	Framework	111
6.3.2	Quantile regression methods	111
6.3.3	Conformal methods: add-on to traditional probabilistic approaches .	113
6.4	Adaptiveness as a wrapper around individual forecasts	114
6.4.1	Online aggregation based strategies	115
6.4.2	Adaptive conformal approaches	115
6.5	Application and results	119
6.5.1	Setting and evaluation	119
6.5.2	Results	120
6.6	Conclusion and perspectives	122
<hr/>		
6.A	Results on the CRPS	124

6.1 Introduction

Electricity price forecasting (EPF) plays a major role for electricity companies as a fundamental entry for trading decisions or energy management operations. As electricity can not be stored, electricity prices are highly volatile which make EPF a particularly difficult task (Weron, 2014; Lago et al., 2021).

The increase of renewable production in many countries (RTE, 2022; IEA, 2022a), the development of storage devices or more generally demand response programs (e.g., electrical vehicle smart charging (Nassar et al., 2022), electric water heater management (Amabile et al., 2021; Moreno et al., 2023)) simultaneously entails a need for good EPF and generates more complexity for price modelling. Furthermore, prices can be affected by fortuitous events such as Covid-19 pandemic in 2020-2021 (IEA, 2021), the stress corrosion issue which affected French nuclear power plants in 2022 or the crisis of the gas markets triggered by Russia's invasion of Ukraine (IEA, 2022b). Trading and more generally energy management decisions require risk management tools which are based on probabilistic EPF (Bunn et al., 2016). This supports the advancement of adaptive probabilistic approaches for forecasting prices, which can continuously learn and adjust to the evolving behaviors of EP, resulting in accurate and reliable probabilistic forecasts.

The literature on EPF is growing rapidly and most papers deals with point forecasts (Weron, 2014; Lago et al., 2021). We focus on short term (day-ahead) EPF as the mainstay of short-term power trading in Europe is the day-ahead market. As proposed in (Lago et al., 2021), models used for forecasting electricity prices can be categorized as either statistical, machine learning or hybrid models.

Statistical models are dominated by auto-regressive models and their variants, in particular the state of the art Lasso Estimated AutoRegressive (LEAR) model proposed by Uniejewski et al. (2016) and recently used as state of the art benchmark in (Lago et al., 2021; Tschora et al., 2022). It consists in a high dimensional ARX model where the fitting process is done by minimizing an elastic net regularization. The high dimension (around 250 parameters) comes from a large number of lags of prices and forecasts of variable of interests (generation, zonal prices, consumption). As highlighted by Lago et al. (2021) pre-processing of EP such as log transformations or more generally variance stabilizing transformations (Uniejewski et al., 2018) are a common practice to deal with heavy tailed distribution. Regarding non-stationarity of the prices, regime switching ARX models are proposed in (Nitka et al., 2021). Marcjasz et al. (2018) propose to average a set of point forecasts obtained from learning with different time windows to derive probabilistic forecasts.

The utilisation of **machine learning** tools including deep learning approaches for electricity price forecasting (EPF) has grown over the past decade. Recent studies (Tschora et al., 2022; Jędrzejewski et al., 2022) reveal that complex ML methods such as deep neural networks can achieve better forecasting performances than the LEAR model at the cost of significantly higher computational cost. The relatively important dimension of these models require a significant amount of data for their calibration, making them poor candidate to adapt to abrupt changes in price distribution (Çağatay Berke Bozlak and Yaşar, 2024). Yang et al. (2023) show how graphical neural network could be used to model spatial

dependency to forecast the day-ahead electricity prices of the Nord Pool market.

Probabilistic price forecasting is progressively becoming more popular in the forecasting literature following the GEFCom2014 energy forecasting competition (Hong et al., 2016). This is a natural goal as the final objective EPF is to optimize a financial risk criteria (Bjorgan et al., 1999; Deschatre et al., 2021). Most of the previous parametric statistical models are based on statistical assumptions and could be naturally extended to produce probabilistic forecast (more or less accurate as we will explore in this paper). Relaxing distributional assumption, non parametric regression models such as quantile regression have been investigated (Uniejewski and Weron, 2021). In Loizidis et al. (2024), machine learning models coupled with bootstrap methods are compared with classical time series models for German and Finnish day-ahead market. Marcjasz et al. (2023) recently proposed a distributional network that outperforms state-of-the-art benchmarks. Nickelsen and Müller (2024) present a Bayesian forecasting framework for the German continuous intraday market and show that orthogonal matching pursuit methods can outperform LEAR. Cornell et al. (2024) propose quantile regression with varying training-length periods and model averaging to forecast prices of the South Australia region of the Australian National Electricity Market.

PEPF models face many pitfalls: extreme price spikes, non-stationarity due to exogenous factors inducing time-varying mean and/or volatility. Conformal methods (Vovk et al., 1999; Papadopoulos et al., 2002; Vovk et al., 2005) and more specifically adaptive conformal methods, proposed for example by Gibbs and Candès (2021); Zaffran et al. (2022), are a way to adapt PEPF models in a very general way. It can be applied to any of the previously cited PEPFs to improve them. We propose to extend the work of Zaffran et al. (2022) to forecast electricity prices in France during the turbulent period 2020-2022. Another framework allowing to adapt PEPF models is online aggregation under expert advice (Cesa-Bianchi and Lugosi, 2006), which was successfully used in financial non-stationary environments (Remlinger et al., 2023; Berrisch and Ziel, 2024a). Our aim is to investigate if and how it is possible to make adaptive an existing probabilistic forecasting algorithm. This approach is driven by an operational concern: proposing a plug-in tool that can be applied to any underlying model eases its integration in the current pipeline.

Contributions We list below our main contributions:

- **New data:** we study the recent turbulent period 2020-2022 and we add a new feature, the nuclear availability
- **Benchmark:** we consider state-of-the-art PEPF methods, their windowed versions (rolling window estimation) and benchmark them on this new dataset
- **Analysis** of the improvements (or not) of existing **online conformal methods**
- Suggestion of **novel online conformal strategy** coined OSSCP-horizon
- Unified framework of **sequential aggregation** of all these probabilistic forecasting

- **Understanding the benefits** of these 2 frameworks of probabilistic post-processing (i.e. CP and aggregation) and how they can help each other: *sequential aggregation with conformalized expert is the best*

6.2 Data presentation and insightful new explanatory variables

6.2.1 Dataset's description

The considered dataset spans approximately 6 years of observations at a hourly frequency, from January 11th, 2016 to December 31st, 2021, and is decomposed of a training set (from January 11th, 2016 to December 31st, 2018) to estimate the parameters of the models, a validation test (year 2019) to estimate the hyperparameters, and a test set (years 2020 and 2021) to evaluate the performances (see Figure 6.1). We consider the task of forecasting day-ahead (DAH) prices on the French EPEX market. As the 24 hours of day d are fixed from EUPHEMIA¹'s market clearing at 12:00pm of day $d - 1$, the features considered to predict each of them are selected so that they are available before 12:00pm of day $d - 1$. More precisely the dataset contains the following features, for a target at day d , hour h :

- the 24 French DAH prices at days $d - 1$ and $d - 7$;

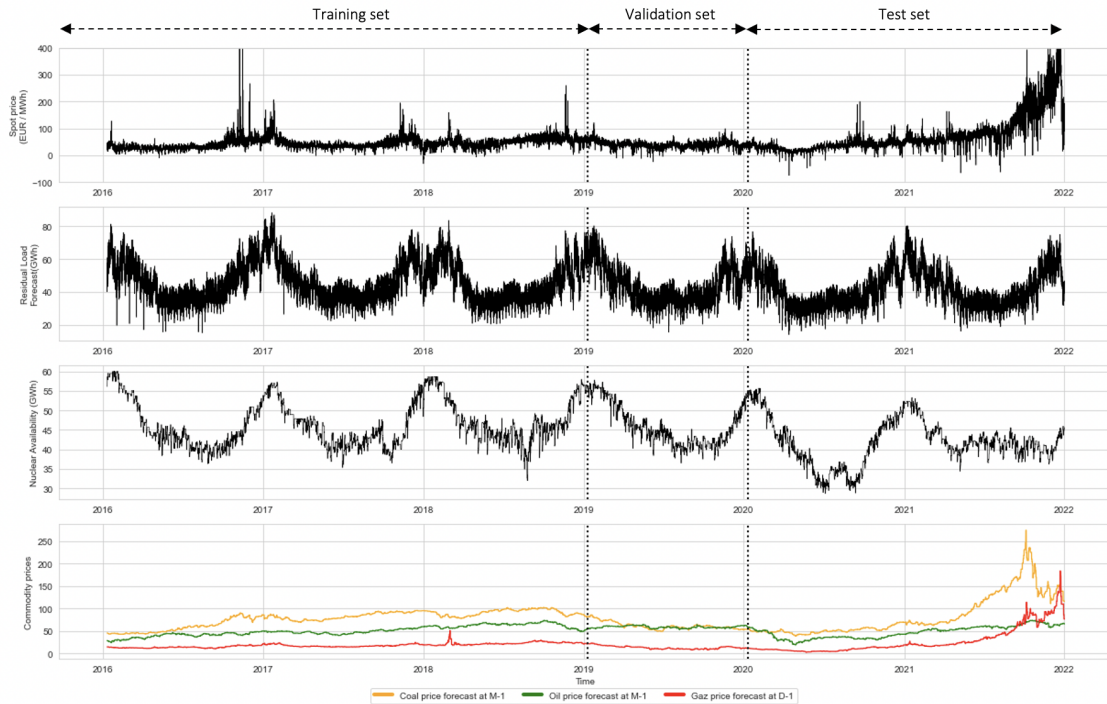


Figure 6.1: Evolution of the Spot prices (first panel), Residual Load (second panel), Nuclear availability (third panel) and commodity prices (last panel) from 2016 to 2021 (x -axis).

¹EUPHEMIA is the algorithm that solves the market coupling problem for the Central West European region, used by EPEX to compute the day-head power prices

- the observed daily price of Gas on the French PEG market at $d - 1$ and the month-ahead futures prices for Oil (Brent) and Coal (CIF ARA Argus-McCloskey);
- the forecasted residual load signal built with data available before 12pm at $d - 1$: the load forecasts for the 24 hours of day d , estimated on day $d - 2$, minus the renewable production forecasts (i.e., wind and solar forecasts estimated on day $d - 2$, and the observed run-of-river electricity on $d - 2$);
- the availability of French nuclear electricity on day d , i.e. the announced available capacity of nuclear generation;
- the observed electricity generation from all production types at $d - 2$ and $d - 7$ (in the case of nuclear energy, the production is divided by the nuclear availability);
- the EUR vs. GBP and EUR vs. USD exchange rate (last observed at $d - 1$);
- the total electricity volume exchanges between France and all its neighbors (observed at $d - 2$);
- the specific electricity volume exchanges between France and Germany (observed at $d - 2$);
- dummy variables, including dummy variables for French holidays (as a percentage of the total population concerned), holiday bridges, weekends, and weekdays;
- the time of year as a sine and cosine function, as well as a clock variable to capture a possible trend.

6.2.2 First point forecast and feature importance

The proposed dataset comprises features classically used to forecast electricity prices, and also a new feature, the nuclear availability, for we intuit that nuclear availability has a significant impact on DAH prices due to the French energy mix.

At first we proceed a point forecast exercise, with Lasso CV and Random forest models, to detect the most important features and highlight the relevance of the proposed new variables. Here, the meaning of the term “feature importance” varies according to the model: in the case of Lasso CV, it refers to the value of the coefficient associated to a given feature, whereas for Random Forest it refers to the Mean Decrease in Impurity (MDI).

In Figure 6.2, we observe the top 20 mean feature importances over both models trained in 2020. Spot price at H-23 of the previous day is the “most important” feature for the Lasso CV model. This is coherent with what is found in (Maciejowska et al., 2022; Ziel and Weron, 2018). The MDI-based importances computed for the Random Forest suggests the same conclusion, even though high correlation between all $d-1$ spot prices makes the interpretation harder. The Lasso CV model, which allows for a better modelisation with highly correlated features, suggests that gas prices and nuclear availability have a high explanatory power. This speaks in favour of an inclusion of these features in EPF prediction models, at least in the case of the French market.

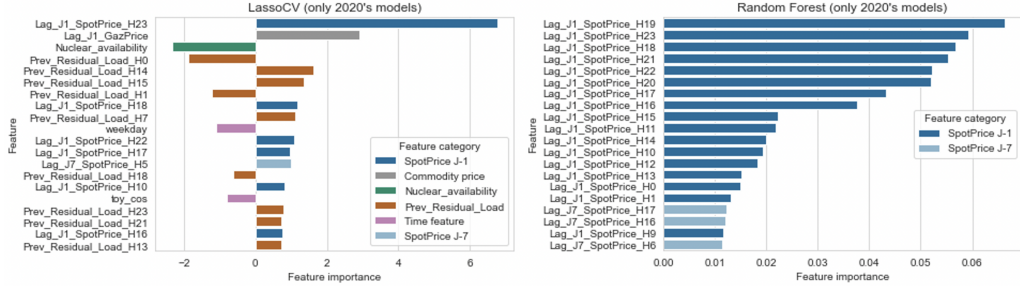


Figure 6.2: Feature (y -axis) importance (x -axis) for Lasso CV (left panel) and Random Forest (right panel) models. The colors are associated with a type of feature.

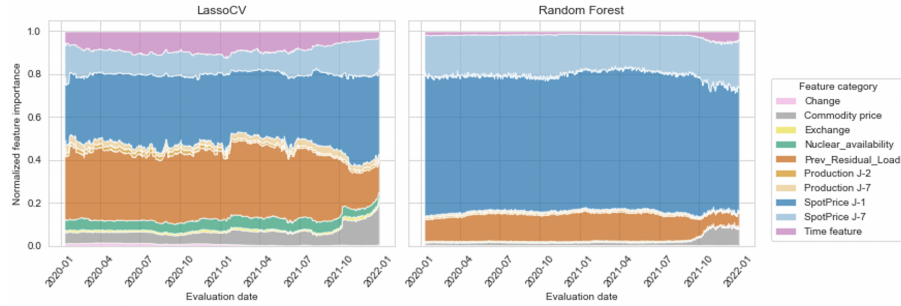


Figure 6.3: Evolution of normalized feature importance (y -axis) for Lasso CV (left panel) and Random Forest (right panel) models over the whole test period (x -axis). The colors are associated with the features.

We also compute the feature importance of both model over every days in the test period and observe the evolution in the predominance of the various feature groups. To do so, we first aggregate features into groups: “Change” for all exchange rates, “Commodity price” for gas, coal and oil prices, “Exchange” for all hourly power volumes exchanges, and the rest of features groups are hourly features aggregated at a daily level. The group aggregation consists in summing up the absolute importance value of all features belonging to this group, then normalize these values by the total sum over all groups. Figure 6.3 represents the evolution we obtain. We observe a considerable change in the relative group’s explanatory power: for both the Random Forest and Lasso model, we observe a significant increase in the aggregated explanatory power of the commodity prices, at the expense of the residual load forecast. This indicates an important distribution shift in the relationships between the times series by September 2021.

6.3 Probabilistic forecasting methods

Notations Given the nature of the data and in particular the hourly patterns, we will build one model per hour, as explained in Section 6.5.1. From now on, the temporal index t is used and it elapses at a daily rate (i.e., for a given hour h). $t = 1$ corresponds to the beginning of the training data, $t = T_0$ marks the end of the training data and $t = T_1$ refers to the last test observation to be predicted. In other words, we aim at predicting the French spot prices between $T_0 + 1$ and T_1 , corresponding to the years 2020 and 2021 (see Figure 6.1).

6.3.1 Framework

One objective of probabilistic forecast is to build *Prediction Intervals* (PIs) for a variable Y_t depending on the covariates X_t . Let $\alpha \in [0, 1]$ be a *miscoverage rate*. A PI at the $1 - \alpha$ level is expected to contain at least $1 - \alpha$ of the realisations: $\mathbb{P}(Y_t \in \text{PI}_{1-\alpha}(X_t)) \geq 1 - \alpha$, while being as small as possible. In order to retrieve as much information as possible about the distribution of Y_t , one can consider multiple values of the miscoverage rate α .

A PI can be characterized by two “point forecasts”: its lower ($\ell(X)$) and upper ($u(X)$) bounds. A natural choice for the PI is $\ell(X) = Q_{\alpha/2}(X)$ and $u(X) = Q_{1-\alpha/2}(X)$, where Q_β is the β -th quantile of the cumulative function distribution (c.d.f.) of the price conditionally to the covariates used to forecast.

However, in practice, these true Q are never known and we have to estimate them, e.g., using quantile regression (Koenker, 2005). This approach is detailed in Section 6.3.2.

Another path is to post-process individual predictors (see Section 6.3.3). The individual predictors can either estimate the mean as in point forecasting and the post-processing step will turn them into PI, or directly estimate a conditional quantile (as described in Section 6.3.2).

6.3.2 Quantile regression methods

We present here the quantile regression methods that we retained for our benchmark study. These methods were chosen for their good performance on time series data, and in particular on electricity related data. They are all quite easy to fit automatically and have a relatively low computational cost (this is a key asset due to the intensive benchmark including rolling window estimation).

6.3.2.1 DESCRIPTION OF THE METHODS

Basics on Quantile Regression (QR) QR (Koenker, 2005) replaces the usual quadratic loss by the *pinball loss* to forecast a conditional quantile of the distribution of Y (i.e. the price) given the features X :

$$\min_{g \in \mathcal{G}} \mathbb{E} [\rho_\beta(Y - g(X)) | X = x],$$

for any x , with ρ_β the *pinball loss* of level β : $\rho_\beta(y - \hat{y}) = (1 - \beta)|y - \hat{y}| \mathbf{1}\{y \leq \hat{y}\} + \beta|y - \hat{y}| \mathbf{1}\{y \geq \hat{y}\}$, and \mathcal{G} the class of regressors considered, e.g. linear models, Lasso (QLR-Lasso), additive non-linear models (QGAM) or gradient boosting regressors (QGB).

Quantile Linear Regression (Linear QR) and Quantile Lasso (Lasso QR) The class of regressors \mathcal{G} is restricted to linear models. For Lasso QR, We perform a Lasso selection process (Tibshirani, 1996) to deal with the pretty high number of covariates, the class of regressors is thus the linear models on all possible subsets of covariates.

Quantile Generalized Additive Models (QGAM) Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1986) consists in explaining the conditional expectation

$\mu(X)$ of Y over X with a semi-parametric additive structure. The estimation of GAMs is based on a (regularized) mean squared error (MSE) criterion. Our objective is to use GAMs for a QR problem. One could replace the MSE by the pinball loss function in the estimation process as described in the previous paragraph. However, [Fasiolo et al. \(2020\)](#) demonstrate that the pinball loss is statistically sub-optimal in this framework and proposes a procedures based on the smooth Extended Log-F loss instead.

Quantile Random Forests (QRF) [Meinshausen \(2006\)](#) adapts Random Forests to the QR task. The same forest is built than for mean-regression, that is a forest grown in order to minimize the mean squared error. However, to adapt to the quantile task at hand, the final decision rule for prediction now corresponds to evaluating an empirical conditional quantile (conditional on the fact that the features of the test point belongs to the corresponding leaves).

Quantile (tree based) Gradient Boosting (QGB) Gradient boosting machine ([Friedman, 2001](#)) are widely used in the forecasting community where it has demonstrated excellent performance for different applications on tabular data ([Grinsztajn et al., 2022](#)) or time series ([Makridakis et al., 2022](#)). As for the Random Forests, the regressors are here regression trees. The boosting algorithm consists in adding a sequence of simple models (called weak learners and trained on a subsample randomly selected of the training set) obtained by sequentially fitting a quantile regression tree to the residuals by minimizing the pinball loss, which is a key difference with QRF.

6.3.2.2 OPERATIONAL PIPELINE

We explore these prediction methods through their implementation in the Python package `scikit-learn` package ([Pedregosa et al., 2011](#)) for linear quantile regression, Lasso and QGB. QRF are implemented through `scikit-garden`. The QGAM are implemented in the R package ([Fasiolo et al., 2021](#)).

All of these models depend on hyper-parameters, and QGAM additionally requires an exact formula. In particular, we optimized for the regularizer (Lasso), the number of trees and their maximum depth (QRF and QGB), as well as the learning rate and fraction of samples (QGB), and the formula (QGAM). Their estimation is based on grid-searching on the validation set after estimation of mean-regression models on the training set, as illustrated in Figure 6.1. Therefore, the formula of the QGAM is the same for all quantiles. It includes:

- *linear effects*: for the indicator of the week days;
- *univariate non-linear terms*: the announced French nuclear availability, the lagged 2 days of the fossil hard coal and observed nuclear productions, the square root of the lagged one day of the Gaz prices, cosin and sin of the time of year;
- *functional smooth effects*: as proposed in [Amara-Ouali et al. \(2023\)](#) in the context of electricity load forecasting, we model the lagged (one day and one week) prices and

the load forecast effects via a functional smooth effect. It allows to capture the effect of these functional (in function of time) covariates over the price at a given instant of the day.

In this paper we do not consider online re-estimation of the hyperparameters, which in practice is very time consuming and statistically challenging. We study the performance of operational fixed prediction models that can be made adaptive through a plugged-in layer, useful when facing non-stationarity without completely retraining them.

Also, as illustrated in the preliminary results of Figure 6.4, before September 2021, only QRF and QGAM achieved *validity*. We explore strategies to recover validity in Section 6.3.3. What is more, none of the probabilistic methods attain the target coverage level after September 2021. Indeed, the high explosion of the prices after this date, both in average and in variability, calls for more adaptive strategies, that we discuss in Section 6.4. Note that the standard rolling training procedure did adapt to this change as illustrated by the lengths of the PIs after September 2021, but more adaptiveness is required given the strength of the shift and variability.

6.3.3 Conformal methods: add-on to traditional probabilistic approaches

Conformal Prediction (CP) (Vovk et al., 1999; Papadopoulos et al., 2002; Vovk et al., 2005) builds PI around any kind of prediction models. These intervals are valid (achieving marginal nominal coverage) in finite samples under the only assumption of exchangeability of the data. Therefore, CP has to be seen as an add-on protective layer to existing probabilistic

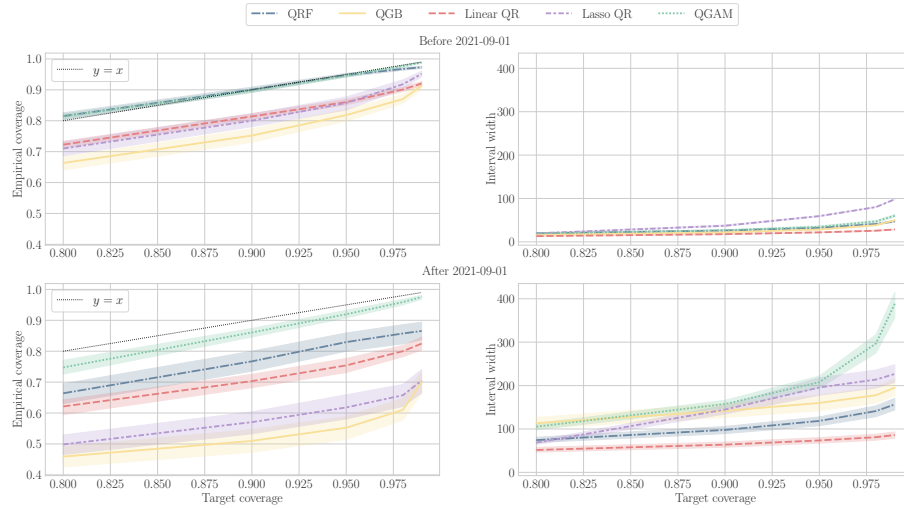


Figure 6.4: PIs's performance of individual probabilistic forecasts at test time, before September 2021 (top row) and after September 2021 (bottom row), for various target coverage levels (x -axis). The left column represents the average empirical coverage: the closest to the $y = x$ line the better, and above it is best. The right column represents the average interval width: the lower the better. The colors and shapes are associated with the models. The shaded regions correspond to the 5% and 95% empirical quantiles after bootstrapping 500 times the test time series, see Section 6.5.1 for details.

(or not) forecasts, that is able to robustify them in terms of validity but whose efficiency and shape will always rely on the quality of the underlying forecast.

Suppose that we have T_0 random variables $\{(X_t, Y_t)\}_{t=1}^{T_0}$. For a given miscoverage rate $\alpha \in [0, 1]$, we aim at building a *marginally valid* PI \hat{C}_α of Y_{T_0+1} , i.e. \hat{C}_α should satisfy:

$$\mathbb{P}\left(Y_{T_0+1} \in \hat{C}_\alpha(X_{T_0+1})\right) \geq 1 - \alpha. \quad (6.1)$$

To achieve this, Split Conformal Prediction (SCP) (Papadopoulos et al., 2002; Lei et al., 2018) randomly splits the T_0 data points into a training set Tr and a calibration set Cal . A regression model $\hat{\mu}$ is then fitted on Tr and used to predict on Cal to obtain a set of conformity scores $\mathcal{S}_{\text{Cal}} = \{S_t := s(X_t, Y_t; \hat{\mu}), t \in \text{Cal}\}$. These scores assess the conformity between the calibration's observed values and the predicted ones: the smaller the better. In the case of regression, they are usually computed using the absolute value of the residuals, i.e. $S_t := s(X_t, Y_t; \hat{\mu}) = |\hat{\mu}(X_t) - Y_t|$. A corrected² $(1 - \tilde{\alpha})$ -th empirical quantile of the conformity scores $Q_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}})$ is obtained, to finally build the prediction interval $\hat{C}_\alpha := \{y : s(X_{T_0+1}, y; \hat{\mu}) \leq Q_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}})\}$. In the standard regression case, it boils down to $\hat{C}_\alpha(X_{T_0+1}) = [\hat{\mu}(X_{T_0+1}) \pm Q_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}})]$. This procedure is guaranteed theoretically to satisfy Equation (6.1) for any model $\hat{\mu}$, any sample size T_0 , as long as the calibration and test data are exchangeable.

Proposed by Romano et al. (2019), Conformalized Quantile Regression (CQR) benefits simultaneously from the adaptiveness of classical QR methods and from the theoretical guarantees ensured by CP. Instead of training a mean regression model on the training set Tr , CQR requires to fit two conditional quantile regression models $\hat{q}_\ell(\cdot), \hat{q}_u(\cdot)$ ³. In this context, the conformity scores now quantify the error made by the fitted PI $\hat{C}(x) := [\hat{q}_\ell(x), \hat{q}_u(x)]$. Precisely, $S_t := s(X_t, Y_t; \hat{q}_\ell, \hat{q}_u) = \max\{\hat{q}_\ell(X_t) - Y_t; Y_t - \hat{q}_u(X_t)\}$. Accordingly, the PI becomes $\hat{C}_\alpha(X_{T_0+1}) = [\hat{q}_\ell(X_{T_0+1}) - Q_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}}), \hat{q}_u(X_{T_0+1}) + Q_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}})]$.

To account for the temporal aspect of time series, an online and sequential version of SCP is usually considered, in which the split leading to Tr and Cal is not random, but constrained so that any point in Tr occurs before any point in Cal (Wisniewski et al., 2020; Zaffran et al., 2022). See Figure 6.5 for an illustration.

6.4 Adaptiveness as a wrapper around individual forecasts

The online setting—in which the environment reveals the true value before the next prediction—allows to post-process individual predictors to adapt to previous errors (e.g., as done in CP). This approach demonstrates all its interest when stationarity – and consequently neither exchangeability – does not hold, as in our case study. One way to implement such a post-processing, coming from the online literature, is online aggregation

²The correction $1 - \tilde{\alpha} = (1 - \alpha)(1 + \frac{1}{\#\text{Cal}})$ is needed to ensure finite sample validity, because of the inflation of the quantiles.

³Usually $\ell = \alpha/2$ and $u = 1 - \alpha/2$, but this is not necessary. Romano et al. (2019) suggest to choose these values by cross-validation, to improve PI's efficiency.

of predictors, as described in Section 6.4.1⁴. Another strategy, within the CP framework, is to modify the calibration step of CP (see Section 6.4.2) and make it adaptive.

6.4.1 Online aggregation based strategies

Adaptive aggregation of *experts* (Cesa-Bianchi and Lugosi, 2006), with $K \in \mathbb{N}^*$ experts denoted $\left(\hat{f}_t^{(k)}(\cdot)\right)_{k \in \llbracket 1, K \rrbracket}$ being various individual forecasters for the prices at time t (that is a corresponding day d on a given hour h) such as the ones introduced in Section 6.3.2, computes an optimal weighted mean of the experts. At each time t (i.e., day d , for a given hour h), the weights $\omega_t^{(k)}$ assigned to expert k depend on all experts' suffered *losses*, i.e. their performances on the previous time steps until $t - 1$. In our case, these performances are evaluated through the pinball loss ρ_β , standard in quantile regression, with the pinball parameter β being the target quantile level. These losses are plugged in the *aggregation rule* Φ , outputting the aggregation weights. Finally, the aggregation rule can include the computation of the gradients of the loss (*gradient trick*, see (Cesa-Bianchi and Lugosi, 2006) for more details). As aggregation rules require bounded experts, a thresholding step is added. Concretely, the aggregated predictor at time t , $\hat{f}_t^\Phi(\cdot)$, is defined by

$$\hat{f}_t^\Phi(X_t) = \sum_{k=1}^K \omega_t^{(k)} f_t^{(k)}(X_t).$$

In our experiments, the different forecasts obtained are aggregated quantile by quantile, using the appropriate pinball loss as a score. The aggregation rule Φ is set to be the Bernstein Online Aggregation (BOA) (Wintenberger, 2017) algorithm, along with the gradient trick. We use the R package OPERA (Gaillard and Goude, 2021) to perform such an aggregation, and reorder the quantiles predicted by the aggregation models to avoid quantile crossing.

Recently, Berrisch and Ziel (2023) proposed an approach that jointly aggregates every quantile forecasting model together and gives directly a probabilistic prediction as an output, instead of performing independent aggregation for each quantile level. Berrisch and Ziel (2023)'s method reduces the number of aggregation parameters to be computed, while yielding preferable probabilistic performances. It is available in the R-Package `profoc` (Berrisch and Ziel, 2024b), compatible with the BOA method with the gradient trick and automatically reordering the predicted quantiles. It has to be noted that we did not explore the full range of tuning possibilities allowed by this method. In our experiments, both approaches performed similarly. Therefore, to avoid overloading the analysis, we present in this paper only the first method.

6.4.2 Adaptive conformal approaches

In addition to online aggregation, we consider another post-processing of individual forecasters which consists in adding a conformal layer on top of them, adaptively. As explained

⁴This does not include Quantile Regression Averaging (QRA) (Nowotarski and Weron, 2014) as it is an offline averaging, thus non-adaptive.

in Section 6.3.3, CP requires exchangeable data, an assumption clearly not satisfied in a time series setting, and even less in our highly non-stationary case study.

The first theoretically grounded result on CP for dependent data is given by [Chernozhukov et al. \(2018\)](#): it shows that when the data is strongly mixing and the learned model is close “enough” to the underlying data generation process then CP guarantees still hold, along with proposing an extension for full CP⁵ under which the previous theorem holds. Again, this is not sufficient to encapsulate our setting.

In practice, Online Sequential Split Conformal Prediction (OSSCP) is often used to take into account the temporal structure, introduced in [Wisniewski et al. \(2020\)](#); [Zaffran et al. \(2022\)](#). The idea is (i) to enforce a sequential split where all the training observations are temporally consecutive, and preceding the ones of the calibration set and (ii) to update this split in order to incorporate the newly observed data points at each prediction step $t + 1$, forgiving the oldest ones, leading to adaptive sets Tr_t and Cal_t . See Figure 6.5 (a) for an illustration. Note that OSSCP does not enjoy any form of theoretical guarantees beyond the exchangeable setting, despite its good empirical performances in the time series framework, as highlighted in ([Zaffran et al., 2022](#)).

6.4.2.1 IMPROVING CP ONLINE ADAPTIVENESS: OSSCP-HORIZON

One drawback of OSSCP is that the set on which the models were fitted can be far from the points on which it will be applied (either calibration or test points). If the temporal data suffers from a strong distribution shift, this may hinder the accuracy of the base learner, and therefore the performances of the PI, both in terms of coverage (the exchangeability assumption is not satisfied anymore) and in terms of efficiency, i.e. interval’s length (as large errors cause large intervals).

In order to avoid high errors on the calibration and test points, we propose a new approach, coined **OSSCP-horizon**. The idea is to ensure that the underlying model is trained on the data just preceding each calibration point: in other words, to only compute test errors of horizon one, as is the forecast horizon. More generally, for any forecasting task at horizon h , **OSSCP-horizon** computes calibration errors of horizon h . See Figure 6.5 (b) for an illustration. Formally, at prediction time $T + 1$, **OSSCP-horizon** thus builds the calibration set as follows:

- For each $X_t \in \text{Cal}_T$, fit quantile regression estimators $\hat{q}_\ell^{-(t)}, \hat{q}_u^{-(t)}$ on⁶

$$\{(X_{t-|\text{Tr}|}, Y_{t-|\text{Tr}|}), \dots, (X_{t-1}, Y_{t-1})\};$$

- Compute the calibration score $S_t = s(X_t, Y_t; \hat{q}_\ell^{-(t)}, \hat{q}_u^{-(t)})$ and add it to the set of scores $\mathcal{S}_{\text{Cal}_T}$.

⁵Full CP is a version of CP that does not require to split the data, at the cost of a bigger computational burden. This is the reason why we do not consider it in this work, along with the fact that full CP can be plugged in on an existing pipeline, making it particularly appealing for operational purposes. The interested reader on full CP can have a look at ([Vovk et al., 2005](#))

⁶For a horizon $h \neq 1$, then $\hat{q}_\ell^{-(t)}, \hat{q}_u^{-(t)}$ are fitted on $\{(X_{t-|\text{Tr}|}, Y_{t-|\text{Tr}|}), \dots, (X_{t-h}, Y_{t-h})\}$.

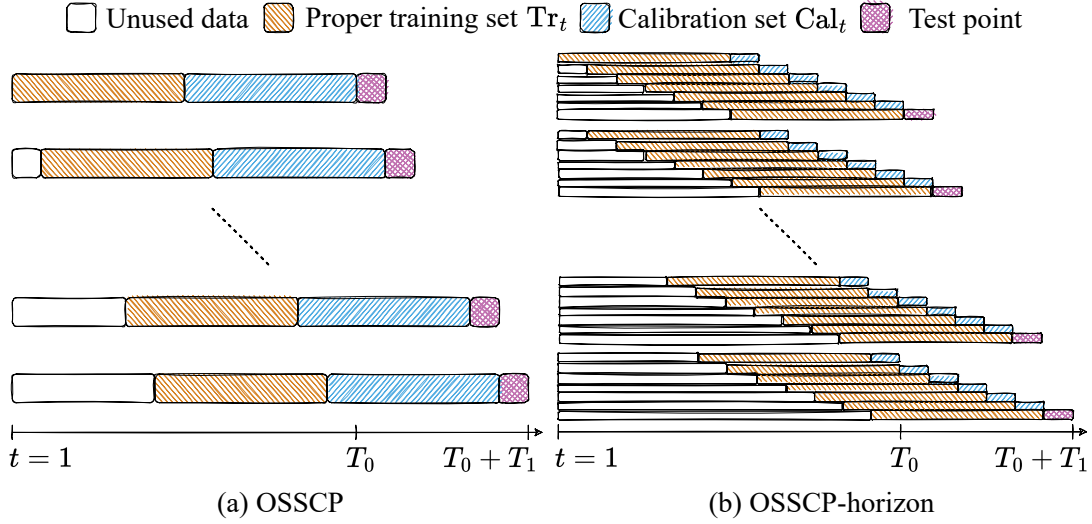


Figure 6.5: Scheme of OSSCP (a) and our proposal (b), OSSCP-horizon, when the horizon is 1.

After having built $\mathcal{S}_{\text{Cal}_T} = \{S_{T-|\text{Cal}|+1}, \dots, S_T\}$, OSSCP-horizon computes the PI for the test point X_{T+1} :

$$\widehat{C}_\alpha(X_{T+1}) := \left[\hat{q}_\ell^{-(T+1)}(X_{T+1}) - Q_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}_T}); \right. \\ \left. \hat{q}_u^{-(T+1)}(X_{T+1}) + Q_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}_T}) \right].$$

Again, while demonstrating empirical improvements upon standard OSSCP in the temporal setting, OSSCP-horizon does not enjoy any form of theoretical guarantees. To theoretically account for the online setting, a popular method is Adaptive Conformal Inference (ACI) (Gibbs and Candès, 2021).

6.4.2.2 ADAPTIVE CONFORMAL INFERENCE (ACI)

Proposed in (Gibbs and Candès, 2021), ACI adapts CP to an arbitrary online setting, including temporal distribution shifts. To do so, ACI recursively updates the *effective* miscoverage rate $\tilde{\alpha} := \alpha_t$ used in the computation of the PI. Set $\alpha_1 = \alpha$. For $t \geq T_0$, and for a chosen $\gamma \geq 0$ the ACI update formula is:

$$\begin{cases} \widehat{C}_{\alpha_t}(X_t) := [\hat{q}_\ell(X_t) - Q_{1-\alpha_t}(\mathcal{S}_{\text{Cal}_t}), \hat{q}_u(X_t) + Q_{1-\alpha_t}(\mathcal{S}_{\text{Cal}_t})] \\ \alpha_{t+1} = \alpha_t + \gamma \left(\alpha - \mathbb{1}_{\{Y_t \notin \widehat{C}_{\alpha_t}(X_t)\}} \right) \end{cases}$$

The underlying idea is the following. If the PI does not cover at time t , then $\alpha_{t+1} \leq \alpha_t$ which increases the size of the PI. Conversely, the size of the interval decreases gently at time $t+1$ when it covers at time t . As noted in (Zaffran et al., 2022), it is possible to have $\alpha_t \geq 1$ or $\alpha_t \leq 0$: the former case is quite rare and produces by convention $\widehat{C}_{\alpha_t} = [\hat{q}_\ell(\cdot), \hat{q}_u(\cdot)]$; however, the latter can happen frequently, especially for a high γ , giving a prediction interval of infinite size ($\widehat{C}_{\alpha_t} \equiv \mathbb{R}$).

The main theoretical result on ACI is that for any sequence $(X_t, Y_t)_t$,

$$\left| \frac{1}{T_1 - T_0} \sum_{t=T_0+1}^{T_1} \mathbb{1}_{\{y_t \in \widehat{C}_{\alpha_t}(X_t)\}} - (1 - \alpha) \right| \leq \frac{2}{\gamma(T_1 - T_0)}.$$

It shows the asymptotically valid frequency of ACI intervals for any arbitrary (possibly adversarial) distribution.

Note that the convergence rate is in γ^{-1} , hence favoring large γ which are the ones leading to more variability and in the extreme case to infinite PIs (discussed previously). This illustrates the need for guidance on how to choose properly γ , and even avoid having to choose it and being able to switch between different γ depending on the current data distribution's evolution.

6.4.2.3 AgACI

The goal of AgACI, proposed in (Zaffran et al., 2022), is precisely to provide a parameter-free method based on ACI, that can adapt to temporal changes in the data distribution adaptively. Given a list of K γ values $\{\gamma_k\}_{k=1}^K$, AgACI works as an adaptive aggregation of experts (Cesa-Bianchi and Lugosi, 2006) (see also Section 6.4.1), with expert k being ACI with parameter γ_k . At each prediction step t , it performs two independent aggregations of the K ACI intervals $\widehat{C}_{\alpha_{t,k}}(\cdot) \stackrel{\text{not.}}{=} [\hat{b}_{t,k}^{(\ell)}(\cdot), \hat{b}_{t,k}^{(u)}(\cdot)]$, one for each bound, and outputs $\tilde{C}_t(\cdot) \stackrel{\text{not.}}{=} [\tilde{b}_t^{(\ell)}(\cdot), \tilde{b}_t^{(u)}(\cdot)]$. According to Zaffran et al. (2022), the standard different aggregation rules gave similar results. In this work, we restrict ourselves to the setting of (Zaffran et al., 2022), that is BOA, with the gradient trick.

6.4.2.4 LATEST RELATED WORKS

Since the analysis presented in this paper was performed, the line of research on adaptive and online conformal approaches has been expanding fast. Recent developments include: Gibbs and Candès (2023) improving on ACI by online aggregation on a grid of different γ , similarly to AgACI, at the crucial difference that the aggregation is on the value of α_t and not on the lower and upper bounds independently (Section 6.5.2 highlights why we argue in favor of different aggregations); Bastani et al. (2022) who achieve stronger coverage guarantees (conditional on the effective level, and conditional on specified subsets of the explanatory variables); Bhatnagar et al. (2023) enjoy anytime regret bound, by leveraging tools from the strongly adaptive regret minimization literature; Angelopoulos et al. (2023) who extend upon ACI ideas by relying on control theory to add more information on the temporal structure; Angelopoulos et al. (2024) proposing to use adaptive learning rates γ_t in ACI.

Our goal in this analysis is to deeply investigate the improvements, or not, brought by conformal as one of the layers for probabilistic forecasts with an operational lens. Therefore, we restricted the study to OSSCP, OSSCP-horizon, and AgACI as it has already shown benefits on electricity prices and does not require to select any hyper-parameter (Zaffran et al., 2022). Indeed, it allows us to easily understand what is the cause of the improved or declined performance. Furthermore, the most recent works are either complex structures (thus less interpretable) or depend on hyper-parameter tuning, making them more costly to implement in operational use.

6.5 Application and results

6.5.1 Setting and evaluation

Experimental details In order to span a wide range of the price distribution function, we vary the PIs' miscoverage level $1 - \alpha > 0.6$. For the final probabilistic forecasts, the overall training set comprises 4 years of data, from 2016 to 2019 included (i.e. merging the training and validation sets).

Due to training time constraints, we trained and evaluated the considered models on hours 3, 8, 13, 18, and 23 of every day. These 5 hours encompass best the different phases of hourly electricity prices in a given day, while uniformly covering the 24 hours of the day.

Finally, due to the high non-stationarity, we trained each of the base models presented in Section 6.3.2 on different window sizes: approximately 4 years, 3 years, 2 years, 1 year, 270 days, 180 days, and 90 days. For the sake of clarity, for each analysis performed, the largest window size will be selected and presented in this paper. In the same vein, the calibration size of the conformal approaches (Sections 6.3.3 and 6.4.2) varies among 25%, 50% and 75% of the overall windowed training set. Again, to ease interpretation of our results, we present here only the results for a calibration set of proportion 50% (except if stated otherwise) as it allows for an intermediary adaptation speed, hence being a good trade-off between up-to-date quantile regression models and calibration set large enough to perform the estimation of the highly non-stationary conformal correction. We recall that in the i.i.d. setting a general rule of thumb for the calibration size is around 25% (Sesia and Candès, 2020). In our study, the impact of non-stationarity induces a need for a trade-off between adaptivity and the calibration window length.

Evaluation procedure The main challenge of evaluating a probabilistic forecast is that the true distribution of the underlying process cannot be observed. Hence, it is impossible

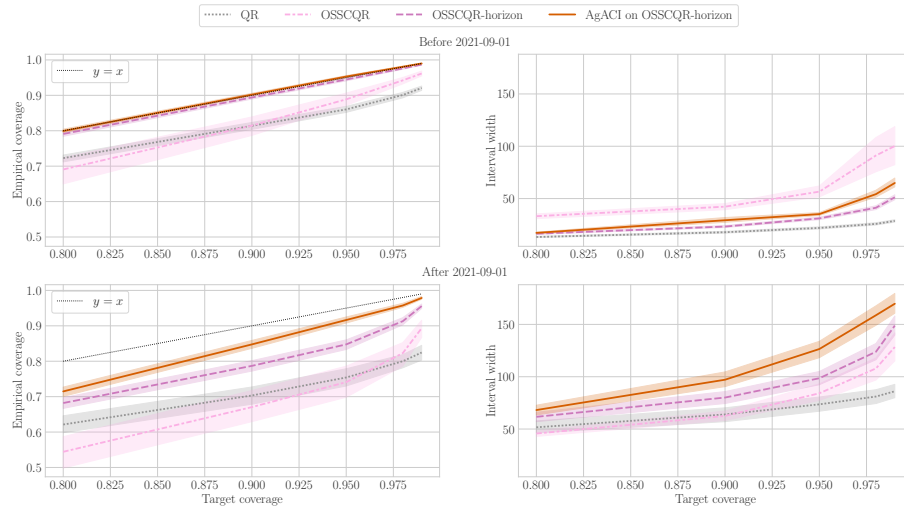


Figure 6.6: PIs's performances with different levels of conformalisation on the **quantile linear model**, before September 2021 (top row) and after September 2021 (bottom row), for various target coverage levels (x -axis). The colors and shapes are associated with the conformalisation layers. The shaded regions correspond to the 5% and 95% empirical quantiles after bootstrapping 500 times the test time series.

to compare the estimated distribution with the actual distribution of the true spot prices. This is not the case for a sequence of PIs $\left(\left[\hat{b}^{(\ell)}(\cdot), \hat{b}^{(u)}(\cdot)\right]\right)_t$ that can be evaluated through:

- **empirical average coverage**,

$\frac{1}{T_1 - T_0} \sum_{t=T_0+1}^{T_1} \mathbb{1} \left\{ y_t \in \left[\hat{b}^{(\ell)}(x_t), \hat{b}^{(u)}(x_t)\right] \right\}$, that should be close and above to the target level $1 - \alpha$ for *validity* (also known as *reliability*),

- **empirical average length**, $\frac{1}{T_1 - T_0} \sum_{t=T_0+1}^{T_1} \hat{b}^{(u)}(x_t) - \hat{b}^{(\ell)}(x_t)$, for *efficiency*⁷ (also known as *sharpness*).

For each of these metrics, confidence intervals are constructed by time series bootstrapping (non-overlapping moving block bootstrap) (Kunsch, 1989; Politis and Romano, 1994).

Results on the CRPS are provided in 6.A. Indeed, our goal is really to compare PIs and not predictive distributions. Therefore, the forecasts' objective is truly to **be as sharp as possible while satisfying validity**.

6.5.2 Results

Impact of the conformalisations In Figures 6.6 and 6.7 we represent the performance of Linear Quantile Regression and Quantile Random Forest respectively, with various layers of conformalisation. The display choice of these two base models is motivated by the fact that they represent a diverse range of modelisation.

In both cases, we observe that a naive conformalisation – in the form of OSSCP – does not allow to achieve the nominal coverage level, neither before nor after September 2021.

Yet, our proposal OSSCP-**horizon** does improve drastically the coverage level: before September 2021 it manages to reach the target level while improving the lengths of the PIs, and after September 2021 it allows to reduce the gap with the target considerably (linear model), while recovering the approximatively satisfactory performances of the individual QRF that was deteriorated by OSSCP.

Finally, making the conformalisation even more adaptive through the use of AgACI especially enhances validity after September 2021. Yet, it has to be noted that it seems to be insufficiently adaptive to perfectly reach the target level.

Analysis of various aggregations Therefore, we go further and add another adaptive post-processing layer by performing online aggregation. In Figure 6.8 we compare the performances of various aggregations, each of them considering a different set of experts (individual forecasts, OSSSCP-**horizon** forecasts, AgACI forecasts, and all of them). As a baseline, we add the uniform average of all of these experts. For each of the aggregation, we compared aggregating forecasts with a unique window size for training with aggregating forecasts with multiple training window size (hence augmenting the number of experts in

⁷Indeed, achieving exactly $1 - \alpha$ coverage can be trivially done by outputting $1 - \alpha$ of the time \mathbb{R} and the empty set otherwise, which is critically uninformative. Thus, one wants to attain *validity* while minimizing the size of the resulting intervals, that is maximizing *efficiency*.

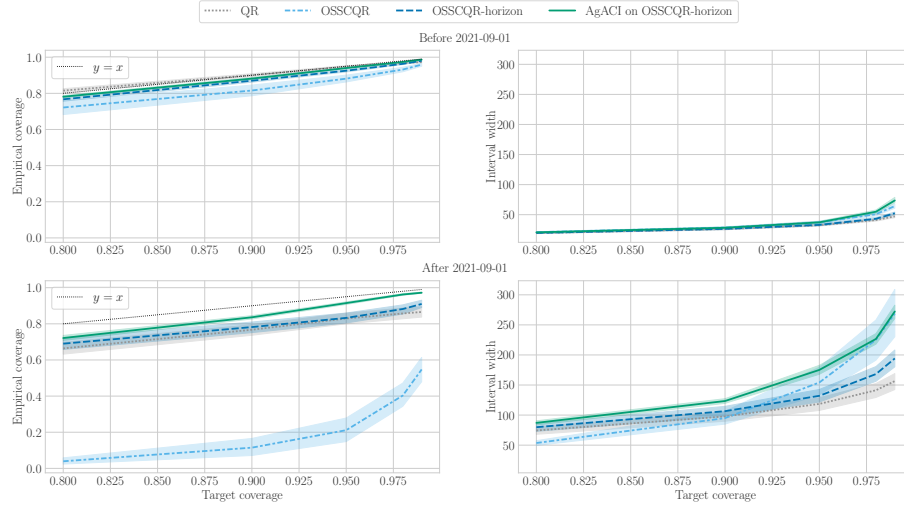


Figure 6.7: Same caption than Figure 6.6 but for the **quantile random forest model**.

the set). This latter strategy is usually referred to as *windowing* (Marcjasz et al., 2018). We selected the best aggregation (namely aggregating AgACI forecasts with windowing) and, for the sake of readability and for coherence, we displayed in Figure 6.8 all the aggregations with windowing. It has to be noted that there is a lot of variability, as it can be seen in Figure 6.8, and that for some aggregation the best choice was in fact without windowing.

Figure 6.8 highlights that online aggregation improves considerably the robustness to non-stationarity in terms of validity. Furthermore, after September 2021, online aggregation on AgACI forecasts enhances the sharpness of the forecasts with respect to the uniform average, that has similar coverage. This can be explained by the fact that the individual performances degrade in this non-stationary environment, leading to aggregation's weights close to uniform so as to minimise the risk (as we will also see in the next analysis).

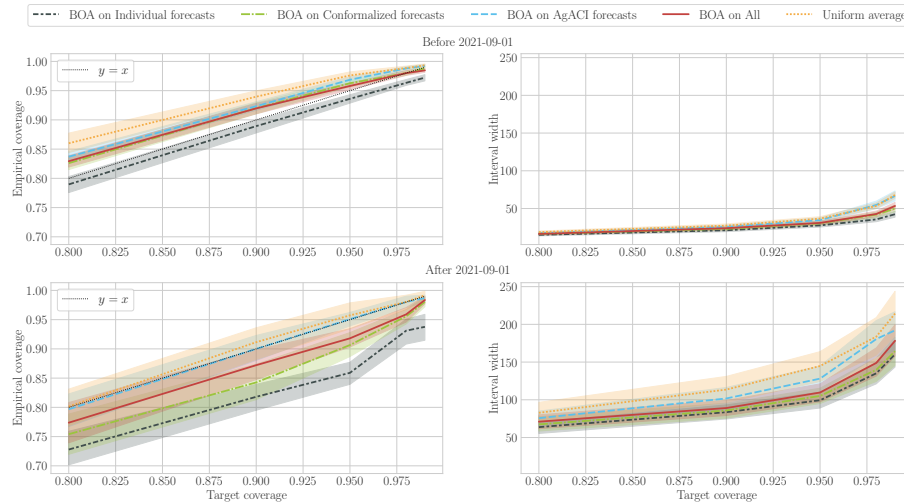


Figure 6.8: PIs's performances of online aggregation on multiple set of experts with windowing, before September 2021 (top row) and after September 2021 (bottom row), for various target coverage levels (x -axis). The colors and shapes are associated with the set of experts. The shaded regions correspond to the 5% and 95% empirical quantiles after bootstrapping 500 times the test time series.

Analysis of aggregation of various AgACI: applying the best conformalisation possible (AgACI) on each model and then aggregating them

In Figure 6.9 we represent the evolution of the weights associated to each of the AgACI (the color representing the base model, and the shade of it indicating the calibration percentage) with time x -axis, for various coverage level (columns). To improve readability, we display these weights for the aggregation without windowing.

The first striking observation is the presence of temporal ruptures in the weights' distribution. They are informative as they are associated with domain phenomena, which depend on the considered bound (lower or upper). Particularly, the first one happening is the big negative spike in Easter 2020 (April 13, 2020, see top row of Figure 6.1) due to both the public holiday and the Covid-19 lockdown. This especially affects the lower bound. The second one occurs in the second fortnight of September 2020 when the first extreme positive peaks take place, impacting the upper bound. These positive spikes are mainly due to a very low wind generation in France (less than 1 GW) and more generally in Europe, along with a French nuclear production well below its level of previous years at the same time. The last significant rupture is around October 2021, when spot prices start to rise drastically and get more and more volatile, corresponding to the increase in level and volatility of gas and carbon emission prices. This one affects both the lower and upper bounds. In particular, the weights' distribution becomes uniform after this rupture, which is expected in a setting where the aggregation tries to minimize the risk with experts performing poorly.

The second observation is that the methods on which the aggregation places the most of the weights is different depending on the bound: remarkably, at the levels 0.95 and 0.98, the lower bound places high mass on quantile random forests, while the upper bound relies more on qgam. This can be explained by the fact that the various methods depend differently on the provided features: additive models such as qgam or linear ones have a great extrapolation ability, while random forests and gradient boosting benefit from more flexibility on features' interaction modeling. This idea is also reflected in Figures 6.2 and 6.3 comparing the feature importance in Lasso with the one of Random forest.

Lastly, for high levels of coverage such as 0.95 and 0.98, the aggregation also places weights on different training size depending on the bound. While the upper bound favors small training size, the lower bound encourages large training size. This might be due to the effective sample size which is required to appropriately learn the lower quantiles of the prices, which are less impacted by the non-stationarity; while the upper bound is particularly complex to model, and having more data points correct the predictive model through conformalisation might be a better usage of the available data.

These three key observations argue in favor aggregating independently the upper and lower bounds.

6.6 Conclusion and perspectives

In this study, we have analysed the performances of a wide range of probabilistic methods in a particularly challenging task: forecasting electricity spot prices in France in 2020 and

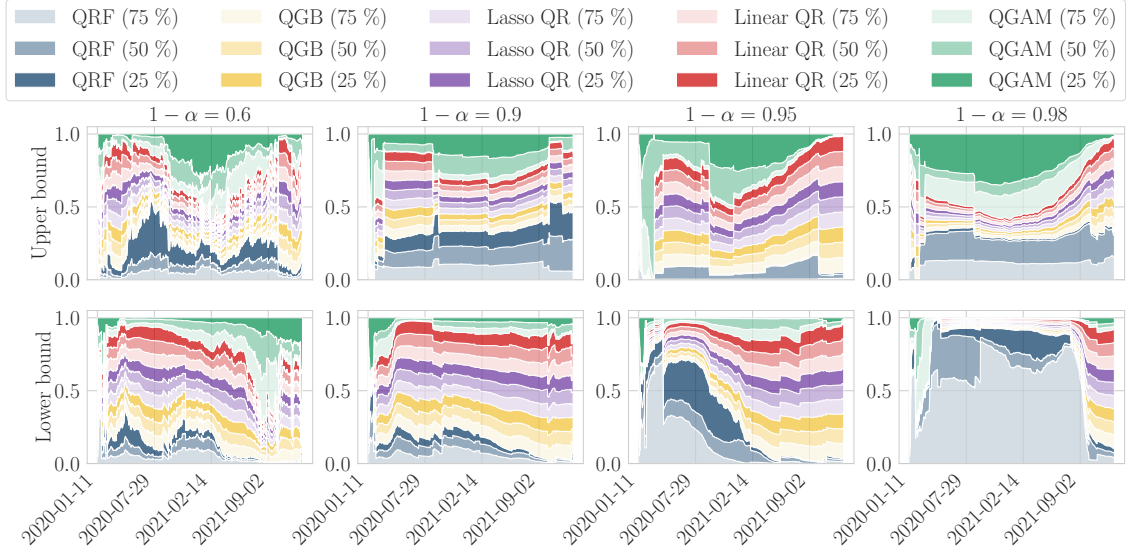


Figure 6.9: Temporal evolution (x -axis) of the weights associated with each expert in the online aggregation, for different values of (columns). The top row (resp. bottom row) shows the weights assigned for the upper (resp. lower) bound forecast. The colors correspond to the base model on which AgACI is applied to, and the transparency to the proportion of training data kept for actually fitting these base models.

2021. On the design, we have highlighted the importance of including the new explanatory variable corresponding to the nuclear plants' availability. We were also able to bring new insights into the post-processing of individual forecasts, such as conformalisation or aggregation. Indeed, our extensive experiments demonstrate that *i*) conformalisation, when appropriately done as through **OSSCP-horizon**, considerably improves PI's quality despite the non-stationarity, *ii*) online aggregation of experts is extremely powerful in terms of adaptiveness bringing enhanced PI's performances and taking advantage of windowing, *iii*) combining both conformalisation and online aggregation appears on this data set to be the best strategy, and most importantly sheds light on many domain phenomena thanks to great interpretability.

There are many avenues for future works. From the electricity lens, the prices have continued to evolve significantly since 2022 and pursuing the study on newer data would undoubtedly yield new knowledge. Speaking of which, our study did not investigate the crucial question of peaks and extreme forecasts, dominant in electricity prices. Works on online procedure tailored for extremes have already been deployed ([Himych et al., 2024](#)), and it might be relevant to see how it can be paired with conformal approaches. Another natural perspective that would deepen our understanding on the benefits of conformalisation is to conformalize the aggregated models as suggested in [Susmann et al. \(2024\)](#), as opposed to aggregating the conformalized models which is what we performed. It would also be interesting to assess the performances of the most recent online conformal algorithms (listed in Section 6.4.2.4), that might be better suited for non-stationarity. Finally, our angle of approach is to showcase the advantages of black-box plugs-in such as CP and aggregation. It is attractive to couple it with recent developments that enhance the interpretability of complex statistical models, such as [Wood et al. \(2022\)](#).

Appendix to Adaptive Probabilistic Forecasting of French Electricity Spot Prices in 2020 and 2021

6.A Results on the CRPS

To assess the performance of a probabilistic method on the overall range of quantiles, one can use the Continuous Ranked Probability Score (CRPS). This score is originally described in terms of the predictive CDS $\hat{F}_{d,h}$:

$$CRPS(\hat{F}_{d,h}, y_{d,h}) = \int_{-\infty}^{\infty} \left(\hat{F}_{d,h}(y|x_{d,h}) - \mathbb{1}_{\{y_{d,h} \leq y\}} \right)^2 dy.$$

Interestingly, the CRPS can be reformulated (to a multiplicative constant) as :

$$CRPS(\hat{F}_{d,h}, y_{d,h}) = \int_0^1 \rho_{\alpha} \left(y_{d,h}, \hat{F}_{d,h}^{-1}(\alpha) \right) d\alpha,$$

where $\hat{F}_{d,h}^{-1}(\alpha)$ actually corresponds to the predicted value at quantile α . By approximating this integral as a Riemann sum, we can transform pinball scores over multiple quantiles into one single metric.

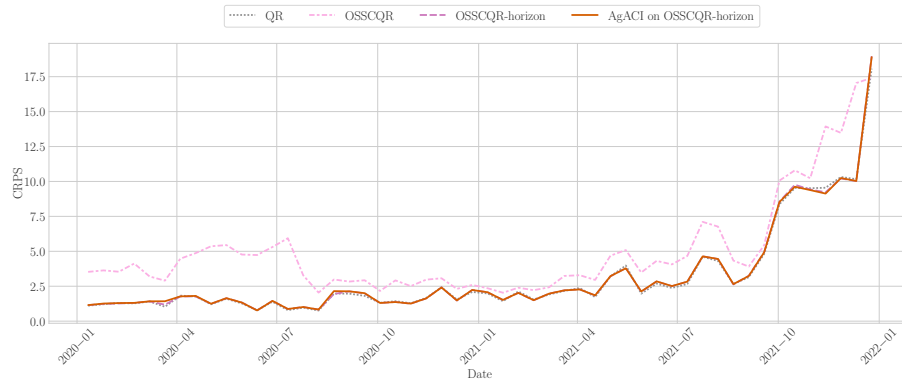


Figure 6.10: PIs's CRPS with different levels of conformalisation on the **quantile linear model**, depending on the time. The colors and shapes are associated with the conformalisation layers.

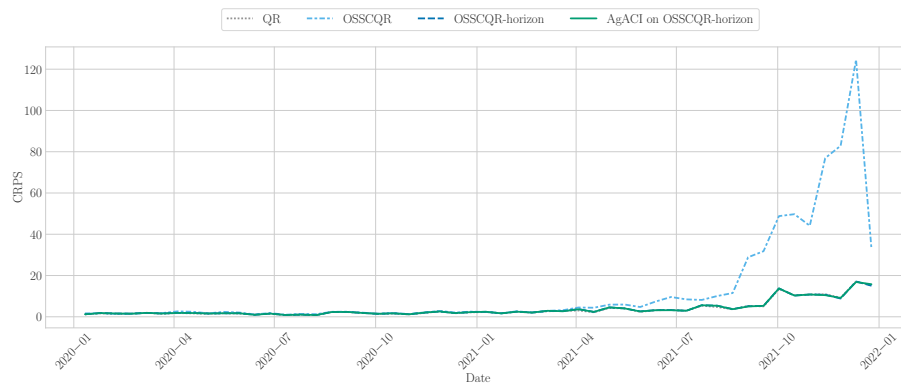


Figure 6.11: Same caption than Figure 6.10 but for the **quantile random forest model**.

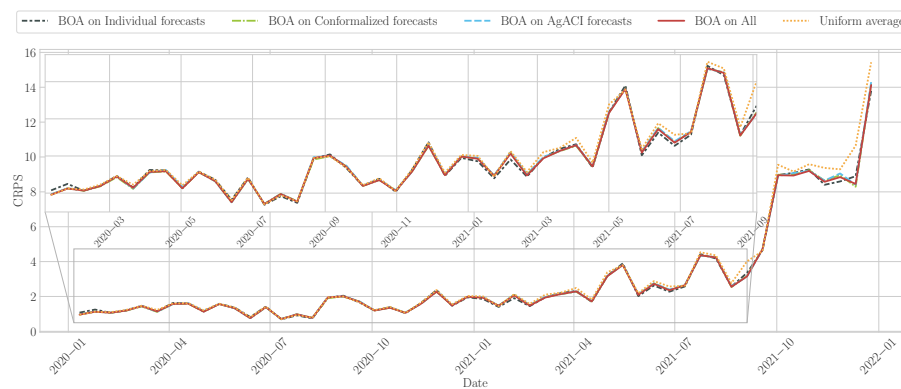
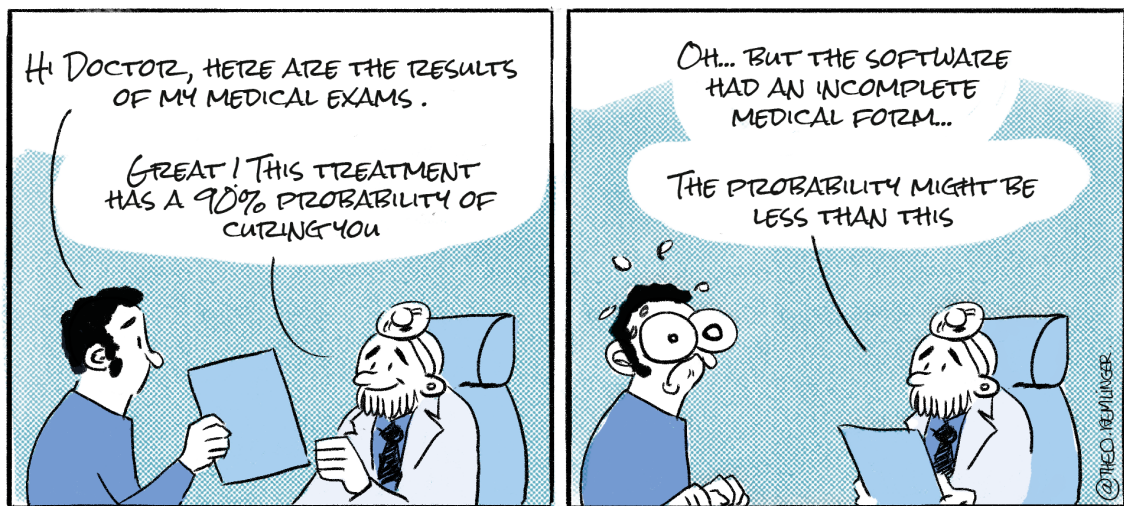


Figure 6.12: PIs's CRPS of online aggregation on multiple set of experts with windowing, depending on the time. The colors and shapes are associated with the set of experts.

Part III

Missing Values



Chapter 7

Conformal Prediction with Missing Values

Conformal prediction is a theoretically grounded framework for constructing predictive intervals. We study conformal prediction with missing values in the covariates – a setting that brings new challenges to uncertainty quantification. We first show that the marginal coverage guarantee of conformal prediction holds on imputed data for any missingness distribution and almost all imputation functions. However, we emphasize that the average coverage varies depending on the pattern of missing values: conformal methods tend to construct prediction intervals that under-cover the response conditionally to some missing patterns. This motivates our novel generalized conformalized quantile regression framework, missing data augmentation, which yields prediction intervals that are valid conditionally to the patterns of missing values, despite their exponential number. We then show that a universally consistent quantile regression algorithm trained on the imputed data is Bayes optimal for the pinball risk, thus achieving valid coverage conditionally to any given data point. Moreover, we examine the case of a linear model, which demonstrates the importance of our proposal in overcoming the heteroskedasticity induced by missing values. Using synthetic and data from critical care, we corroborate our theory and report improved performance of our methods.

Contents

7.1	Introduction	129
7.2	Background	131
7.3	Warm-up: marginal coverage with NAs	132
7.4	Challenge: NAs induce heteroskedasticity	133
7.5	Achieving mask-conditional-validity (MCV)	134
7.5.1	Missing Data Augmentation (MDA)	135
7.5.2	Theoretical guarantees in finite sample	137
7.6	Towards asymptotic individualized coverage	138
7.7	Empirical study	139
7.7.1	Synthetic experiments: Gaussian linear data	140
7.7.2	Semi-synthetic experiments	141
7.7.3	Predicting the level of platelets for trauma patients	142
7.8	Conclusion and perspectives	143
7.A	Detailed perspective discussion	144
7.B	Illustration and details on CQR (Romano et al., 2019) procedure	145
7.C	Impute-then-predict+conformalization	147
7.D	Gaussian linear model	148
7.E	Finite sample algorithms	151
7.F	Infinite data results	156
7.G	Experimental study	159

7.1 Introduction

By leveraging increasingly large data sets, statistical algorithms and machine learning methods can be used to support high-stakes decision-making problems such as autonomous driving, medical or civic applications, and more. To ensure the safe deployment of predictive models it is crucial to quantify the uncertainty of the resulting predictions, communicating the limits of predictive performance. Uncertainty quantification attracts a lot of attention in recent years, particularly methods that are based on Conformal Prediction (CP) (Vovk et al., 2005; Papadopoulos et al., 2002; Lei et al., 2018). CP provides controlled predictive regions for any underlying predictive algorithm (e.g., neural networks and random forests), in finite samples with no assumption on the data distribution except for the exchangeability of the train and test data. More precisely, for a *miscoverage rate* $\alpha \in [0, 1]$, CP outputs a *marginally valid* prediction interval \hat{C}_α for the test response Y given its corresponding covariates X , that is:

$$\mathbb{P}(Y \in \hat{C}_\alpha(X)) \geq 1 - \alpha. \quad (7.1)$$

Split CP (Papadopoulos et al., 2002; Lei et al., 2018) achieves Eq. (7.1) by keeping a hold-out set, the *calibration set*, used to evaluate the performance of a fixed predictive model.

At the same time, as the volume of data increases, the volume of missing values also increases. There is a vast literature on this topic (Little, 2019; Josse and Reiter, 2018), and a recent survey even identified more than 150 different implementations (Mayer et al., 2019). Missing values create additional challenges to the task of supervised learning, as traditional machine learning algorithms can not handle incomplete data (Josse et al., 2019; Le Morvan et al., 2020b,a, 2021; Ayme et al., 2022; Van Ness et al., 2022). One of the most popular strategies to deal with missing values suggests imputing the missing entries with plausible values to get completed data, on which any analysis can be performed. The drawback of this “impute-then-predict” approach is that single imputation can distort the joint and marginal distribution of the data. Yet, Josse et al. (2019); Le Morvan et al. (2020b, 2021) showed that such impute-then-predict strategies are Bayes consistent, under the assumption that a universally consistent learner is applied on an imputed data set. However, this line of work focuses on point prediction with missing values that aim to predict the most likely outcome. In contrast, our goal is quantifying predictive uncertainty, which was not explored with missing values although its enormous importance.

Contributions.

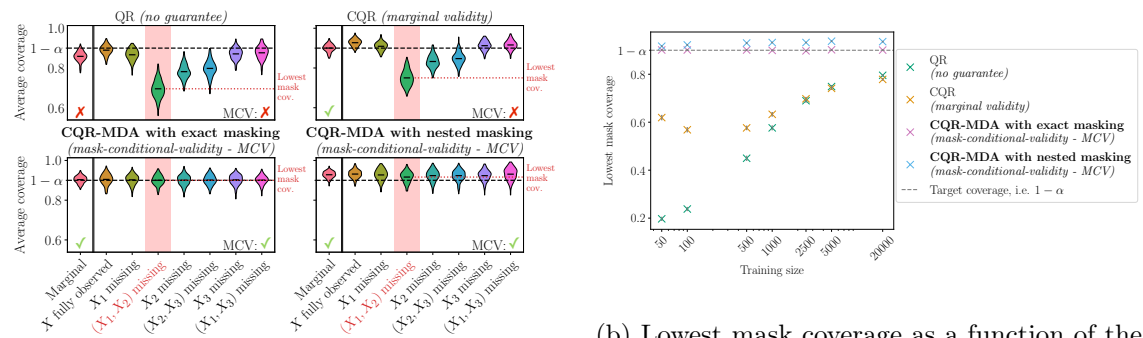
We study CP with missing covariates. Specifically, we study downstream quantile regression (QR) based CP, like CQR (Romano et al., 2019), on impute-then-predict strategies. Still, the proposed approaches also encapsulate other regression basemodels, and even classification.

After setting background in Section 7.2, our first contribution is showing that CP on impute-then-predict is *marginally valid* regardless of the model, missingness distribution, and imputation function (Section 7.3).

Then, we focus on the specificity of uncertainty quantification *with missing values*. In Section 7.4, we describe how different masks (i.e. the set of observed features) introduce additional heteroskedasticity: *the uncertainty on the output strongly depends on the set of predictive features observed*. We therefore focus on achieving valid coverage *conditionally on the mask*, coined MCV – Mask-Conditional-Validity. MCV is desirable in practice, as occurrence of missing values are linked to important attributes (see Section 7.5).

Traditional approaches such as QR and CQR fail to achieve MCV because they do not account for this core connection between missing values and uncertainty. This is illustrated on synthetic data in Figure 7.1. In Figure 7.1a, a toy example with only 3 features, thus $2^3 - 1 = 7$ possible masks, shows how the coverage of QR and CQR varies depending on the mask. Both methods dramatically undercover when the most important variable (X_2) is missing, and the loss of coverage worsens when additional features are missing. In particular, for each method, one mask (X_1 and X_2 missing, highlighted in red) leads to the *lowest mask coverage*. Achieving MCV corresponds to a lowest mask coverage greater than $1 - \alpha$. In Figure 7.1b, the dimension is 10: instead of the $2^{10} - 1 = 1023$ different masks, we only report the lowest mask coverage for increasing sample sizes. It highlights that QR (green \times) and CQR (orange \times) do not meet the lowest mask coverage target of 90%, even for large sample sizes.

This motivates our second contribution: we show in Section 7.5 how to form prediction intervals that are MCV. This is highly challenging since there are exponentially many possible patterns to consider. Therefore, the naive solution to perform a calibration for each mask would fail as in finite samples, we often observe test samples with a mask that have low (or even null) frequency of appearance in the calibration set. To tackle this issue, we suggest two conformal methods that share the same core idea of missing data augmentation (MDA): the calibration data is artificially masked to match the mask of the point we consider at test time. The first method, *CP-MDA with exact masking*, relies on building an ideal calibration set for which the data points have the exact same mask as



(a) Coverage of the predictive intervals depending on which features are missing, among the 3 features. Evaluation over 200 runs.

(b) Lowest mask coverage as a function of the training size. Results evaluated over 100 repetitions, and the (tiny) error bars correspond to standard errors.

Figure 7.1: Methods are Quantile Regression (QR), Conformalized Quantile Regression (CQR), and two novel procedures **CP-MDA-Exact** and **CP-MDA-Nested**, on top of CQR. Settings are given in Section 7.7, in a nutshell: data follows a Gaussian linear model where missing values are independent of everything else and of proportion 20%; the dimension of the problem is 3 in Figure 7.1a while in 7.1b it is 10.

of the test point. We show its MCV under exchangeability and Missing Completely At Random assumptions. Our second method, *CP-MDA with nested masking*, does not require such an ideal calibration set. Instead, we artificially construct a calibration set in which the data points have *at least* the same mask as the test point, i.e., this artificial masking results in calibration points having possibly more missing values than the test point. We show the latter method also achieves the desired coverage conditional on the mask, but at the cost of an additional assumption for validity: stochastic domination of the quantiles. Figure 7.1 illustrates those findings: both methods are MCV, as their lowest mask coverage is above $1 - \alpha$.

Our third contribution further supports our design choice to use QR. We show that QR on impute-then-predict strategy is Bayes-consistent – it can achieve the strongest form of coverage conditional on the observed test features (Section 7.6).

Lastly, we support our proposal using both (semi)-synthetic experiments and real medical data (Section 7.7). The code to reproduce our experiments is available on [GitHub](#).

7.2 Background

Background on missing values. Consider a data set with n exchangeable realizations of the random variable $(X, M, Y) \in \mathbb{R}^d \times \{0, 1\}^d \times \mathbb{R}$: $\{(X^{(k)}, M^{(k)}, Y^{(k)})\}_{k=1}^n$, where X represents the features, M the missing pattern, or mask, and Y an outcome to predict. For $j \in \llbracket 1, d \rrbracket$, $M_j = 0$ when X_j is observed and $M_j = 1$ when X_j is missing, i.e. **NA** (Not Available). We note $\mathcal{M} = \{0, 1\}^d$ the set of masks. For a pattern $m \in \mathcal{M}$, $X_{\text{obs}(m)}$ is the random vector of observed components, and $X_{\text{mis}(m)}$ is the random vector of unobserved ones. For example, if we observe **(NA, 6, 2)** then $m = (1, 0, 0)$ and $X_{\text{obs}(m)} = (6, 2)$. Our goal is to predict a new outcome $Y^{(n+1)}$ given $X_{\text{obs}(M^{(n+1)})}^{(n+1)}$ and $M^{(n+1)}$.

Assumption A1 (exchangeability). The random variables $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^{n+1}$ are exchangeable.

Following [Rubin \(1976\)](#), we consider three well-known missingness mechanisms.

Definition 7.2.1 (Missing Completely At Random (MCAR)).

For any $m \in \mathcal{M}$, $\mathbb{P}(M = m|X) = \mathbb{P}(M = m)$.

Definition 7.2.2 (Missing At Random (MAR)).

For any $m \in \mathcal{M}$, $\mathbb{P}(M = m|X) = \mathbb{P}(M = m|X_{\text{obs}(m)})$.

Definition 7.2.3 (Missing Non At Random (MNAR)). If the missing data is not MAR, it is MNAR. Thus, its probability distribution depends on X , including the missing values.

Impute-then-predict. As most predictive algorithms can not directly handle missing values, we impute the incomplete data using an imputation function Φ which maps observed values to themselves and missing values to a function of the observed values. With notations from [Le Morvan et al. \(2021\)](#) we note $\varphi^m : \mathbb{R}^{|\text{obs}(m)|} \rightarrow \mathbb{R}^{|\text{mis}(m)|}$ the imputation function which takes as input observed values and outputs imputed values, i.e.

plausible values, given a mask $m \in \mathcal{M}$. Then, the imputation function Φ belongs to $\mathcal{F}^I := \left\{ \Phi : \mathbb{R}^d \times \mathcal{M} \rightarrow \mathbb{R}^d : \forall j \in \llbracket 1, d \rrbracket, \Phi_j(X, M) = X_j \mathbb{1}_{M_j=0} + \varphi_j^M(X_{\text{obs}(M)}) \mathbb{1}_{M_j=1} \right\}$. Additionally, \mathcal{F}_∞^I is the restriction of \mathcal{F}^I to \mathcal{C}^∞ functions which include deterministic imputation, such as mean imputation or imputation by regression. The imputed data set is formed by the realizations of the n random variables $(\Phi(X, M), M, Y)$. In practice, Φ is obtained as the result of an algorithm \mathcal{I} trained on $\{(X^{(k)}, M^{(k)})\}_{k=1}^{n+1}$.

Assumption A2 (Symmetrical imputation). The imputation function Φ is the output of an algorithm \mathcal{I} treating its input data points symmetrically: $\mathcal{I}((X^{(\sigma(k))}, M^{(\sigma(k))})_{k=1}^{n+1}) \stackrel{(d)}{=} \mathcal{I}((X^{(k)}, M^{(k)})_{k=1}^{n+1})$ conditionally on $(X^{(k)}, M^{(k)})_{k=1}^{n+1}$ and for any permutation σ on $\llbracket 1, n+1 \rrbracket$.

Assumption A2 is very mild and satisfied by all existing imputation methods for exchangeable data. In particular, it is valid for iterative regression imputation which allows out-of-sample imputation.

Background on (split) conformal prediction. Split, or inductive, CP (SCP) (Papadopoulos et al., 2002; Lei et al., 2018) builds predictive regions by first splitting the n points of the training set into two disjoint sets $\text{Tr}, \text{Cal} \subset \llbracket 1, n \rrbracket$, to create a *proper training set*, Tr , and a *calibration set*, Cal . On the proper training set, a model \hat{f} (chosen by the user) is fitted, and then used to predict on the calibration set. *Conformity scores* $S_{\text{Cal}} = \{(s(X^{(k)}, Y^{(k)}))_{k \in \text{Cal}}\}$ are computed to assess how well the fitted model \hat{f} predicts the response values of the calibration points. For example, Conformalized Quantile Regression (CQR, Romano et al., 2019) fits two quantile regressions \hat{q}_{low} and \hat{q}_{upp} , on the proper training set. The conformity scores are defined by $s(x, y) = \max(\hat{q}_{\text{low}}(x) - y, y - \hat{q}_{\text{upp}}(x))$. Finally, a corrected $(1 - \tilde{\alpha})$ -th quantile of these scores $\hat{Q}_{1-\tilde{\alpha}}(S_{\text{Cal}})$ is computed (called *correction term*) to define the predictive region: $\hat{C}_\alpha(x) := \{y \text{ such that } s(y, \hat{f}(x)) \leq \hat{Q}_{1-\tilde{\alpha}}(S_{\text{Cal}})\}$.¹ An illustration of CQR is provided in Section 7.B.

This procedure satisfies Eq. (7.1) for any \hat{f} , any (finite) sample size n , as long as the data points are exchangeable.² Moreover, if the scores are almost surely distinct, the coverage holds almost exactly: $\mathbb{P}(Y \in \hat{C}_\alpha(X)) \leq 1 - \alpha + \frac{1}{\#\text{Cal}+1}$.

For more details on SCP, we refer to Angelopoulos and Bates (2023); Vovk et al. (2005), as well as to Manokhin (2022).

7.3 Warm-up: marginal coverage with NAs

A first idea to get valid predictive intervals $\hat{C}_\alpha(X, M)$ in the presence of missing values M is to apply CP in combination with impute-then-predict, which we refer to as *impute-then-predict+conformalization*. More details on this approach are given in Section 7.C.1 for both classification and regression tasks, although our main focus is regression. It turns out that such a simple approach is marginally (exactly) valid.

¹The correction $\alpha \rightarrow \tilde{\alpha}$ is needed because of the inflation of quantiles in finite sample (see Lemma 2 in Romano et al. (2019) or Section 2 in Lei et al. (2018)).

²Only the calibration and test data points need to be exchangeable.

Definition 7.3.1 (Marginal validity). A method outputting intervals \widehat{C}_α is marginally valid if the following lower bound is satisfied, and exactly valid if the following upper bound is also satisfied:

$$1 - \alpha \underset{\text{validity}}{\leq} \mathbb{P} \left(Y^{(n+1)} \in \widehat{C}_\alpha \left(X^{(n+1)}, M^{(n+1)} \right) \right) \underset{\text{exact validity}}{\leq} 1 - \alpha + \frac{1}{\#\text{Cal} + 1}.$$

Indeed, symmetric imputation preserves exchangeability.

Lemma 7.3.1 (Imputation preserves exchangeability). *Let A1 hold. Then, for any missing mechanism, for any imputation function Φ satisfying A2, the imputed random variables $(\Phi(X^{(k)}, M^{(k)}), M^{(k)}, Y^{(k)})_{k=1}^{n+1}$ are exchangeable.*

Note that if we replace A1 by an i.i.d. assumption, the imputed data set is only exchangeable but not i.i.d. without further assumptions on \mathcal{I} . Indeed, even simple mean imputation breaks independence.

Proposition 7.3.1 ((Exact) validity of impute-then-predict+conformalization). *If A1 and A2 are satisfied, impute-then-predict+conformalization is marginally valid. If moreover the scores are almost surely distinct, it is exactly valid.*

This is an important first positive result (proved in Section 7.C.2) showing that CP applied on an imputed data set has the same validity properties as on complete data, regardless of the missing value mechanism (MCAR, MAR or MNAR) and of the symmetric imputation scheme. Note that similar propositions could be derived for full CP (Vovk et al., 2005) and Jackknife+ (Barber et al., 2021b).

Proposition 7.3.1 complements the work by Yang (2015), that also guarantees *marginal* coverage for full CP, with the striking difference of having a complete training data.

7.4 Challenge: NAs induce heteroskedasticity

To better understand the interplay between missing values and conditional coverage with respect to the mask, we consider an illustrative example of a Gaussian linear model.

Model 7.4.1 (Gaussian linear model). The data is generated according to a linear model and the covariates are Gaussian conditionally to the pattern:

- $Y = \beta^T X + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \perp (X, M)$, $\beta \in \mathbb{R}^d$.
- for all $m \in \mathcal{M}$, there exist μ^m and Σ^m such that $X|(M = m) \sim \mathcal{N}(\mu^m, \Sigma^m)$.

In particular, Model 7.4.1 is verified when X is Gaussian and the missing data is MCAR. Model 7.4.1 is more general: it even includes MNAR examples (Ayme et al., 2022).

Proposition 7.4.1 (Oracle intervals). *The oracle predictive interval is defined as the smallest valid interval knowing $X_{\text{obs}(M)}$ and M . Under Model 7.4.1, its length only depends on the mask. For any $m \in \mathcal{M}$ this oracle length is:*

$$\mathcal{L}_\alpha^*(m) = 2q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sqrt{\beta_{\text{mis}(m)}^T \Sigma_{\text{mis|obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2}. \quad (7.2)$$

See Section 7.D for the definition of $\mu_{\text{mis|obs}}^m$ and $\Sigma_{\text{mis|obs}}^m$ and the quantiles of $Y|(X_{\text{obs}(m)}, M = m)$.

Eq. (7.2) stresses that even when the noise of the generative model is homoskedastic, *missing values induce heteroskedasticity*. Indeed, the covariance of the conditional distribution of $Y|(X_{\text{obs}(m)}, M = m)$ depends on m . Furthermore, the uncertainty increases when missing values are associated with larger regression coefficients (i.e. the most predictive variables): if $\beta_{\text{mis}(m)}$ is large, then $\mathcal{L}_\alpha^*(m)$ is also large, as $\Sigma_{\text{mis}|\text{obs}}^m$ is positive. In the extreme case where all the variables are missing, i.e. $m = (1, \dots, 1)$, $\mathcal{L}_\alpha^*(m) = 2q_{1-\frac{\alpha}{2}}^{N(0,1)} \sqrt{\beta \Sigma^m \beta^T + \sigma_\varepsilon^2} = q_{1-\frac{\alpha}{2}}^Y - q_{\frac{\alpha}{2}}^Y$. On the contrary, if $m = (0, \dots, 0)$ (that is all X_j are observed), $\beta_{\text{mis}(m)}$ is empty and $\mathcal{L}_\alpha^*(m) = 2q_{1-\frac{\alpha}{2}}^{N(0,1)} \sigma_\varepsilon = q_{1-\frac{\alpha}{2}}^\varepsilon - q_{\frac{\alpha}{2}}^\varepsilon$. We illustrate this induced heteroskedasticity and the impact of the predictive power in Figure 7.1a, and in Section 7.D along with a discussion emphasizing that even with the Bayes predictor for the conditional mean, mean-based CP does not yield intervals that are MCV.

The above analysis motivates the following two design choices we make in this work. First, we advocate working with QR models rather than classic regression ones, as the former can handle heteroskedastic data. Second, we recommend providing the mask information to the model in addition to the input covariates, as the mask may further encourage the model to construct an interval with a length adaptive to the given mask. Therefore, we focus on CQR (Romano et al., 2019)³, an adaptive version of SCP, and concatenate the mask to the features. However, the predictive intervals of this procedure may not necessarily provide valid coverage conditionally on the masks, especially in finite samples as shown in Figure 7.1b (orange crosses). This is because the quality of the prediction at some (X, M) depends strongly on M , as there is an exponential number of patterns (2^d) for a finite training size, whereas the correction term is calculated independently of the masks.

7.5 Achieving mask-conditional-validity (MCV)

We now aim at achieving *mask-conditional-validity* (MCV) defined as follows using an ordering on the masks.

Definition 7.5.1 (Included masks). Let $(\hat{m}, \check{m}) \in \mathcal{M}^2$, $\hat{m} \subset \check{m}$ if for any $j \in \llbracket 1, d \rrbracket$ such that $\hat{m}_j = 1$ then $\check{m}_j = 1$, i.e. \check{m} includes at least the same missing values than \hat{m} .

Definition 7.5.2 (MCV). A method is MCV if for any $m \in \mathcal{M}$ the following lower bound is satisfied, and exactly MCV if for any $m \in \mathcal{M}$ the following upper bound is also satisfied:

$$1 - \alpha \underset{\text{valid}}{\leq} \mathbb{P} \left(Y^{(n+1)} \in \hat{C}_\alpha \left(X^{(n+1)}, m \right) \mid M^{(n+1)} = m \right) \underset{\text{exactly valid}}{\leq} 1 - \alpha + \frac{1}{\#\text{Cal}^m + 1},$$

where $\text{Cal}^m = \{k \in \text{Cal} \text{ such that } m^{(k)} \subset m\}$.

On the relevance of MCV. In a medical application context, it is very common to have missing data completely at random (MCAR) when a measurement device fails or the medical team forgot to fill out some forms. As a general rule, from an *equity standpoint*, a patient whose data is missing should not be penalized (because of “bad luck”) by being

³Note that our proposed framework is not based on CQR, this is only one instance of it.

assigned a prediction interval that is less likely to include the true response than if the data were complete.

Furthermore, the mask can also be linked to an external unobserved feature corresponding to a meaningful category. Consider the problem of predicting a disease among a population. Aggregating data from multiple hospitals with different practices and measurement devices can imply different features are observed for each patient. This can be viewed as a MCAR setting when *identically distributed* patients⁴ are assigned an hospital at random. Patterns are then linked to the cities, that themselves are related to socio-economical data.

Overall, the missing patterns form *meaningful categories* and *ensuring MCV yields more equitable treatment*. Therefore, a method achieving marginal coverage by systematically failing on a given pattern, even in a MCAR setting, is not suitable. Finally, in non-MCAR cases, the pattern may be exactly related to critical discriminating features.

7.5.1 Missing Data Augmentation (MDA)

To obtain a MCV procedure, we suggest [modifying the calibration set](#) according to the [mask of the test point](#), while the training step is unchanged. More precisely, the mask of the test point is applied to the calibration set, as illustrated in Figure 7.2. The rationale is to mimic the missing pattern of the test point by artificially augmenting the calibration set with that mask. It ensures that the correction term is computed using data with (at least) the same missing values as the test point. We refer to this strategy as *CP with Missing Data Augmentation* (CP-MDA), and derive two versions of it. Algorithms 12 and 13 are written using CQR as the base conformal procedure, but they work with any conformal method as we describe in Section 7.E.1.

Algorithm 12 – CP-MDA-Exact. CP-MDA with *exact masking* consists of keeping the *artificially masked calibration points* (l. 7) that have exactly the same missing pattern as the *test point* (l. 5). Then Algorithm 12 performs as impute-then-predict+conformalization: impute the calibration set (l. 10), predict on it and get the calibration scores (l. 11), compute their quantile to obtain the correction term (l. 14), and finally impute and predict the test point with the fixed fitted model by adding and subtracting the correction term (l. 15) to

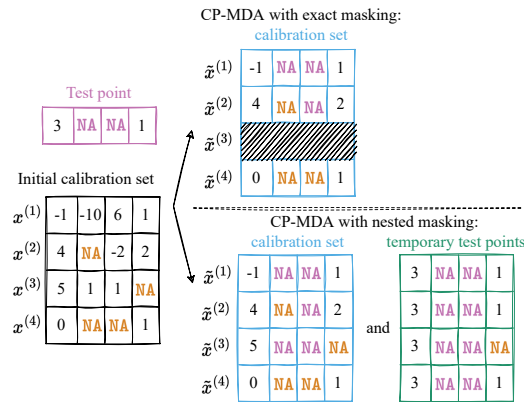


Figure 7.2: CP-MDA illustration. [Augmented calibration set](#) according to one [test point](#). For CP-MDA-Nested, the [augmented masks of the calibration set](#) are also [applied temporarily to the test point](#).

⁴say, for example young children whose input/output distribution is *not* dependent on the neighborhood.

Algorithm 12 CP-MDA-Exact (with CQR)

Input: Imputation algorithm \mathcal{I} , quantile regression algorithm \mathcal{QR} , significance level α , training set $\{(x^{(k)}, m^{(k)}, y^{(k)})\}_{k=1}^n$, test point $(x^{(\text{test})}, m^{(\text{test})})$

Output: Prediction interval $\hat{C}_\alpha(x^{(\text{test})}, m^{(\text{test})})$

- 1: Randomly split $\{1, \dots, n\}$ into 2 disjoint sets Tr & Cal
- 2: Fit the imputation function: $\Phi(\cdot) \leftarrow \mathcal{I}(\{(x^{(k)}, m^{(k)})\}, k \in \text{Tr})$
- 3: Impute the training set: $\forall k \in \text{Tr}, x_{\text{imp}}^{(k)} = \Phi(x^{(k)}, m^{(k)})$
- 4: Fit \mathcal{QR} :

$$\begin{aligned}\hat{q}_{\frac{\alpha}{2}}(\cdot) &\leftarrow \mathcal{QR}\left(\left\{\left(x_{\text{imp}}^{(k)}, y^{(k)}\right), k \in \text{Tr}\right\}, \alpha/2\right) \\ \hat{q}_{1-\frac{\alpha}{2}}(\cdot) &\leftarrow \mathcal{QR}\left(\left\{\left(x_{\text{imp}}^{(k)}, y^{(k)}\right), k \in \text{Tr}\right\}, 1-\alpha/2\right)\end{aligned}$$

// Generate an augmented calibration set:

- 5: $\text{Cal}^{(\text{test})} = \{k \in \text{Cal} \text{ such that } m^{(k)} \subset m^{(\text{test})}\}$
- 6: **for** $k \in \text{Cal}^{(\text{test})}$ **do**
- 7: $\tilde{m}^{(k)} = m^{(\text{test})}$ // Additional masking
- 8: **end for** Augmented calibration set generated. //
- 9: **for** $k \in \text{Cal}^{(\text{test})}$ **do**
- 10: Impute the calibration set: $x_{\text{imp}}^{(k)} = \Phi(x^{(k)}, \tilde{m}^{(k)})$
- 11: Set $s^{(k)} = \max(\hat{q}_{\frac{\alpha}{2}}(x_{\text{imp}}^{(k)}) - y^{(k)}, y^{(k)} - \hat{q}_{1-\frac{\alpha}{2}}(x_{\text{imp}}^{(k)}))$
- 12: **end for**
- 13: Set $S = \{s^{(k)}, k \in \text{Cal}^{(\text{test})}\}$
- 14: Compute $\hat{Q}_{1-\tilde{\alpha}}(S)$, the $1-\tilde{\alpha}$ -th empirical quantile of S , with $1-\tilde{\alpha} := (1-\alpha)(1+1/\#S)$
- 15: Set $\hat{C}_\alpha(x^{(\text{test})}, m^{(\text{test})}) = \left[\hat{q}_{\frac{\alpha}{2}} \circ \Phi(x^{(\text{test})}, m^{(\text{test})}) - \hat{Q}_{1-\tilde{\alpha}}(S); \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(x^{(\text{test})}, m^{(\text{test})}) + \hat{Q}_{1-\tilde{\alpha}}(S) \right]$

the initial conditional quantile estimates. Note that Algorithm 12 is described for one test point for simplicity but extends easily to many test points. The computations are then shared: the training part (l. 1-4) is common to any test point and the correction term (l. 5-14) can be reused for any new test point with the same mask.

In high dimensions, many calibration points may be discarded when applying CP-MDA-Exact since it is likely that their missing patterns would not be included in the one of the test point.⁵ This limitation brings us to the second algorithm we propose, CP-MDA-Nested.

Algorithm 13 – CP-MDA-Nested. CP-MDA with *nested masking* avoids the removal of calibration points whose masks are not included in that of the test point. Instead, we apply the mask of the test point to the calibration points, and so we keep all the observations (l. 3). Next, we impute the masked calibration points (l. 6) before computing their scores $s^{(k)}$ (l. 7). Then, for each calibration point, the fitted quantile regressors are used to predict on the test point with a temporary mask, which matches the mask of the given augmented calibration point. These predictions are corrected with the score of the calibration point (l. 8-9) and stored in two bags $Z_{\frac{\alpha}{2}}$ for the lower interval boundary, and $Z_{1-\frac{\alpha}{2}}$ for the upper interval boundary (l. 11-12). The prediction is finally obtained by

⁵Yet, these discarded points could be used for training but this comes at the cost of fitting a different model for each pattern; such a path is reasonable if the data is scarce.

Algorithm 13 CP-MDA-Nested (with CQR)**Input:** Same as Algorithm 12**Output:** Same as Algorithm 12

```

1: Compute lines 1 to 4 of Algorithm 12
   // Generate an augmented calibration set:
2: for  $k \in \text{Cal}$  do Additional nested masking
3:    $\tilde{m}^{(k)} = \max(m^{(\text{test})}, m^{(k)})$ 
4: end for Augmented calibration set generated. //
5: for  $k \in \text{Cal}$  do
6:   Impute the calibration set:  $x_{\text{imp}}^{(k)} := \Phi(x^{(k)}, \tilde{m}^{(k)})$ 
7:   Set  $s^{(k)} = \max(\hat{q}_{\frac{\alpha}{2}}(x_{\text{imp}}^{(k)}) - y^{(k)}, y^{(k)} - \hat{q}_{1-\frac{\alpha}{2}}(x_{\text{imp}}^{(k)}))$ 
8:   Set  $z_{\frac{\alpha}{2}}^{(k)} = \hat{q}_{\frac{\alpha}{2}} \circ \Phi(x^{(\text{test})}, \tilde{m}^{(k)}) - s^{(k)}$ 
9:   Set  $z_{1-\frac{\alpha}{2}}^{(k)} = \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(x^{(\text{test})}, \tilde{m}^{(k)}) + s^{(k)}$ 
10: end for
11: Set  $Z_{\frac{\alpha}{2}} = \{z_{\frac{\alpha}{2}}^{(k)}, k \in \text{Cal}\}$ 
12: Set  $Z_{1-\frac{\alpha}{2}} = \{z_{1-\frac{\alpha}{2}}^{(k)}, k \in \text{Cal}\}$ 
13: Compute  $\hat{Q}_{\tilde{\alpha}}(Z_{\frac{\alpha}{2}})$ 
14: Compute  $\hat{Q}_{1-\tilde{\alpha}}(Z_{1-\frac{\alpha}{2}})$ 
15: Set  $\hat{C}_{\alpha}(x^{(\text{test})}, m^{(\text{test})}) = [\hat{Q}_{\tilde{\alpha}}(Z_{\frac{\alpha}{2}}); \hat{Q}_{1-\tilde{\alpha}}(Z_{1-\frac{\alpha}{2}})]$ 

```

taking the α quantiles of the bags Z (l. 13-15).

The rationale for predicting on temporary test points with the mask of a given augmented calibration point is that we want to treat the test and calibration points in the same way.⁶ We should note that this method may tend to achieve conservative coverage, since the augmented calibration set may have masks that overly include the missing pattern of the test point, i.e., the augmented points may have more missing values than the test point.

7.5.2 Theoretical guarantees in finite sample

Let us consider the following assumptions.

Assumption A3 (Y is not explained by M). $(Y \perp\!\!\!\perp M)|X$.

Assumption A4 (Stochastic domination of the quantiles). Let $(\tilde{m}, \check{m}) \in \mathcal{M}^2$. If $\tilde{m} \subset \check{m}$ then for any $\delta \in [0, 0.5]$:

- $q_{1-\delta/2}^{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})} \leq q_{1-\delta/2}^{Y|(X_{\text{obs}(\check{m})}, M=\check{m})}$,
- $q_{\delta/2}^{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})} \geq q_{\delta/2}^{Y|(X_{\text{obs}(\check{m})}, M=\check{m})}$.

A4 grasps the underlying intuition that the conditional distribution of $Y|(X_{\text{obs}(m)}, M = m)$ tends to have larger deviations when the number of observed variables is smaller, in concordance with the intuition that observing predictive variables reduce the conditional randomness of $Y|X_{\text{obs}}$.

⁶This motivation is similar to the one of Jackknife+ (Barber et al., 2021b) and out-of-bags methods (Gupta et al., 2022).

The following theorems (proved in Section 7.E) state the finite sample guarantees of CP-MDA.

Theorem 7.5.1 (MCV of CP-MDA). *Assume the missing mechanism is MCAR, and A1 to A3. Then:*

1. *CP-MDA-Exact is MCV;*
2. *if the scores are almost surely distinct, CP-MDA-Exact is exactly MCV;*
3. *if A4 also holds, CP-MDA-Nested is MCV, up to a technical minor modification of the output.*

The challenge in proving MCV of CP-MDA-Nested is that the augmented calibration and test points are not exchangeable conditional on the mask and thus may result in under-coverage. However, by imposing A4 we prove that this violation of exchangeability still leads to MCV (and often conservative MCV) (see Lemma 7.E.1). We conjecture that CP-MDA-Nested attains MCV (without any modification), as also supported by experiments. However, we could not prove it without making an independence assumption which we prefer to avoid as exchangeability is key to imputation methods. Instead, we prove in Theorem 7.E.2 the MCV of any variant outputting $[\hat{Q}_{\hat{\alpha}}(Z_{\frac{\alpha}{2}}^{\tilde{m}}); \hat{Q}_{1-\hat{\alpha}}(Z_{1-\frac{\alpha}{2}}^{\tilde{m}})]$ for $Z_{\frac{\alpha}{2}}^{\tilde{m}}$ the subset of $Z_{\frac{\alpha}{2}}$ composed with points using mask \tilde{m} at l. 6-9.

Theorem 7.5.2 (Marginal validity of CP-MDA). *Under the same assumptions as Theorem 7.5.1 (i) CP-MDA-Exact is marginally valid; (ii) if A4 also holds, CP-MDA-Nested is marginally valid (with the same caveats as in Theorem 7.5.1).*

7.6 Towards asymptotic individualized coverage

Achieving validity conditionally on the mask is an important step towards conditional coverage: in practice one aims at the strongest coverage conditional on *both* X and M . Lei and Wasserman (2014); Vovk (2012); Barber et al. (2021a) studied a related question (without considering missing patterns) and concluded that it is impossible to achieve *informative* intervals satisfying conditional coverage, $\mathbb{P}(Y \in \hat{C}_{\alpha}(x)|X = x) \geq 1 - \alpha$ for any $x \in \mathcal{X}$ in the distribution-free and finite samples setting. Still, we can analyze the asymptotic regime, similarly to Theorem 1 of Sesia and Candès (2020), which proves the asymptotic conditional validity of CQR (without the presence of missing values) under consistency assumptions on the underlying quantile regressor. Here, by contrast, we study the asymptotic conditional validity of the impute-then-predict+conformalization procedure, by analyzing the consistency of impute-then-regress in Quantile Regression (QR). That is, we aim at showing that we satisfy the required assumption of consistency to invoke Theorem 1 of Sesia and Candès (2020). The proofs of this section are given in Section 7.F.

To analyze the consistency of impute-then-predict procedures for QR, we extend the work of Le Morvan et al. (2021) on mean regression. QR with missing values, for a quantile level β , aims at solving

$$\min_{f: \mathcal{X} \times \mathcal{M} \rightarrow \mathbb{R}} \mathcal{R}_{\ell_{\beta}}(f) := \mathbb{E}[\ell_{\beta}(Y, f(X, M))], \quad (7.3)$$

with ℓ_β the pinball loss $\ell_\beta(y, \hat{y}) = \rho_\beta(y - \hat{y})$ and $\rho_\beta(u) = \beta|u|\mathbb{1}_{\{u \geq 0\}} + (1 - \beta)|u|\mathbb{1}_{\{u \leq 0\}}$.

An associated ℓ_β -Bayes predictor minimizes Eq. (7.3). Its risk is called the ℓ_β -Bayes risk, noted $\mathcal{R}_{\ell_\beta}^*$. Impute-then-predict procedure in QR aims at solving

$$\min_{g: \mathcal{X} \rightarrow \mathbb{R}} \mathcal{R}_{\ell_\beta, \Phi}(g) := \mathbb{E}[\ell_\beta(Y, g \circ \Phi(X, M))], \quad (7.4)$$

for Φ any imputation. Let $g_{\ell_\beta, \Phi}^* \in \arg \min_g \mathcal{R}_{\ell_\beta, \Phi}(g)$. The following proposition states that $\mathcal{R}_{\ell_\beta, \Phi}(g_{\ell_\beta, \Phi}^*) = \mathcal{R}_{\ell_\beta}^*$ and the consistency of a universal learner.

Proposition 7.6.1 (ℓ_β -consistency of an universal learner). *Let $\beta \in [0, 1]$. If X admits a density on \mathbb{R}^d , then, for almost all imputation function $\Phi \in \mathcal{F}_\infty^I$, (i) $g_{\ell_\beta, \Phi}^* \circ \Phi$ is ℓ_β -Bayes-optimal (ii) any universally consistent algorithm for QR trained on the data imputed by Φ is ℓ_β -Bayes-consistent (i.e., asymptotically in the training set size).*

Note that this QR case does not require $\mathbb{E}[\varepsilon | X_{\text{obs}(M)}, M] = 0$, contrary to the quadratic loss case (Le Morvan et al., 2021). We conclude our asymptotic analysis of conditional coverage with Corollary 7.6.1.

Corollary 7.6.1. *For any missing mechanism, for almost all imputation function $\Phi \in \mathcal{F}_\infty^I$, if $F_{Y|(X_{\text{obs}(M)}, M)}$ is continuous, a universally consistent quantile regressor trained on the imputed data set yields asymptotic conditional coverage.*

In words, the intervals obtained by taking Bayes predictors of levels $\alpha/2$ and $1 - \alpha/2$ are exactly valid conditionally to both the mask M and the observed variables $X_{\text{obs}(M)}$, if $F_{Y|(X_{\text{obs}(M)}, M)}$ is continuous. Importantly, while this result is asymptotic, it holds for *any* missing mechanism and it considers individualized conditional coverage.

7.7 Empirical study

Setup. In all experiments, the data are imputed using iterative regression (**iterative ridge** implemented in Scikit-learn, Pedregosa et al. (2011)).⁷ We compare the performance of our CQR-MDA-Exact and CQR-MDA-Nested (that is CP-MDA based on CQR) to CQR as well as to a vanilla QR (without any calibration). The predictive models are fitted on the imputed data concatenated with the mask. Without concatenating the mask to the features, the mask-conditional coverage of QR is worsened, as demonstrated in Section 7.4. The prediction algorithm is a Neural Network (NN), fitted to minimize the pinball loss (Sesia and Romano, 2021, see Section 7.G.1 for details). For the vanilla QR, we use both the training and calibration sets for training.

Synthetic and semi-synthetic experiments. We designed the training and calibration data to have 20% of MCAR values. To evaluate the test marginal coverage $\mathbb{P}(Y \in \hat{C}_\alpha(X, M))$, missing values are introduced in the test set according to the same distribution as on the training and calibration sets. Then, to compute an estimator of $\mathbb{P}(Y \in \hat{C}_\alpha(X, m) | M = m)$ for each $m \in \mathcal{M}$, we fix to a constant the number of observations

⁷Theoretical results hold for any symmetric imputation. In practice, constant, mean and MICE imputations gave similar results.

per pattern, to ensure that the variability in coverage is not impacted by $\mathbb{P}(M = m)$. All experiments are repeated 100 times with different splits.

7.7.1 Synthetic experiments: Gaussian linear data

Data generation. The data is generated with $d = 10$ according to Model 7.4.1, with $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, \dots, 1)^T$ and $\Sigma = \varphi(1, \dots, 1)^T(1, \dots, 1) + (1 - \varphi)I_d$, $\varphi = 0.8$, Gaussian noise $\varepsilon \sim \mathcal{N}(0, 1)$ and the following regression coefficients⁸ $\beta = (1, 2, -1, 3, -0.5, -1, 0.3, 1.7, 0.4, -0.3)^T$. Here, the oracle intervals are known (Proposition 7.4.1).

Lowest and highest mask coverage, and associated length. Figures 7.1b and 7.8 (Section 7.G.2) and Figure 7.9 (Section 7.G.2) show the lowest and highest mask coverage and their associated length as a function of the training set size. The calibration size is fixed to 1000 and the test set contains 2000 points with the mask leading to the lowest coverage (here it corresponds to cases where only X_4 is observed) and 2000 points with the mask leading to the highest coverage (here it corresponds to all the variables observed). These figures highlight that:

- **CQR** and **QR** conditional coverage improve when the training size increases (Corollary 7.6.1);
- **Both versions of CQR-MDA** are MCV (Theorem 7.5.1);
- **CQR-MDA-Exact** is exactly MCV as highest and lowest mask coverage are exactly 90% (Theorem 7.5.1);
- **CQR-MDA-Exact**'s lengths converge to the oracle ones with increasing training size, showing it is not conservative, while **CQR-MDA-Nested** is overly conservative.

Coverage and length by mask size. Figure 7.3 displays the average coverage and intervals' length as a function of the pattern size, i.e., the performance metrics are aggregated by the masks with the same number of missing variables; the first violin plot of each panel corresponds to the marginal coverage (see Section 7.G.2 for QR results). Note that only the pattern sizes are presented and not the patterns themselves as there are $2^d = 1024$ possible masks.⁹ For each pattern size, 100 observations are drawn according to the distribution of $M|\text{size}(M)$ in the test set. The training and calibration sizes are respectively 500 and 250 (Figure 7.11 contains the results for other sizes). Figure 7.3 shows that:

- **CQR** is marginally valid (Proposition 7.3.1);
- **CQR** and **QR** undercover with an increasing number of missing values. This can be explained because their length nearly does not vary with the size of the missing pattern, despite having the mask concatenated with the features;
- **Both versions of CQR-MDA** are marginally valid (Th. 7.5.2) and mask(-size)-conditionally-valid (Th. 7.5.1);

⁸For dimension 3, in Figure 7.1a, the same model is used, keeping only the 3 first features and their associated parameters.

⁹Note that in practice the relationship between the coverage and the number of missing values is not necessarily monotonic as a mask with only one missing value can lead to more uncertainty than a mask with many missing values, see Section 7.D.

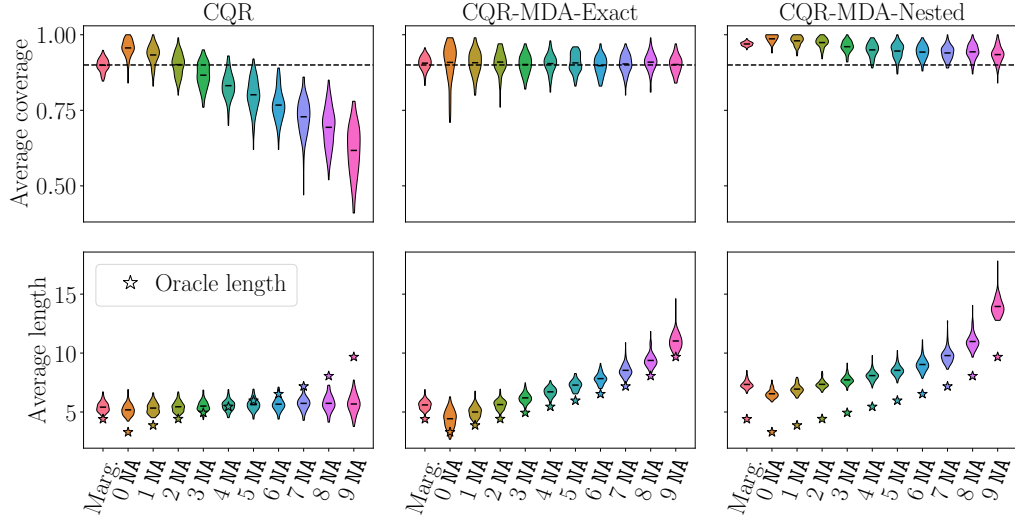


Figure 7.3: Average coverage (top) and length (bottom) as a function of the number of missing values (NA). The first violin plot shows the marginal coverage. $\#Tr = 500$ and $\#Cal = 250$. The marginal test set includes 2000 observations. The mask-conditional test set includes 100 individuals for each missing data pattern size.

- **CQR-MDA-Exact** is exactly mask(-size)-conditionally-valid (Theorem 7.5.1) and its length is close to the oracle ones. It has more variability for the patterns with few missing values as for these masks $Cal^{(test)}$ is smaller.

Similar experiments with 40% of missing values are available in Section 7.G.3. Briefly, it corresponds to a setting where CP-MDA-Nested is preferable over CP-MDA-Exact as the former outputs smaller intervals and is less variable.

7.7.2 Semi-synthetic experiments

We consider 6 benchmark real data sets for regression: `meps_19`, `meps_20`, `meps_21` (MEPS), `bio`, `bike` and `concrete` (Dua and Graff, 2017), where we introduce missing values in their quantitative features, each of them having a probability 0.2 of being missing (i.e. it is a MCAR mechanism), as in the synthetic experiments. Note that therefore some patterns have a low (or null) frequency of appearance in the training sets of `bio` and `concrete`. The sample sizes for training, calibration, and testing, and simulation details are provided in Section 7.G.4, along with results for smaller training and calibration sets.

Figure 7.4 depicts the results by combining *validity* and *efficiency* (length) for `meps_19`, `bio`, `concrete`, and `bike`, where this graph follows the visualization used in Zaffran et al. (2022). The results for `meps_20` and `meps_21` are given in Section 7.G.4, as they are similar to `meps_19`. Each of the panels in Figure 7.4 summarizes the results for one data set, with the average coverage shown in the x -axis and the average length in the y -axis. A method is mask-conditionally-valid if all the markers of its color are at the right of the vertical dotted line (90%). The design of Figure 7.4 requires a different interpretation than Figure 7.3 (or the subsequent Figure 7.5). For each method we report, for the pattern having the highest (or lowest) coverage, its length and coverage. However, as this pattern may depend on the method, the length for the highest/lowest should not be directly compared between methods. We observe that:

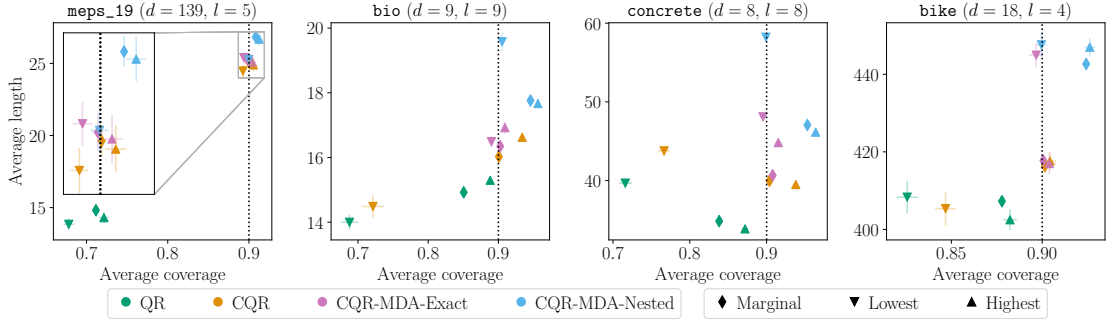


Figure 7.4: Validity and efficiency with missing values for 4 data sets (panels) with d features, including l quantitative ones in which missing values are introduced with probability 0.2. Colors represent the methods. Diamonds (\diamond) represent marginal coverage while the patterns giving the lowest and highest mask coverage are represented with triangles (∇ and \blacktriangle). Vertical dotted lines represent the target coverage.

- **CQR** is marginally valid (orange \diamond , Proposition 7.3.1), but not MCV as the lowest mask coverage (orange ∇) is far below 90% (*bio*, *concrete*, and *bike* data sets);
- **CQR-MDA-Exact** is marginally valid (purple \diamond , Theorem 7.5.2). It is also exactly MCV, as the lowest (purple ∇) and highest (purple \blacktriangle) mask coverages are about 90% (Theorem 7.5.1);
- **CQR-MDA-Nested** is marginally valid (blue \diamond , Theorem 7.5.2). It is also MCV, as the lowest (blue ∇) mask coverage is larger than 90% (Theorem 7.5.1).

7.7.3 Predicting the level of platelets for trauma patients

We study the applicability and robustness of CPMDA on the critical care TraumaBase[®] data. We focus on predicting the level of platelets of severely injured patients upon arrival at the hospital. This level is directly related to the occurrence of hemorrhagic shock and is difficult to obtain in real-time: predicting it accurately could be crucial to anticipate the need for transfusion and blood resources. In addition, this prediction task appears to be challenging as Jiang et al. (2022) achieved an average relative prediction error ($\|\hat{y} - y\|^2 / \|y\|^2$) that is no lower than 0.23. This highlights the need for reliable uncertainty

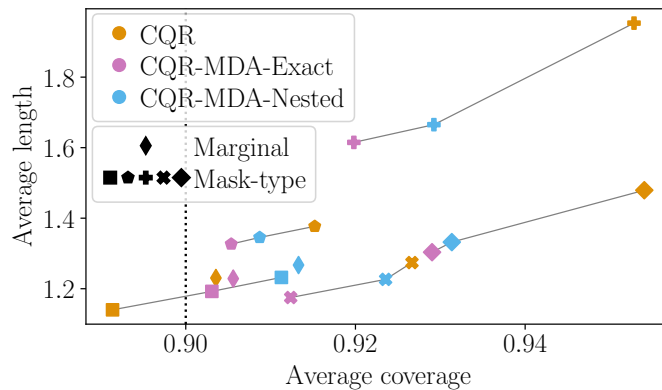


Figure 7.5: Average coverage and length on the TraumaBase[®] analysis. See the caption of Figure 7.4 for details. Other symbols than diamond correspond to computing the average per mask. Each individual's prediction is obtained by using 15390 observations for training, and 7694 for calibration.

quantification.

After applying inclusion and exclusion criteria obtained by medical doctors and following the pipeline of Sportisse et al. (2020) described in Section 7.G.5, we left with a subset of 28855 patients and 7 features. Missing values vary from 0% to 24% by features, with a total average of 7%.

Results. The results are summarized in Figure 7.5, where we use different markers to denote the different masks. To ensure a fair comparison between the conformal methods, we only keep the missing patterns for which there are more than 200 individuals; this excludes 7 patterns. Finally, since we found that the vanilla QR tends to be overly conservative, we refer to Section 7.G.5 for its results. Figure 7.5 shows that all conformal approaches achieve marginal coverage higher than the desired 90% level (diamonds \blacklozenge). Furthermore, for each mask (each set of linked markers) **CQR-MDA** improves coverage compared to **CQR** by approaching 90%, and efficiency by reducing the average length. Noticeably, for the pattern corresponding to all features observed (squares \blacksquare), **CQR-MDA** has a coverage rate above 90% while **CQR** is below the target level. Therefore, we believe **CQR-MDA** should be recommended as it improves upon the vanilla impute-then-regress+CQR approach.

7.8 Conclusion and perspectives

In this paper, we study the interplay between uncertainty quantification and missing values. We show that missing values introduce heteroskedasticity in the prediction task. This brings challenges on how to provide uncertainty estimators that are valid conditionally on the missing patterns, which are addressed by this work. Our analysis leaves several directions open: (1) obtaining results *beyond the MCAR assumption* for CP-MDA, both theoretically and numerically, (2) extending the (numerical) analysis to non-split approaches, (3) investigating the numerical performances of other conditional CP approaches (such as Sesia and Candès (2020); Izbicki et al. (2020, 2022); Lin et al. (2021)), (4) studying the impact of the imputation on QR with finite samples. A more detailed discussion on these directions is provided in Section 7.A.

Appendix to Conformal Prediction with Missing Values

The appendices are organized as follows.

Section 7.A provides a more detailed discussion on open directions and perspectives.

Section 7.B describes CQR, used in the paper.

Section 7.C provides an explicit description of impute-then-predict+conformalization (Section 7.C.1), along with its proof of validity, that is the proofs for Section 7.3 (Section 7.C.2).

Then, Section 7.D contains the proofs for the Gaussian linear model oracle intervals presented in Section 7.4 (Section 7.D.1), along with the discussion on how mean-based approaches fail (Section 7.D.2).

Section 7.E gives the general statement of CP-MDA-Exact (Section 7.E.1), and the proofs of the validity theorems for CP-MDA-Exact (Section 7.E.2), along with the theoretical study of CP-MDA-Nested (Section 7.E.3).

Section 7.F provides all the proofs about consistency and asymptotic conditional coverage presented in Section 7.6.

Finally, Section 7.G contains all the details for the experimental study and additional results completing Section 7.7. More precisely, Section 7.G.1 gives more details about the settings. Section 7.G.2 contains results on synthetic data with 20% of MCAR missing values, while Section 7.G.3 shows the results on synthetic data when the proportion of MCAR missing values is 40%. Section 7.G.4 describes the real data sets used for the semi-synthetic experiments, and presents the remaining results. Section 7.G.5 presents the real medical data set (TraumaBase®), the pipeline and settings used and the results obtained by QR on this data set.

7.A Detailed perspective discussion

First, obtaining results *beyond the MCAR assumption* for CP-MDA. On the numerical side, preliminary experiments show promising results, indicating CP-MDA’s robustness, but a detailed numerical study is needed. On the theoretical side, understanding the limits of CP-MDA validity is of high importance. Results without assumptions on the missingness distribution seem impossible to obtain. Even with MAR data, the task of pointwise prediction can be very challenging if the output distribution strongly depends on the pattern (Ayme et al., 2022). As the impossibility results of conditional validity (Lei and

Wasserman, 2014; Vovk, 2012; Barber et al., 2021a), assumptions on the missing mechanism are needed.

Second, extending the (numerical) analysis to non-split approaches (e.g., based on the Jackknife) would be relevant, as it could improve the base model and therefore how the heteroskedasticity is taken into account. Note that CP-MDA can be written to take into account this splitting strategy, and thus our theoretical results on MCV would directly extend.

Third, investigating the numerical performances of other conditional CP approaches (such as Sesia and Candès (2020); Izbicki et al. (2020, 2022); Lin et al. (2021)) within the MDA framework is of interest. In this paper, we analyze empirically the instance of CP-MDA on top of CQR as it is the simplest version of QR based CP, but the theory and motivation of this work is not specific to CQR. Exactly as CQR, none of the aforementioned methods would provide MCV if used out of the box. But if combined with CP-MDA, then all of them will be granted MCV.

Finally, while our approach is to be agnostic to the imputation chosen (similarly to CP being agnostic to the underlying model), an interesting research path is to study the impact of the imputation on QR with finite samples.

7.B Illustration and details on CQR (Romano et al., 2019) procedure

Figure 7.6 provides a visualization and step by step description of CQR.

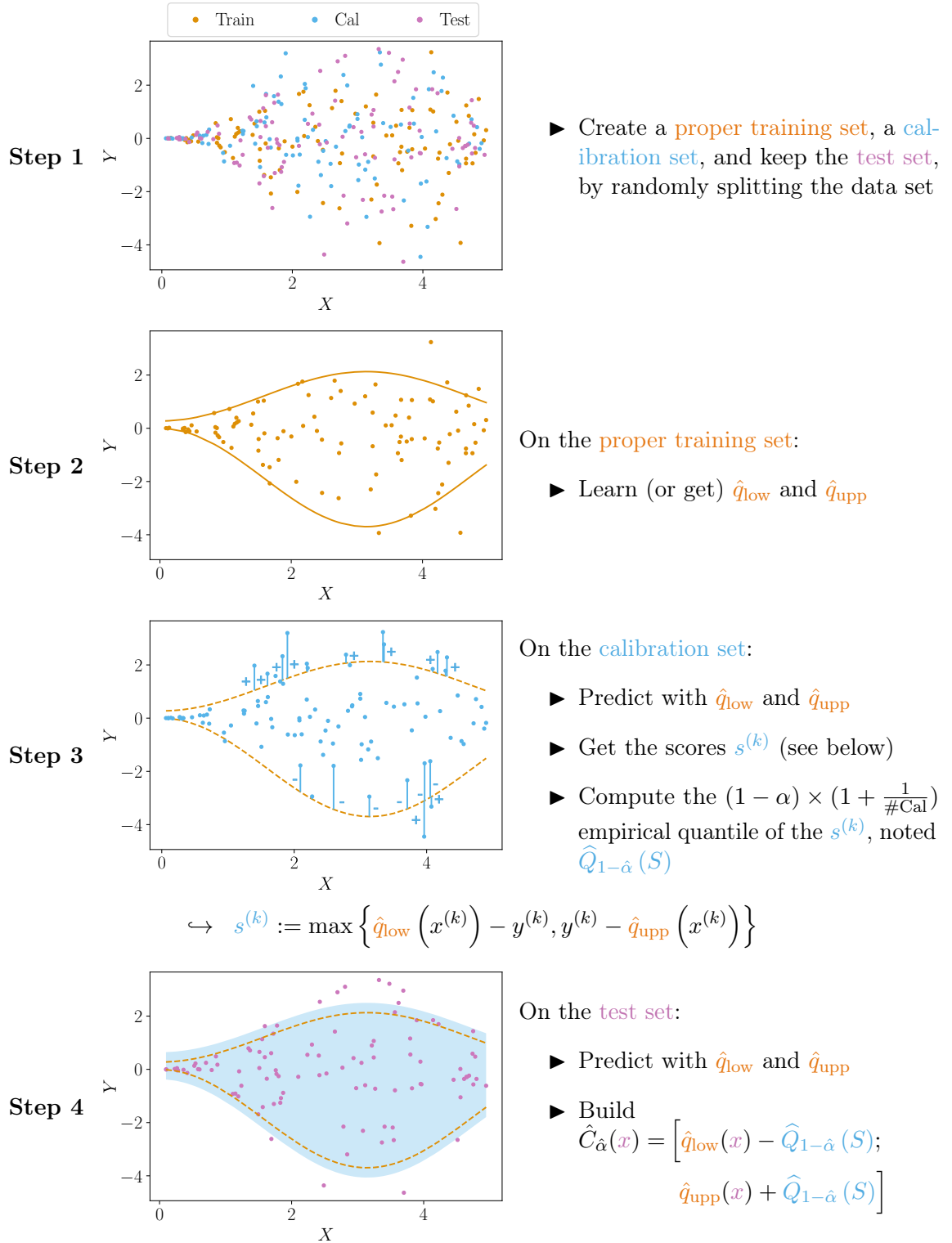


Figure 7.6: Schematic illustration of Conformalized Quantile Regression (CQR) (Romano et al., 2019).

7.C Impute-then-predict+conformalization

7.C.1 Description of the algorithm

Algorithm 14 SCP on impute-then-predict

Input: Imputation algorithm \mathcal{I} , predictive algorithm \mathcal{A} , conformity score function s , significance level α , training set $\{(X^{(1)}, M^{(1)}, Y^{(1)}), \dots, (X^{(n)}, M^{(n)}, Y^{(n)})\}$.

Output: Prediction interval $\hat{C}_\alpha(X, M)$.

- 1: Randomly split $\{1, \dots, n\}$ into two disjoint sets Tr and Cal.
 - 2: Fit the imputation function: $\Phi(\cdot) \leftarrow \mathcal{I}(\{(X^{(k)}, M^{(k)}), k \in \text{Tr}\})$
 - 3: Impute the data set: $\{X_{\text{imp}}^{(k)}\}_{k=1}^n := \{\Phi(X^{(k)}, M^{(k)})\}_{k=1}^n$
 - 4: Fit algorithm \mathcal{A} : $\hat{g}(\cdot) \leftarrow \mathcal{A}(\{(X_{\text{imp}}^{(k)}, Y^{(k)}), k \in \text{Tr}\})$
 - 5: **for** $k \in \text{Cal}$ **do**
 - 6: Set $S^{(k)} = s(Y^{(k)}, \hat{g}(X_{\text{imp}}^{(k)}))$, the conformity scores
 - 7: **end for**
 - 8: Set $\mathcal{S}_{\text{Cal}} = \{S^{(k)}, k \in \text{Cal}\}$
 - 9: Compute $\hat{Q}_{1-\alpha^{\text{SCP}}}(\mathcal{S}_{\text{Cal}})$, the $1 - \alpha^{\text{SCP}}$ -th empirical quantile of \mathcal{S}_{Cal} , with $1 - \alpha^{\text{SCP}} := (1 - \alpha)(1 + 1/\#\text{Cal})$.
 - 10: Set $\hat{C}_\alpha(X, M) = \{y \text{ such that } s(y, \hat{g} \circ \Phi(X, M)) \leq \hat{Q}_{1-\alpha^{\text{SCP}}}(\mathcal{S}_{\text{Cal}})\}$.
-

Similarly, Algorithm 12 can be written to include any underlying predictive algorithm (regression or classification) and any score function.

7.C.2 Proof of exchangeability after imputation

In this subsection, we provide a more formal statement of Lemma 7.3.1 and Proposition 7.3.1 in respectively Lemma 7.C.1 and Proposition 7.C.1. To that end, we introduce a notion of symmetrical imputation on a set \mathcal{T} , for $\mathcal{T} \subset \llbracket 1, n+1 \rrbracket$.

Assumption A5 (Symmetrical imputation on a set \mathcal{T}).

For a given set of points $\{X^{(k)}, M^{(k)}, Y^{(k)}\}_{k \in \mathcal{T}}$ the imputation function Φ is the output of an algorithm \mathcal{I} that treats the data points in \mathcal{T} symmetrically: $\mathcal{I}(\{X^{(k)}, M^{(k)}, Y^{(k)}\}_{k \in \mathcal{T}}) \stackrel{(d)}{=} \mathcal{I}(\{X^{(\sigma(k))}, M^{(\sigma(k))}, Y^{(\sigma(k))}\}_{k \in \mathcal{T}})$ conditionally to $\{X^{(k)}, M^{(k)}, Y^{(k)}\}_{k \in \mathcal{T}}$ and for any permutation σ on $\llbracket 1, \#\mathcal{T} \rrbracket$.

Lemma 7.C.1 (Imputation preserves exchangeability). *Let A1 hold. Then, for any missing mechanism, for any imputation function Φ satisfying A5, the imputed random variables $(\Phi(X^{(k)}, M^{(k)}), M^{(k)}, Y^{(k)})_{k \in \mathcal{T}}$ are exchangeable.*

Proposition 7.C.1 ((Exact) validity of impute-then-predict+conformalization). *If A1 is satisfied, then we have the following three results.*

1. **Full CP:** if A5 is satisfied for $\mathcal{T} = \llbracket 1, n+1 \rrbracket$ (i.e., the imputation algorithm treats all points symmetrically), then impute-then-predict+Full CP is marginally valid. If moreover the scores are almost surely distinct, it is exactly valid.

OR

2. **Jackknife+** if [A5](#) is satisfied for $\mathcal{T} = \llbracket 1, n+1 \rrbracket$ (i.e., the imputation algorithm treats all points symmetrically), then impute-then-predict+Jackknife+ is marginally valid (of level $1 - 2\alpha$).

OR

3. **SCP** with the split $\llbracket 1, n+1 \rrbracket = \text{Tr} \cup \text{Cal} \cup \text{Test}$ and if [A5](#) is satisfied for $\mathcal{T} = \text{Cal} \cup \text{Test}$ (i.e., the imputation treats all points in $\text{Cal} \cup \text{Test}$ symmetrically) then impute-then-predict+conformalization is marginally valid. If moreover the scores are almost surely distinct, it is exactly valid.

Remark 7.C.1 (Imputation choices for SCP). In the latter case, for SCP, the coverage result can be derived conditionally on Tr , thus the coverage results holds for: (i) any deterministic imputation function (conditionally on Tr) (that is any arbitrary function of Tr), or (ii) any stochastic imputation function treating Cal and Test symmetrically (iii) any combination of both.

Proof of Lemma 7.C.1.

Φ is the output of an imputing algorithm \mathcal{I} trained on $\left\{ (X^{(k)}, M^{(k)}, Y^{(k)})_{k \in \mathcal{T}} \right\}$.

Assume $(X^{(k)}, M^{(k)}, Y^{(k)})_{k \in \mathcal{T}}$ are exchangeable ([A1](#)).

Thus, if \mathcal{I} treats the data points in \mathcal{T} symmetrically, $(\Phi(X^{(k)}, M^{(k)}), M^{(k)}, Y^{(k)})_{k \in \mathcal{T}}$ are exchangeable (see proof of Theorem 1b in ([Barber et al., 2023](#)) for example).

□

Proof of Proposition 7.C.1. Proposition 7.C.1 is a consequence of Lemma 7.C.1 with different choices of \mathcal{T} , that enable to apply the following results:

1. Full CP: [Vovk et al. \(2005\)](#), also re-stated in [Barber et al. \(2023\)](#)
2. Jackknife+: [Barber et al. \(2021b\)](#)
3. SCP: [Lei et al. \(2018\)](#) or [Papadopoulos et al. \(2002\)](#) and [Angelopoulos and Bates \(2023\)](#) for a generic version with any score function (note that the coverage is proved conditionally on Tr).

□

7.D Gaussian linear model

7.D.1 Distribution of $Y|(X_{\text{obs}(m)}, M)$ and oracle intervals

Proposition 7.D.1 (Distribution of $Y|(X_{\text{obs}(M)}, M)$ ([Le Morvan et al., 2020b](#))). *Under Model 7.4.1, for any $m \in \{0, 1\}^d$:*

$$Y|(X_{\text{obs}(m)}, M = m) \sim \mathcal{N}(\tilde{\mu}^m, \tilde{\Sigma}^m),$$

with:

$$\begin{aligned}
\tilde{\mu}^m &= \beta_{\text{obs}(m)}^T X_{\text{obs}(m)} + \beta_{\text{mis}(m)}^T \mu_{\text{mis|obs}}^m \\
\mu_{\text{mis|obs}}^m &= \mu_{\text{mis}(m)}^m + \Sigma_{\text{mis}(m), \text{obs}(m)}^m (\Sigma_{\text{obs}(m), \text{obs}(m)}^m)^{-1} (X_{\text{obs}(m)} - \mu_{\text{obs}(m)}^m), \\
\tilde{\Sigma}^m &= \beta_{\text{mis}(m)}^T \Sigma_{\text{mis|obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2 \\
\Sigma_{\text{mis|obs}}^m &= \Sigma_{\text{mis}(m), \text{mis}(m)}^m - \Sigma_{\text{mis}(m), \text{obs}(m)}^m (\Sigma_{\text{obs}(m), \text{obs}(m)}^m)^{-1} \Sigma_{\text{obs}(m), \text{mis}(m)}^m.
\end{aligned}$$

Proposition 7.D.2 (Oracle intervals). *Under Model 7.4.1, for any $m \in \{0, 1\}^d$, for any $\delta \in (0, 1)$:*

$$q_\delta^{Y|(X_{\text{obs}(m)}, M=m)} = \beta_{\text{obs}(m)}^T X_{\text{obs}(m)} + \beta_{\text{mis}(m)}^T \mu_{\text{mis|obs}}^m + q_\delta^{\mathcal{N}(0,1)} \sqrt{\beta_{\text{mis}(m)}^T \Sigma_{\text{mis|obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2},$$

and the oracle predictive interval length is given by:

$$\mathcal{L}_\alpha^*(m) = 2q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sqrt{\beta_{\text{mis}(m)}^T \Sigma_{\text{mis|obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2}. \quad (7.5)$$

Proof. Using multivariate Gaussian conditioning (Eaton, 1983), for any subset of indices $L \in \llbracket 1, d \rrbracket$:

$$X_K | (X_L, M) \sim \mathcal{N}(\mu_{K|L}^M, \Sigma_{K|L}^M), \quad (7.6)$$

with $K = \bar{L}$ (the complement indices) and:

$$\begin{aligned}
\mu_{K|L}^M &= \mu_K^M + \Sigma_{K,L}^M \Sigma_{L,L}^{M^{-1}} (X_L - \mu_L^M), \\
\Sigma_{K|L}^M &= \Sigma_{K,K}^M - \Sigma_{K,L}^M \Sigma_{L,L}^{M^{-1}} \Sigma_{L,K}^M.
\end{aligned}$$

Given that $Y = \beta^T X + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \perp (X, M)$, the following holds:

$$Y | (X_L, M) \stackrel{(d)}{=} (\beta^T X + \varepsilon) | (X_L, M) \stackrel{(d)}{=} \beta_L^T X_L + (\varepsilon + \beta_K^T X_K) | (X_L, M)$$

and by Equation (7.6), $\beta_K^T X_K | (X_L, M) \sim \mathcal{N}(\beta_K^T \mu_{K|L}^M, \beta_K^T \Sigma_{K|L}^M \beta_K)$, and $(\varepsilon | (X_L, M)) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, and $(\beta_K^T X_K \perp \varepsilon) | (X_L, M)$. Thus:

$$Y | (X_L, M) \sim \mathcal{N}(\beta_L^T X_L + \beta_K^T \mu_{K|L}^M, \beta_K^T \Sigma_{K|L}^M \beta_K + \sigma_\varepsilon^2).$$

Consequently, for any $\delta \in (0, 1)$:

$$q_\delta^{Y|(X_L, M)} = \beta_L^T X_L + \beta_K^T \mu_{K|L}^M + q_\delta^{\mathcal{N}(0,1)} \sqrt{\beta_K^T \Sigma_{K|L}^M \beta_K + \sigma_\varepsilon^2}. \quad (7.7)$$

For any pattern $m \in \{0, 1\}^d$, applying Equation (7.7) with $K = \text{mis}(m) = \overline{\text{obs}(m)}$, $L = \text{obs}(m)$, we have, for any $\delta \in (0, 1)$:

$$q_\delta^{Y|(X_{\text{obs}(m)}, M=m)} = \beta_{\text{obs}(m)}^T X_{\text{obs}(m)} + \beta_{\text{mis}(m)}^T \mu_{\text{mis|obs}}^m + q_\delta^{\mathcal{N}(0,1)} \sqrt{\beta_{\text{mis}(m)}^T \Sigma_{\text{mis|obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2},$$

and:

$$\mathcal{L}_\alpha^*(m) = 2 \times q_{1-\alpha/2}^{\mathcal{N}(0,1)} \times \sqrt{\beta_{\text{mis}(m)}^T \Sigma_{\text{mis|obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2},$$

with:

$$\begin{aligned}
\mu_{\text{mis|obs}}^m &= \mu_{\text{mis}(m)}^m + \Sigma_{\text{mis}(m), \text{obs}(m)}^m (\Sigma_{\text{obs}(m), \text{obs}(m)}^m)^{-1} (X_{\text{obs}(m)} - \mu_{\text{obs}(m)}^m), \\
\Sigma_{\text{mis|obs}}^m &= \Sigma_{\text{mis}(m), \text{mis}(m)}^m - \Sigma_{\text{mis}(m), \text{obs}(m)}^m (\Sigma_{\text{obs}(m), \text{obs}(m)}^m)^{-1} \Sigma_{\text{obs}(m), \text{mis}(m)}^m.
\end{aligned}$$

□

7.D.2 Discussion on how mean-based approaches fail

Under Model 7.4.1, the Bayes predictor for a quadratic loss in presence of missing values – $\mathbb{E}[Y | (X_{\text{obs}(M)}, M)]$ – is fully characterized (Le Morvan et al., 2020b,a; Ayme et al., 2022).

Figure 7.7 is obtained by generating the data according to Model 7.4.1 with $d = 3$, $\beta = (1, 2, -1)^T$ and $\sigma_\varepsilon = 1$, with multivariate Gaussian X and MCAR mechanism ($X \perp\!\!\!\perp M$) (which is a particular case of Model 7.4.1 with $\mu^m \equiv \mu$ and $\Sigma^m \equiv \Sigma$). The left panel represents the method *Oracle mean + SCP* where SCP is applied on the regressor being the Bayes predictor for the mean with absolute residuals as the score function. The first violin plot represents the marginal coverage whereas the other 7 represent conditional coverage with respect to the different possible patterns: conditional on observing all the variables, on observing all the variables except X_1 , except X_2 etc (see Section 7.7 for details on the simulation process).

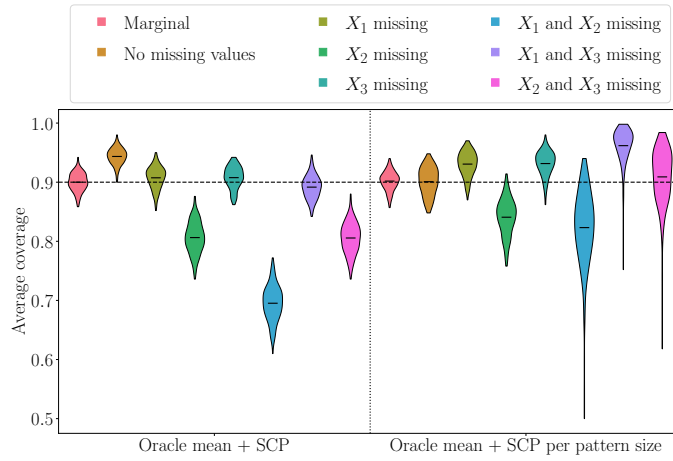


Figure 7.7: Calibration set contains 500 points. Test size for each pattern is of 500 individuals and for marginal is of 2000. 200 repetitions allow to display violin plots, the horizontal black line representing the mean.

SCP on a (oracle) mean regressor lacks of conditional coverage with respect to the mask. Figure 7.7 (left) highlights that even with the best mean regressor (the Bayes predictor) and an homoskedastic noise, usual SCP intervals:

- over-cover when there are no missing values;
- cover less for a mask \check{m} than for a mask \hat{m} when $\hat{m} \subset \check{m}$ (e.g. $\hat{m} = (1, 0, 0)$ only X_1 is missing, $\check{m} = (1, 1, 0)$ that is X_1 and X_2 are missing);
- cover less when the most informative variable (X_2) is missing.

To tackle this issue, one could calibrate conditionally to the missing data patterns. This is in the same vein as calibrating conditionally to the categories of a categorical variable or to different groups (Romano et al., 2020a). This strategy is not viable as there are 2^d patterns: the number of subsets grows exponentially with the dimension, implying the creation of subsets with too little data to perform the calibration. As an alternative, one could consider to perform calibration conditionally to the pattern size (e.g. when $d = 3$, either 0 missing value, 1 or 2). This is possible as there are only d different pattern sizes.

Calibrating by pattern size does not provide validity conditionally to the missing data patterns. Figure 7.7 (right) shows the coverages of *Oracle mean + SCP per pattern size* where SCP is applied on the Bayes predictor for the mean and the calibration is protected by pattern size. The previous statements still hold with this strategy, even if the coverage disparities are smaller. Therefore, it is not enough to calibrate per pattern size.

7.E Finite sample algorithms

7.E.1 General statement of Algorithm 12

We provide in Algorithm 15 a general statement of CP-MDA-Exact handling any learning algorithm (both regression and classification) and any score function.

Algorithm 15 CP-MDA-Exact

Input: Imputation algorithm \mathcal{I} , predictive algorithm \mathcal{A} , conformity score function s_g parametrized by a model g , significance level α , training set $\{(X^{(k)}, M^{(k)}, Y^{(k)})\}_{k=1}^n$, test point $(X^{(\text{test})}, M^{(\text{test})})$.

Output: Prediction interval $\hat{C}_\alpha(x^{(\text{test})}, m^{(\text{test})})$.

- 1: Randomly split $\{1, \dots, n\}$ into two disjoint sets Tr and Cal.
 - 2: Fit the imputation function: $\Phi(\cdot) \leftarrow \mathcal{I}(\{(X^{(k)}, M^{(k)})\}_{k \in \text{Tr}})$
 - 3: Impute the training set: $\{X_{\text{imp}}^{(k)}\}_{k \in \text{Tr}} := \{\Phi(X^{(k)}, M^{(k)})\}_{k \in \text{Tr}}$
 - 4: Fit algorithm \mathcal{A} : $\hat{g}(\cdot) \leftarrow \mathcal{A}(\{(X_{\text{imp}}^{(k)}, Y^{(k)})\}_{k \in \text{Tr}})$
 - // Generate an augmented calibration set:
 - 5: $\text{Cal}^{(\text{test})} = \{k \in \text{Cal} \text{ such that } M^{(k)} \subset M^{(\text{test})}\}$
 - 6: **for** $k \in \text{Cal}^{(\text{test})}$ **do**
 - 7: $\widetilde{M}^{(k)} = M^{(\text{test})}$ Additional masking
 - 8: **end for**
 - Augmented calibration set generated. //
 - 9: Impute the calibration set: $\{X_{\text{imp}}^{(k)}\}_{k \in \text{Cal}^{(\text{test})}} := \{\Phi(X^{(k)}, \widetilde{M}^{(k)})\}_{k \in \text{Cal}^{(\text{test})}}$
 - 10: **for** $k \in \text{Cal}^{(\text{test})}$ **do**
 - 11: Set $S^{(k)} = s_{\hat{g}}(Y^{(k)}, X_{\text{imp}}^{(k)})$, the conformity scores
 - 12: **end for**
 - 13: Set $\mathcal{S}_{\text{Cal}} = \{S^{(k)}, k \in \text{Cal}^{(\text{test})}\}$
 - 14: Compute $\hat{Q}_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}})$, the $1 - \tilde{\alpha}$ -th empirical quantile of \mathcal{S}_{Cal} , with $1 - \tilde{\alpha} := (1 - \alpha)(1 + 1/\#\mathcal{S}_{\text{Cal}})$.
 - 15: Set $\hat{C}_\alpha(X^{(\text{test})}, M^{(\text{test})}) = \{y \text{ such that } s_{\hat{g}}(y, \Phi(X^{(\text{test})}, M^{(\text{test})})) \leq \hat{Q}_{1-\tilde{\alpha}}(\mathcal{S}_{\text{Cal}})\}$.
-

7.E.2 Mask-conditional validity of CP-MDA-Exact

Before proving the results, we introduce a slightly stronger notion of mask-conditional validity, when the calibration set is itself of random cardinality.

Definition 7.E.1 (Mask-conditional-validity-random-calibration-size). A method is mask-conditionally-valid with a random calibration size $\#\text{Cal}$ if for any $m \in \mathcal{M}$, the lower bound is satisfied, and exactly mask-conditionally-valid if for any $m \in \mathcal{M}$, $1 \leq c \leq n$, the upper

bound is also satisfied:

$$1 - \alpha \leq \underset{\text{valid}}{\mathbb{P}} \left(Y^{(n+1)} \in \widehat{C}_\alpha \left(X^{(n+1)}, m \right) \mid M^{(n+1)} = m, \# \text{Cal} = c \right) \underset{\text{exactly valid}}{\leq} 1 - \alpha + \frac{1}{c+1}.$$

We start by proving Theorem 7.E.1 that implies the result on CP-MDA-Exact in Theorem 7.5.1.

Theorem 7.E.1. *[Conditional validity of CP-MDA-Exact with calibration of random cardinality] Assume the missing mechanism is MCAR, and that Assumptions A1 to A3 hold. Then:*

- CP-MDA-Exact is valid with a random calibration size $\# \text{Cal}$ conditionally to the missing patterns;
- if the scores $S^{(k)}$ are almost surely distinct, CP-MDA-Exact is exactly mask-conditionally-valid with a random calibration size $\# \text{Cal}$.

Proof of Theorem 7.E.1. Let Tr and Cal be two disjoint sets on $\llbracket 1, n \rrbracket$. Let \hat{g} be some model. Given A1, the sequence $\left\{ (X^{(k)}, M^{(k)}, Y^{(k)})_{k \in \text{Cal}}, (X^{(\text{test})}, M^{(\text{test})}, Y^{(\text{test})}) \right\}$ is exchangeable. Therefore, the sequence $\left\{ (X^{(k)}, Y^{(k)})_{k \in \text{Cal}}, (X^{(\text{test})}, Y^{(\text{test})}) \right\}$ is also exchangeable.

Let m in \mathcal{M} . We define $\text{Cal}^m = \{k \in \text{Cal} \text{ such that } M^{(k)} \subset m\}$.

Let $c \in \llbracket 1, \# \text{Cal} \rrbracket$.

As the $M \perp\!\!\!\perp X$ (missingness is MCAR) and $(M \perp\!\!\!\perp Y) \mid X$ (Assumption A3), then $M \perp\!\!\!\perp (X, Y)$, and $\# \text{Cal}^m \perp\!\!\!\perp (X^{(k)}, Y^{(k)})_{k \in \text{Cal}}, (X^{(\text{test})}, Y^{(\text{test})})$. It follows that the sequence $\left\{ (X^{(k)}, Y^{(k)})_{k \in \text{Cal}^m}, (X^{(\text{test})}, Y^{(\text{test})}) \right\}$ is exchangeable conditionally to $\# \text{Cal}^m = c$.

Similarly, $M^{(\text{test})} \perp\!\!\!\perp (X^{(k)}, Y^{(k)})_{k \in \text{Cal}}, (X^{(\text{test})}, Y^{(\text{test})})$.

Thus the sequence $\left\{ (X^{(k)}, M^{(\text{test})}, Y^{(k)})_{k \in \text{Cal}^m}, (X^{(\text{test})}, M^{(\text{test})}, Y^{(\text{test})}) \right\}$ is exchangeable conditionally to $\# \text{Cal}^m = c$ and $M^{(\text{test})} = m$.

Therefore, we can now invoke Proposition 7.3.1 in combination with Lemma 1 of Romano et al. (2020a) to conclude the proof. But we can state a more rigorous version here, since in fact Cal^m is a random variable (as discussed in Definition 7.E.1).

Since the algorithm \mathcal{I} treats the calibration and test data points symmetrically (A5 with $\mathcal{T} = \text{Cal} \cup \text{Test}$), A5 also holds for any $\mathcal{T}' \subset \mathcal{T}$. Therefore, by Lemma 7.C.1 the sequence $\left\{ (\Phi(X^{(k)}, M^{(\text{test})}), M^{(\text{test})}, Y^{(k)})_{k \in \text{Cal}^m}, (\Phi(X^{(\text{test})}, M^{(\text{test})}), M^{(\text{test})}, Y^{(\text{test})}) \right\}$ is exchangeable conditionally to $\# \text{Cal}^m = c$ and $M^{(\text{test})} = m$.

The conclusion follows from usual arguments (Papadopoulos et al., 2002; Lei et al., 2018; Angelopoulos and Bates, 2023).

Precisely, $\left\{ (s_{\hat{g}}(Y^{(k)}, \Phi(X^{(k)}, M^{(\text{test})})))_{k \in \text{Cal}^m}, s_{\hat{g}}(Y^{(\text{test})}, \Phi(X^{(\text{test})}, M^{(\text{test})})) \right\}$ is exchangeable conditionally to $\# \text{Cal}^m = c$ and $M^{(\text{test})} = m$. Therefore,

$$\begin{aligned} & \mathbb{P} \left(s_{\hat{g}}(Y^{(\text{test})}, \Phi(X^{(\text{test})}, M^{(\text{test})})) \right. \\ & \quad \left. \leq \widehat{Q}_{1-\tilde{\alpha}}((s_{\hat{g}}(Y^{(k)}, \Phi(X^{(k)}, M^{(\text{test})})))_{k \in \text{Cal}^m}) \mid M^{(\text{test})} = m, \# \text{Cal}^m = c \right) \geq 1 - \alpha, \end{aligned}$$

and if the $\left((s_{\hat{g}}(Y^{(k)}, \Phi(X^{(k)}, M^{(\text{test})})))_{k \in \text{Cal}^m}, s_{\hat{g}}(Y^{(\text{test})}, \Phi(X^{(\text{test})}, M^{(\text{test})}))\right)$ are almost surely distinct (i.e. have a continuous distribution) then (Lei et al., 2018; Romano et al., 2019):

$$\begin{aligned} & \mathbb{P}\left(s_{\hat{g}}(Y^{(\text{test})}, \Phi(X^{(\text{test})}, M^{(\text{test})}))\right. \\ & \quad \left. \leq \widehat{Q}_{1-\tilde{\alpha}}((s_{\hat{g}}(Y^{(k)}, \Phi(X^{(k)}, M^{(\text{test})})))_{k \in \text{Cal}^m}) \middle| M^{(\text{test})} = m, \# \text{Cal}^m = c\right) \leq 1 - \alpha + \frac{1}{c+1}. \end{aligned}$$

This proves the first two points (with respect to Definition 7.E.1) of Theorem 7.5.1, by observing that

$$\begin{aligned} \left\{Y^{(\text{test})} \in \widehat{C}_{\alpha}(X^{(\text{test})}, M^{(\text{test})})\right\} &= \left\{s_{\hat{g}}(Y^{(\text{test})}, \Phi(X^{(\text{test})}, M^{(\text{test})}))\right. \\ & \quad \left. \leq \widehat{Q}_{1-\tilde{\alpha}}\left(\left(s_{\hat{g}}(Y^{(k)}, \Phi(X^{(k)}, M^{(\text{test})})))_{k \in \text{Cal}^m}\right)\right\}. \end{aligned}$$

□

Then, the proof of Theorem 7.5.2 (marginal validity of the CP-MDA-Exact) is direct by marginalizing the result of Theorem 7.5.1. □

7.E.3 Validities of CP-MDA-Nested.

Next, we give more details on the results on CP-MDA-Nested.

7.E.3.1 MASK-CONDITIONAL-VALIDITY OF CP-MDA-NESTED.

Let $m \in \mathcal{M}$.

We start by describing the links between CP-MDA-Nested and CP-MDA-Exact. CP-MDA-Exact can be re-written in the same way as CP-MDA-Nested, but keeping the subselection step of l. 5.

Indeed, first mention that the output of Algorithm 12 can be written in the following ways:

$$\begin{aligned} \bullet \quad \widehat{C}_{\alpha}(X^{(\text{test})}, m^{(\text{test})}) &= \left[\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, m^{(\text{test})}) - \widehat{Q}_{1-\tilde{\alpha}}(S); \right. \\ & \quad \left. \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, m^{(\text{test})}) + \widehat{Q}_{1-\tilde{\alpha}}(S)\right] \\ \bullet \quad \widehat{C}_{\alpha}(X^{(\text{test})}, m^{(\text{test})}) &= \left[\widehat{Q}_{\tilde{\alpha}}\left(\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, m^{(\text{test})}) - S_{\text{Cal}^{(\text{test})}}\right); \right. \\ & \quad \left. \widehat{Q}_{1-\tilde{\alpha}}\left(\hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, m^{(\text{test})}) + S_{\text{Cal}^{(\text{test})}}\right)\right] \\ \bullet \quad \widehat{C}_{\alpha}(X^{(\text{test})}, m^{(\text{test})}) &= \left[\widehat{Q}_{\tilde{\alpha}}\left(Z_{\frac{\alpha}{2}}^{m^{(\text{test})}}\right); \widehat{Q}_{1-\tilde{\alpha}}\left(Z_{1-\frac{\alpha}{2}}^{m^{(\text{test})}}\right)\right]. \end{aligned}$$

With $Z_{\frac{\alpha}{2}}^m := \{z_{\frac{\alpha}{2}}^{(k)}, k \in \text{Cal} \text{ and } \tilde{M}^{(k)} = m\}$, and similarly for the upper bag. Recall that we have: $z_{\frac{\alpha}{2}}^{(k)} = \hat{q}_{\frac{\alpha}{2}} \circ \Phi(x^{(\text{test})}, \tilde{m}^{(k)}) - s^{(k)}$.

On the other hand, the output predictive interval of Algorithm 13 is then written as:

$$\bullet \quad \widehat{C}_{\alpha}(X^{(\text{test})}, m^{(\text{test})}) = [\widehat{Q}_{\tilde{\alpha}}\left(Z_{\frac{\alpha}{2}}\right); \widehat{Q}_{1-\tilde{\alpha}}\left(Z_{1-\frac{\alpha}{2}}\right)].$$

With these notations, $Z_{\frac{\alpha}{2}}$ can be partitioned as

$$Z_{\frac{\alpha}{2}} = Z_{\frac{\alpha}{2}}^m \cup \left(\bigcup_{\tilde{m}^{(k)} \supset m} Z_{\frac{\alpha}{2}}^{\tilde{m}^{(k)}} \right). \quad (7.8)$$

With

$$\begin{aligned} Z_{\frac{\alpha}{2}} &= \{Z_{\frac{\alpha}{2}}^{(k)}, k \in \text{Cal}\} \\ Z_{\frac{\alpha}{2}}^{(k)} &= \hat{q}_{\frac{\alpha}{2}} \circ \Phi \left(X^{(\text{test})}, \widetilde{M}^{(k)} \right) - S^{(k)} \\ s^{(k)} &= \max(\hat{q}_{\frac{\alpha}{2}}(x_{\text{imp}}^{(k)}) - y^{(k)}, y^{(k)} - \hat{q}_{1-\frac{\alpha}{2}}(x_{\text{imp}}^{(k)})). \end{aligned}$$

The result of Algorithm 12 implies that for any mask $m \in \mathcal{M}$, we have :

$$\mathbb{P} \left(Y^{(\text{test})} \in \widehat{C}_{\alpha} \left(X^{(\text{test})}, m \right) \mid M^{(\text{test})} = m \right) \geq 1 - \alpha,$$

i.e.

$$\mathbb{P} \left(Y^{(\text{test})} \notin \left[\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, m) - \widehat{Q}_{1-\alpha}(S^m); \right. \right. \quad (7.9)$$

$$\left. \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, m) + \widehat{Q}_{1-\alpha}(S^m) \right] \mid M^{(\text{test})} = m \right) \leq \alpha. \quad (7.10)$$

Where: $Q_{1-\alpha}(S)$ is the $(1-\alpha)(1+1/\#S)$ -quantile of S and $S^m = \{s^{(k)} \text{ for } k \in \text{Cal} \text{ and } \widetilde{M}^{(k)} = m\}$. Equivalently:

$$\mathbb{P} \left(Y^{(\text{test})} \in \left[\widehat{Q}_{\alpha} \left(Z_{\frac{\alpha}{2}}^m \right); \widehat{Q}_{1-\alpha} \left(Z_{1-\frac{\alpha}{2}}^m \right) \right] \mid M^{(\text{test})} = m \right) \geq 1 - \alpha. \quad (7.11)$$

In the following Lemma, we show that for $\tilde{m} \supset m$ the result extends under Assumption A4.

Lemma 7.E.1. Assume Assumption A4. For any $m \in \mathcal{M}$, for any $\tilde{m} \supset m$

$$\mathbb{P} \left[\left(Y^{(\text{test})} \in \left[\widehat{Q}_{\alpha} \left(Z_{\frac{\alpha}{2}}^{\tilde{m}} \right); \widehat{Q}_{1-\alpha} \left(Z_{1-\frac{\alpha}{2}}^{\tilde{m}} \right) \right] \right) \mid M^{(\text{test})} = m \right] \geq 1 - \alpha. \quad (7.12)$$

This inequality shows the conservativeness of the quantiles of the bags resulting from larger missing patterns \tilde{m} than m when the construction of the output of Algorithm 13.

While inequality Equation (7.11) is “tight” in the sense that the probability is almost exactly $1 - \alpha$ (item 2 of Theorem 7.5.1), the proof hereafter shows that Equation (7.12) can be pessimistic in terms of actual coverage, as one may have

$$\mathbb{P}[(Y^{(\text{test})} \notin [\widehat{Q}_{\alpha}(Z_{\frac{\alpha}{2}}^{\tilde{m}}); \widehat{Q}_{1-\alpha}(Z_{1-\frac{\alpha}{2}}^{\tilde{m}})]) \mid M^{(\text{test})} = m] \ll \alpha.$$

More precisely, we have the following inequality:

$$\mathbb{E} \left[\mathbb{P} \left(Y^{(\text{test})} \notin \left[\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \widehat{Q}_{1-\alpha}(S^{\tilde{m}}); \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \widehat{Q}_{1-\alpha}(S^{\tilde{m}}) \right] \mid M^{(\text{test})} = m, X_{\text{obs}(m)}^{(\text{test})} \right) \mid M^{(\text{test})} = m \right] \leq \alpha. \quad (7.13)$$

The interpretation of that Lemma is that the intervals resulting from the prediction on $x^{\text{test}}, \tilde{m}$ (more data hidden) and corrected with the residuals of the calibration points $(X^k, M^k = \tilde{m}, Y^k)$ have a *larger* probability of containing Y^{test} , conditionally to $X_{\text{obs}(m)}$ than the interval built using prediction on x^{test}, m (more data available) and corrected with the residuals of the calibration points $(X^k, M^k = m, Y^k)$ (more data available)

Proof of Lemma 7.E.1. We start by invoking Equation (7.10) for \tilde{m} :

$$\mathbb{P} \left(Y^{(\text{test})} \notin \left[\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}}); \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right] \middle| M^{(\text{test})} = \tilde{m} \right) \leq \alpha. \quad (7.14)$$

Consequently, by the tower property of conditional expectations:

$$\mathbb{E} \left[\mathbb{P} \left(Y^{(\text{test})} \notin \left[\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}}); \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right] \middle| M^{(\text{test})} = \tilde{m}, S^{(\tilde{m})}, X_{\text{obs}(\tilde{m})}^{(\text{test})} \right) \middle| M^{(\text{test})} = \tilde{m} \right] \leq \alpha. \quad (7.15)$$

Observe that $\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}})$ is $\{M^{(\text{test})} = \tilde{m}, S^{(\tilde{m})}, X_{\text{obs}(\tilde{m})}^{(\text{test})}\}$ -measurable.

Moreover, by Assumption A4, we have that for any $\delta \in [0, 0.5]$:

$$\frac{Y|(X_{\text{obs}(m)}, M=m)}{q_{1-\delta/2}} \leq \frac{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})}{q_{1-\delta/2}} \quad (7.16)$$

$$\frac{Y|(X_{\text{obs}(m)}, M=m)}{q_{\delta/2}} \geq \frac{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})}{q_{\delta/2}}. \quad (7.17)$$

In other words the conditional distribution of Y given $X_{\text{obs}(\tilde{m})}$ and $M = \tilde{m}$ “stochastically dominates” the conditional distribution of Y given $X_{\text{obs}(m)}$ and $M = m$.

We thus have, with F_Z denoting the cumulative distribution function of Z : $F_{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})}$ the cumulative distribution function of $Y|(X_{\text{obs}(\tilde{m})}, M = \tilde{m})$:

$$\begin{aligned} & \mathbb{P} \left(Y^{(\text{test})} \notin \left[\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}}); \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right] \middle| M^{(\text{test})} = \tilde{m}, S^{(\tilde{m})}, X_{\text{obs}(\tilde{m})}^{(\text{test})} \right) \\ &= 1 - \left[F_{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})} \left(\hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right) - F_{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})} \left(\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right) \right] \\ &\stackrel{(i)}{\geq} 1 - \left[F_{Y|(X_{\text{obs}(m)}, M=m)} \left(\hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right) - F_{Y|(X_{\text{obs}(m)}, M=m)} \left(\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right) \right] \\ &= \mathbb{P} \left(Y^{(\text{test})} \notin \left[\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}}); \hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right] \middle| M^{(\text{test})} = m, S^{(\tilde{m})}, X_{\text{obs}(m)}^{(\text{test})} \right). \end{aligned} \quad (7.18)$$

At (i) we use (7.17) $F_{Y|(X_{\text{obs}(m)}, M=m)}(\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}})) \leq F_{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})}(\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}}))$, and (7.16): $F_{Y|(X_{\text{obs}(m)}, M=m)}(\hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}})) \geq F_{Y|(X_{\text{obs}(\tilde{m})}, M=\tilde{m})}(\hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}}))$ by A4. Remark that here we assume that $\left(\hat{q}_{1-\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) + \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right) \geq \text{median}(Y^{(\text{test})}|(X_{\text{obs}(\tilde{m})}^{(\text{test})}, M = \tilde{m}))$ and $\left(\hat{q}_{\frac{\alpha}{2}} \circ \Phi(X^{(\text{test})}, \tilde{m}) - \hat{Q}_{1-\alpha}(S^{\tilde{m}}) \right) \leq \text{median}(Y^{(\text{test})}|(X_{\text{obs}(\tilde{m})}^{(\text{test})}, M = \tilde{m}))$.

We obtain Equation (7.13) in Lemma 7.E.1 by plugging (7.18) in (7.15), then Equation (7.12) by the tower property. \square

Theorem 7.E.2. Assume the missing mechanism is MCAR, and that Assumptions A1 to A3 hold. Additionally Assumption A4 is satisfied.

Consider the partition described in Equation (7.8), and consider CP-MDA-Nested running on a test point with missing pattern $m^{(\text{test})}$, with any of the following outputs, instead of l. 15 $\hat{C}_\alpha(x^{(\text{test})}, m^{(\text{test})}) = [\hat{Q}_{\hat{\alpha}}(Z_{\frac{\alpha}{2}}); \hat{Q}_{1-\hat{\alpha}}(Z_{1-\frac{\alpha}{2}})]:$

1. $\widehat{C}_\alpha(x^{(test)}, m^{(test)}) = [\widehat{Q}_{\tilde{\alpha}}(Z_{\frac{\alpha}{2}}^{\tilde{m}}); \widehat{Q}_{1-\tilde{\alpha}}(Z_{1-\frac{\alpha}{2}}^{\tilde{m}})]$ where $\tilde{m} \supset m^{(test)}$ is an arbitrary choice.
2. $\widehat{C}_\alpha(x^{(test)}, m^{(test)}) = [\widehat{Q}_{\hat{\alpha}}(Z_{\frac{\alpha}{2}}^{\hat{m}}); \widehat{Q}_{1-\hat{\alpha}}(Z_{1-\frac{\alpha}{2}}^{\hat{m}})]$ where \hat{m} is a randomly selected pattern in $\{\tilde{m}, \tilde{m} \supset m^{(test)}\}$, possibly with varying probability depending on the cardinality of the sets $Z_{\alpha/2}^{\tilde{m}}$.

Then the resulting algorithm is mask-conditionally-valid.

Proof of Theorem 7.E.2. The proof immediately follows from Equation (7.12), and gives the result without difficulty for any arbitrary pattern or random variable independent of all other randomness.

Extension to a choice that involves the cardinality of the sets $Z_{\alpha/2}^{\tilde{m}}$, leveraging the independence between these cardinals and the coverage properties (same as in the proof of Theorem 7.E.1). \square

Then, the proof of Theorem 7.5.2 (marginal validity of the CP-MDA-Nested) is direct by marginalizing the result of Theorem 7.E.2. \square

7.F Infinite data results

Proposition 7.6.1 (ℓ_β -consistency of an universal learner). *Let $\beta \in [0, 1]$. If X admits a density on \mathcal{X} , then, for almost all imputation function $\Phi \in \mathcal{F}_\infty^I$, the function $g_{\ell_\beta, \Phi}^* \circ \Phi$ is Bayes optimal for the pinball risk of level β .*

Proof of Proposition 7.6.1. The proof starts in the exact same way than [Le Morvan et al. \(2021\)](#), based on their Lemmas A.1 and A.2. For completeness, we copy here the statements of these lemmas without their proof and rewrite the two first parts of the main proof.

Let Φ be an imputation function such that for each missing data pattern m , $\varphi^m \in \mathcal{C}^\infty(\mathbb{R}^{|\text{obs}(m)|}, \mathbb{R}^{|\text{mis}(m)|})$.

Lemma 7.F.1 (Lemma A.1 in [Le Morvan et al. \(2021\)](#)).

Let $\varphi^m \in \mathcal{C}^\infty(\mathbb{R}^{|\text{obs}(m)|}, \mathbb{R}^{|\text{mis}(m)|})$ be the imputation function for missing data pattern m , and let $\mathcal{M}^m = \{x \in \mathbb{R}^d : x_{\text{mis}(m)} = \varphi^m(x_{\text{obs}(m)})\}$. For all m , \mathcal{M}^m is an $|\text{obs}(m)|$ -dimensional manifold.

In Lemma 7.F.1, \mathcal{M}^m represents the manifold in which the data points are sent once imputed by φ^m . Lemma 7.F.1 states that this manifold is of dimension $|\text{obs}(m)|$.

Lemma 7.F.2 (Lemma A.2 in [Le Morvan et al. \(2021\)](#)). *Let m and m' be two distinct missing data patterns with the same number of missing (resp. observed) values $|\text{mis}|$ (resp. $|\text{obs}|$). Let $\varphi^m \in \mathcal{C}^\infty(\mathbb{R}^{|\text{obs}(m)|}, \mathbb{R}^{|\text{mis}(m)|})$ be the imputation function for missing data pattern m , and let $\mathcal{M}^m = \{x \in \mathbb{R}^d : x_{\text{mis}(m)} = \varphi^m(x_{\text{obs}(m)})\}$. We define similarly $\Phi^{(m')}$ and $\mathcal{M}^{(m')}$. For almost all imputation functions φ^m and $\Phi^{(m')}$,*

$$\dim(\mathcal{M}^m \cap \mathcal{M}^{(m')}) = \begin{cases} 0 & \text{if } |\text{mis}| > \frac{d}{2} \\ d - 2|\text{mis}| & \text{otherwise.} \end{cases}$$

Note that, as by Lemma 7.F.1 $\dim(\mathcal{M}^m) = \dim(\mathcal{M}^{(m')}) = |\text{obs}| = d - |\text{mis}|$, Lemma 7.F.2 states that $\dim(\mathcal{M}^m \cap \mathcal{M}^{(m')}) \leq \dim(\mathcal{M}^m) = \dim(\mathcal{M}^{(m')})$.

Now, to prove Proposition 7.6.1 the missing data patterns are ordered as in Le Morvan et al. (2021): the first one will be the one in which all the variables are missing, while the last one will be the one in which all the variables are observed. For two data patterns with the same number of missing variables, the ordering is picked at random. We denote by $m(i)$ the i -th missing data pattern according to this ordering.

We are going to build a function g_Φ which, composed with Φ , will reach the ℓ -Bayes risk.

For each missing data pattern, and starting by $m(1)$ of all variables missing, we can define g_Φ on the data points from the current missing data pattern. More precisely, for each i , g_Φ is built for every imputed data point belonging to $\mathcal{M}^{(m(i))}$ except for those already considered in previous steps (one imputed data point can belong to multiple manifolds):

$$\forall Z = \Phi(X, M) \in \mathcal{M}^{(m(i))} \setminus \bigcup_{k < i} \mathcal{M}^{(m(k))}, \quad g^*(Z) = \tilde{f}^*(\tilde{X})$$

That is, $g_\Phi \circ \Phi(X, M)$ will equal $\tilde{f}^*(X, M)$ except possibly if $\Phi(X, M) = \Phi(\tilde{Y})$ for some \tilde{Y} that has more missing values than X, M . Therefore, for each missing data pattern $m(i)$, $g_\Phi \circ \Phi$ equals \tilde{f}^* except on $\bigcup_{k < i} \mathcal{M}^{(m(k))}$. The question that remains is: what is the dimension of $\mathcal{M}^{(m(i))} \cap (\bigcup_{k < i} \mathcal{M}^{(m(k))})$, these points for which there is no necessarily equality between $g_\Phi \circ \Phi$ and \tilde{f}^* . First, note that $\mathcal{M}^{(m(i))} \cap (\bigcup_{k < i} \mathcal{M}^{(m(k))}) = \bigcup_{k < i} (\mathcal{M}^{(m(i))} \cap \mathcal{M}^{(m(k))})$. For each space in this reunion, there are two cases:

- either $|\text{obs}(m(k))| < |\text{obs}(m(i))|$: using Lemma 7.F.1, $\dim(\mathcal{M}^{(m(k))}) = |\text{obs}(m(k))| < |\text{obs}(m(i))| = \dim(\mathcal{M}^{(m(i))})$. Thus, $\mathcal{M}^{(m(i))} \cap \mathcal{M}^{(m(k))}$ is of measure zero in $\mathcal{M}^{(m(i))}$.
- either $|\text{obs}(m(k))| = |\text{obs}(m(i))|$: using Lemma 7.F.2, $\mathcal{M}^{(m(i))} \cap \mathcal{M}^{(m(k))}$ is of dimension 0 or smaller than $\dim(\mathcal{M}^{(m(i))})$, thus it is of measure zero in $\mathcal{M}^{(m(i))}$.

Therefore, the set of data points for which $g_\Phi \circ \Phi$ does not equal the oracle is of measure 0 for each missing data pattern.

Let $\beta \in [0, 1]$. We can now write down the ℓ_β -risk of this built function:

$$\begin{aligned} \mathbb{E}[\ell_\beta(Y, g^* \circ \Phi(X, M))] &= \mathbb{E}[\rho_\beta(Y - g^* \circ \Phi(X, M))] \\ &= \mathbb{E}\left[\rho_\beta\left(Y - \tilde{f}^*(X, M) + \tilde{f}^*(X, M) - g^* \circ \Phi(X, M)\right)\right] \\ (i) \leq \mathbb{E}\left[\rho_\beta\left(Y - \tilde{f}^*(X, M)\right)\right] &+ \mathbb{E}\left[\rho_\beta\left(\tilde{f}^*(X, M) - g^* \circ \Phi(X, M)\right)\right] \\ &\leq \mathcal{R}_{\ell_\beta}^* + \mathbb{E}\left[\rho_\beta\left(\tilde{f}^*(X, M) - g^* \circ \Phi(X, M)\right)\right], \end{aligned}$$

where (i) holds thanks to the shape of ρ_β . For any $w \in \mathbb{R}$ and any $\lambda \in \mathbb{R}_+$:

$$\begin{aligned} \rho_\beta(\lambda w) &= \beta \lambda |w| \mathbb{1}_{w \geq 0} + (1 - \beta) \lambda |w| \mathbb{1}_{w \leq 0} \\ \rho_\beta(\lambda w) &= \lambda \rho_\beta(w). \end{aligned}$$

Furthermore, ρ_β is convex, thus for any $(u, v) \in \mathbb{R}^2$:

$$\begin{aligned}\rho_\beta \left(\frac{1}{2}u + \frac{1}{2}v \right) &\leq \frac{1}{2}\rho_\beta(u) + \frac{1}{2}\rho_\beta(v) \\ \frac{1}{2}\rho_\beta(u + v) &\leq \frac{1}{2}\rho_\beta(u) + \frac{1}{2}\rho_\beta(v) \\ \rho_\beta(u + v) &\leq \rho_\beta(u) + \rho_\beta(v).\end{aligned}$$

As \tilde{f}^* and $g^* \circ \Phi$ are equals almost everywhere on each missing subspace, it holds that

$$\mathbb{E} \left[\rho_\beta \left(\tilde{f}^*(X, M) - g^* \circ \Phi(X, M) \right) \right] = 0.$$

Indeed, decomposing by pattern one can write:

$$\begin{aligned}\mathbb{E} \left[\rho_\beta \left(\tilde{f}^*(X, M) - g^* \circ \Phi(X, M) \right) \right] &= \\ \sum_{M=m} \mathbb{P}(M = m) \mathbb{E} \left[\rho_\beta \left(\tilde{f}^*(X, M) - g^* \circ \Phi(X, M) \right) \mid M = m \right]\end{aligned}$$

and thus by equality almost everywhere for each pattern every term in this sum is null.

Therefore one obtains:

$$\mathbb{E} [\ell_\beta(Y, g^* \circ \Phi(X, M))] \leq \mathcal{R}_{\ell_\beta}^*.$$

Thus:

$$\mathbb{E} [\ell_\beta(Y, g^* \circ \Phi(X, M))] = \mathcal{R}_{\ell_\beta}^*,$$

and $g^* \circ \Phi$ is Bayes optimal. This implies that $\mathcal{R}_{\ell_\beta, \Phi}^* = \mathcal{R}_{\ell_\beta}^*$. Thus, a universally consistent algorithm learning g_Φ chained with Φ will lead to a Bayes consistent function. \square

Proof of Corollary 7.6.1. Corollary 7.6.1 states that “For any missing mechanism, for almost all imputation function $\Phi \in \mathcal{F}_\infty^I$, if $F_{Y|(X_{\text{obs}(M)}, M)}$ is continuous, a universally consistent quantile regressor trained on the imputed data set yields asymptotic conditional coverage.”.

Let $\beta \in [0, 1]$.

Remark that Proposition 7.6.1 states that for any missing mechanism, for almost all imputation function $\Phi \in \mathcal{F}_\infty^I$ a universally consistent quantile regressor trained on the imputed data set achieves the Bayes risk asymptotically. We will thus show that any ℓ_β -Bayes predictor f_β^* (any function achieving the ℓ_β -Bayes-risk) is such that $\mathbb{P}(Y \leq f_\beta^*(X, M) \mid X_{\text{obs}(M)}, M) = \beta$ if $F_{Y|(X_{\text{obs}(M)}, M)}$ is continuous. Therefore, any two Bayes predictors $f_{\alpha/2}^*$ and $f_{1-\alpha/2}^*$ form an interval $[f_{\alpha/2}^*(X, M); f_{1-\alpha/2}^*(X, M)]$ that achieves conditional coverage (conditionally to $X_{\text{obs}(M)}$ and M).

Let f_β^* be a ℓ_β -Bayes predictor. Then:

$$\begin{aligned}f_\beta^* &\in \arg \min_{f: \mathcal{X} \times \mathcal{M} \rightarrow \mathbb{R}} \mathbb{E} [\rho_\beta(Y - f(X, M))] \\ &= \mathbb{E} [\mathbb{E} [\rho_\beta(Y - f(X, M)) \mid X_{\text{obs}(M)}, M]].\end{aligned}$$

Let $(x, m) \in \mathcal{X} \times \mathcal{M}$. Denote $H_{x,m}(z) := \mathbb{E} [\rho_\beta(Y - z) | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m]$. As $Y \neq z$ almost surely, we have:

$$\begin{aligned}
H'_{x,m}(z) &= \mathbb{E} [-\rho'_\beta(Y - z) | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m] \\
&= \mathbb{E} [-(-\beta \mathbb{1}_{Y-z \geq 0} + (1 - \beta) \mathbb{1}_{Y-z \leq 0}) | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m] \\
&= \mathbb{E} [\beta \mathbb{1}_{Y \geq z} - (1 - \beta) \mathbb{1}_{Y \leq z} | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m] \\
&= \beta \mathbb{P}(Y \geq z | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m) \\
&\quad - (1 - \beta) \mathbb{P}(Y \leq z | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m) \\
&= \beta (1 - \mathbb{P}(Y \leq z | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m)) \\
&\quad - (1 - \beta) \mathbb{P}(Y \leq z | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m) \\
H'_{x,m}(z) &= \beta - \mathbb{P}(Y \leq z | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m).
\end{aligned}$$

Therefore $H'_{x,m}(z) \leq 0$ if and only if $\beta \leq \mathbb{P}(Y \leq z | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m)$.

Thus, z minimizes $H_{x,m}$ if and only if $\beta = \mathbb{P}(Y \leq z | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m)$.

If $F_{Y|(X_{\text{obs}(M)}, M)}$ is continuous, there exists at least a solution, that might not be unique if it is not additionally strictly increasing. Therefore, if $F_{Y|(X_{\text{obs}(M)}, M)}$ is continuous, all the ℓ_β -Bayes predictors can be written as $f_\beta^*(x, m) = q_{x,m}$ with

$$\mathbb{P}(Y \leq q_{x,m} | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m) = \mathbb{P}(Y \leq f_\beta^*(x, m) | X_{\text{obs}(M)} = x_{\text{obs}(m)}, M = m) = \beta.$$

□

7.G Experimental study

7.G.1 Settings detail

Quantile Neural Network. The architecture and optimization of the Quantile Neural Network used in the experiments is taken from [Sesia and Romano \(2021\)](#) (their code is freely available). This is the description provided in the original paper of the neural network: “The network is composed of three fully connected layers with a hidden dimension of 64, and ReLU activation functions. We use the pinball loss to estimate the conditional quantiles, with a dropout regularization of rate 0.1. The network is optimized using Adam [Kingma and Ba \(2014\)](#) with a learning rate equal to 0.0005. We tune the optimal number of epochs by cross validation, minimizing the loss function on the hold-out data points; the maximal number of epochs is set to 2000.”

7.G.2 Gaussian linear results

Figure 7.9 is the analogous of Figure 7.8, but by evaluating the performances on the mask leading to the highest coverage.

Hereafter, we present in Figure 7.10 the exact same figure than Figure 7.3 but with a panel (the first) for vanilla QR. The 3 other methods are displayed again to facilitate the comparison.

Finally, Figure 7.11 is the analogous of Figure 7.10, but for a training set containing 1000 observations and a calibration set containing 500 observations.

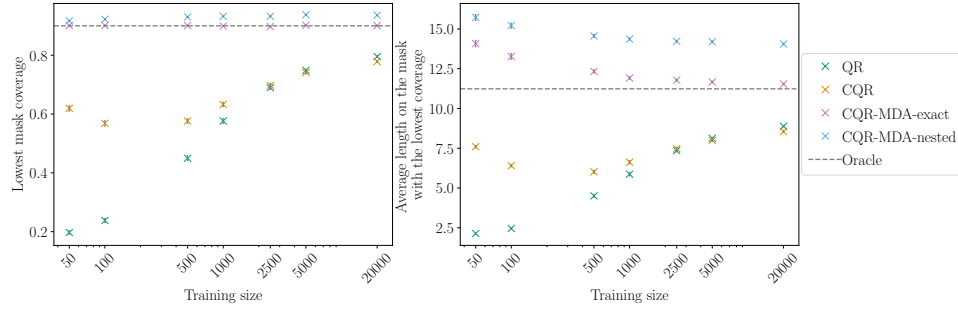


Figure 7.8: Coverage and interval's length for the mask leading to the lowest coverage. Model is NN. Calibration size fixed to 1000. The mask is concatenated in the features. Data is imputed using Iterative Ridge. 100 repetitions allow to display error bars, corresponding to standard error.

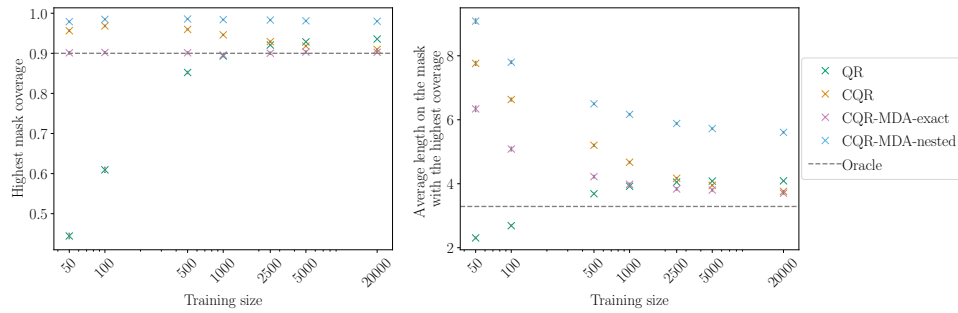


Figure 7.9: Coverage and interval's length for the mask leading to the highest coverage. See caption of Figure 7.8 for the setting.

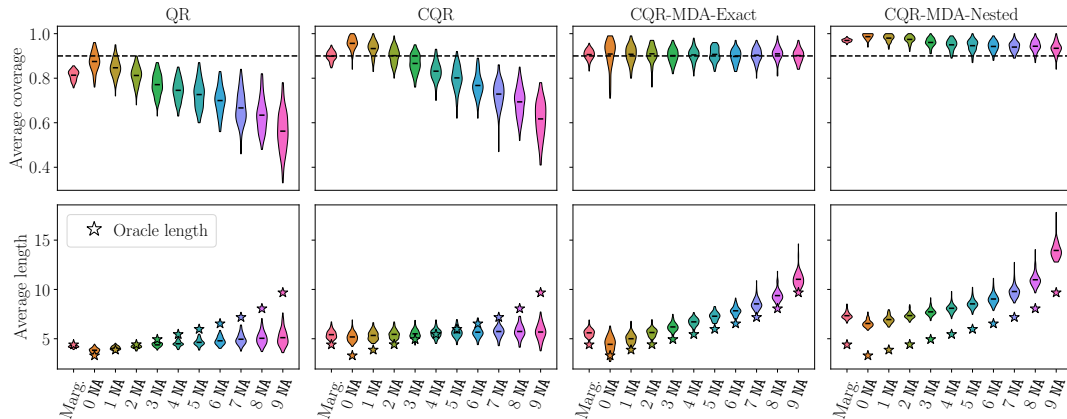


Figure 7.10: Average coverage (top) and length (bottom) as a function of the pattern size, i.e. the number of missing values (NA). First violin plot corresponds to marginal coverage. Stars correspond to the oracle length. Settings are: model is NN, train size is 500, calibration size is 250. The marginal test set includes 2000 observations. The conditional test set includes 100 individuals for each possible missing data pattern size. The mask is concatenated to the features. Data is imputed using Iterative Ridge. 100 repetitions are performed.

7.G.3 Higher proportion of missing values

We present synthetic experiments where the proportion of MCAR missing values is of 40% (instead of 20% in Figure 7.3). Except from this, the settings are exactly the same than the ones of Figure 7.3. Precisely, the data is generated with $d = 10$ according to Model 7.4.1, with $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, \dots, 1)^T$ and $\Sigma = \varphi(1, \dots, 1)^T(1, \dots, 1) +$

$(1 - \varphi)I_d$, $\varphi = 0.8$, Gaussian noise $\varepsilon \sim \mathcal{N}(0, 1)$ and the following regression coefficients $\beta = (1, 2, -1, 3, -0.5, -1, 0.3, 1.7, 0.4, -0.3)^T$. For each pattern size, 100 observations are drawn according to the distribution of $M|\text{size}(M)$ in the test set. The training and calibration sizes are respectively 500 and 250. The experiment is repeated 100 times. The results are displayed in Figure 7.12.

Interestingly, although expected, these experiments lead CP-MDA-Exact to frequently output infinite intervals. This is because the subsampling step with few calibration data – with respect to the dimension and proportion of missing values – reached a point where there are not enough observations for CP-MDA-Exact to calibrate accurately for some patterns.

To compare CP-MDA-Exact and CP-MDA-Nested in this setting, Figure 7.12 is obtained by replacing the infinite intervals by $\max_{k \in Tr \cup Cal} y^{(k)} - \min_{k \in Tr \cup Cal} y^{(k)}$. It highlights that CP-MDA-Exact is less *efficient* (i.e. outputs larger intervals) than CP-MDA-Nested for patterns with less than 4 NAs. With a smaller calibration set or a higher proportion of missing values, this effect would be amplified and generalized to more patterns. Figure 7.12 also emphasizes that CP-MDA-Exact leads to more coverage variability than CP-MDA-Nested, on the patterns for which CP-MDA-Exact does not almost surely cover.

7.G.4 Semi-synthetic experiments

In the semi-synthetic experiments, two settings are examined: one where the training size is small in comparison to the number of parameters of the Neural Network – “Medium” –, and one where the training size is even smaller so that some masks have a really low (or null) frequency of appearance in the training set – “Small”. In both cases, the calibration size is approximately half the training size. Figure 7.4 presented the results for the “Medium” case.

Figure 7.13 represents the results for these settings, using the same parameters than Figure 7.4. For the results on the two other `meps` data sets (`meps_20` and `meps_21`) see Figure 7.14, which repeats the visualisation of `meps_19` to ease comparison.

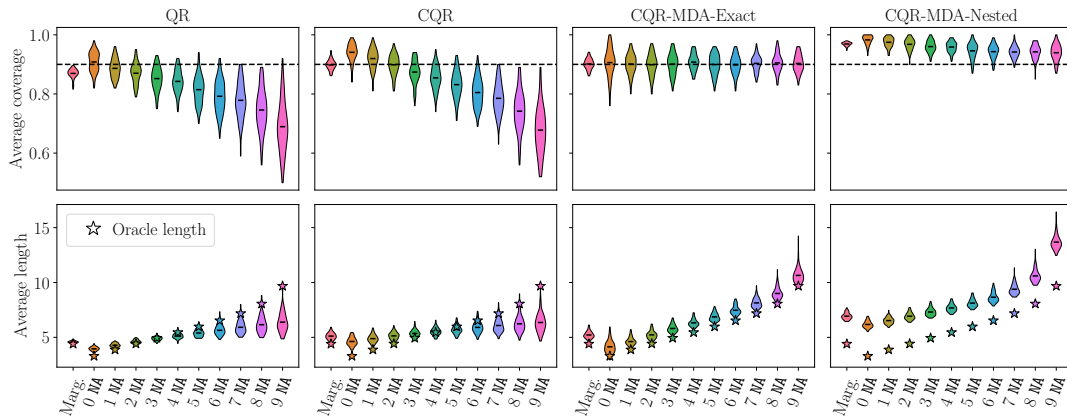


Figure 7.11: Model is NN. Train size is 1000. Calibration size fixed to 500. The marginal test set includes 2000 observations. The conditional test set includes 100 individuals for each possible missing data pattern size. The mask is concatenated in the features. Data is imputed using Iterative Ridge. 100 repetitions are performed.

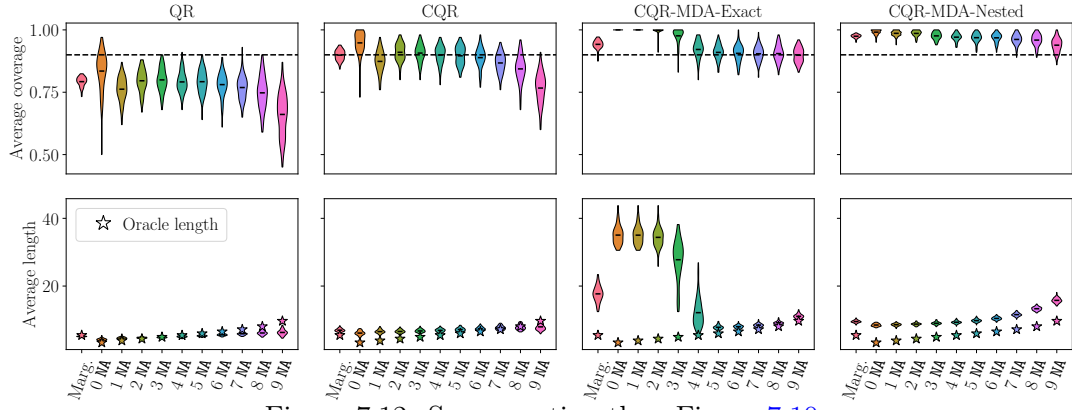


Figure 7.12: Same caption than Figure 7.10.

Table 7.1: Semi-synthetic settings: training and calibration sizes for each of the 6 data sets depending on the setting.

		meps_19 $d = 139$ $l = 5$ $n = 15785$	meps_20 $d = 139$ $l = 5$ $n = 17541$	meps_21 $d = 139$ $l = 5$ $n = 15656$	bio $d = 9$ $l = 9$ $n = 45730$	bike $d = 18$ $l = 4$ $n = 10886$	concrete $d = 8$ $l = 8$ $n = 1030$
Small	Tr size	500	500	500	500	500	330
	Cal size	250	250	250	250	250	100
Medium	Tr size	1000	1000	1000	1000	1000	630
	Cal size	500	500	500	500	500	200

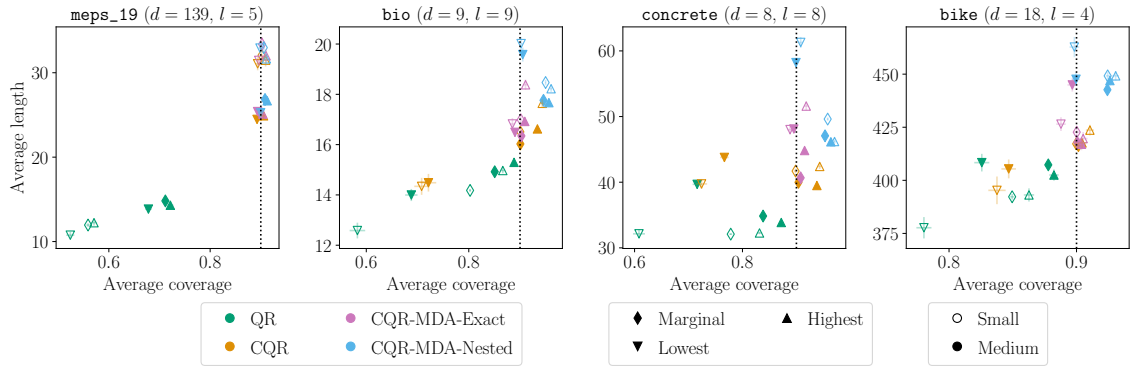


Figure 7.13: Model is NN. The mask is concatenated in the features. Data is imputed using Iterative Ridge. 100 repetitions are performed, allowing to display the standard error as error bars. The vertical dotted lines represent the target coverage of 90%.

7.G.5 Real data

Data set description. Sportisse et al. (2020) selected 7 variables to model the level of platelets, after discussion with medical doctors. Thus, we followed their pipeline. Here are the 7 variables used:

- **Age:** the age of the patient (no missing values);
- **Lactate:** the conjugate base of lactic acid, upon arrival at the hospital (17.66% missing values);

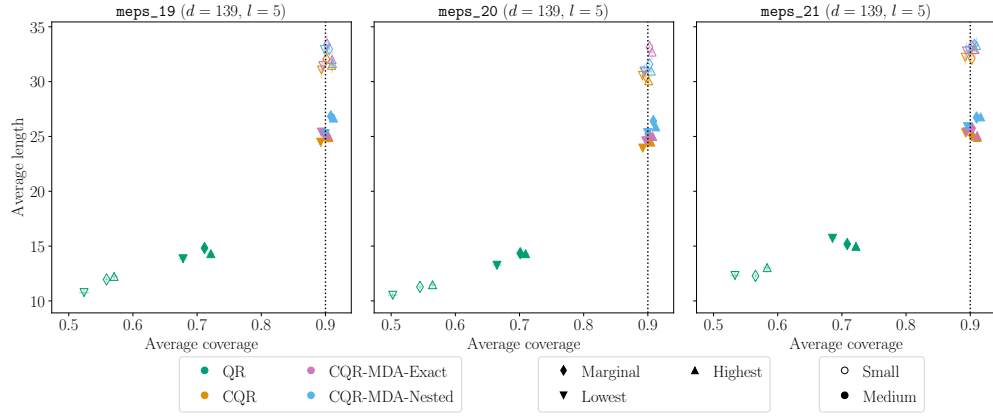


Figure 7.14: Same caption than Figure 7.13.

- **Delta_hemo**: the difference between the hemoglobin upon arrival at hospital and the one in the ambulance (23.82% missing values);
- **VE**: binary variable indicating if a Volume Expander was applied in the ambulance. A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system (2.46% missing values);
- **RBC**: a binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed (0.37% missing values);
- **SI**: the shock index. It indicates the level of occult shock based on heart rate (HR) and systolic blood pressure (SBP), that is $SI = \frac{HR}{SBP}$, upon arrival at hospital (2.09% missing values);
- **HR**: the heart rate measured upon arrival of hospital (1.62% missing values).

Splitting strategy. To study the coverage conditionally on the masks, we must handle the scarcity of some of them. For each individual in the data set, we make only one prediction, this way avoiding too many repetitions of the same test point when computing the average. We split the data set into 5 folds, and predict on each fold by training the procedure on the 4 others, with 15390 observations for training, and 7694 for calibration.

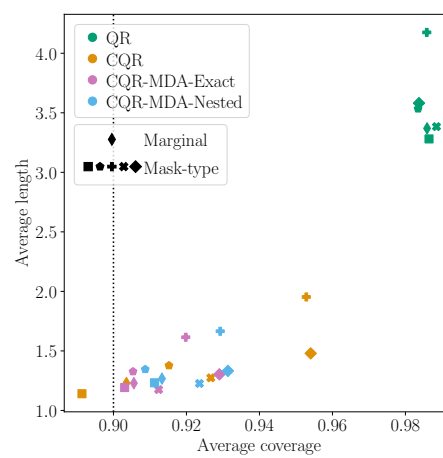


Figure 7.15: Average coverage and length on the TraumaBase[®] data when predicting the platelets level. Colors correspond to the methods. Diamond (◆) corresponds to taking the average among all individuals. Other symbols correspond to computing the average among the individuals having a fixed mask. The vertical dotted line represents the target coverage of 90%. Model is NN. The mask is concatenated to the features. Imputation is Iterative Ridge. Each individual is predicted using 15390 observations for training, and 7694 for calibration.

Chapter 8

Predictive Uncertainty Quantification with Missing Covariates

Predictive uncertainty quantification is crucial in decision-making problems. We investigate how to adequately quantify predictive uncertainty with missing covariates. A bottleneck is that missing values induce heteroskedasticity on the response's predictive distribution given the observed covariates. Thus, we focus on building predictive sets for the response that are valid *conditionally* to the missing values pattern. We show that this goal is impossible to achieve informatively in a distribution-free fashion, and we propose useful restrictions on the distribution class. Motivated by these hardness results, we characterize how missing values and predictive uncertainty intertwine. Particularly, we rigorously formalize the idea that the more missing values, the higher the predictive uncertainty. Then, we introduce a generalized framework, coined **CP-MDA-Nested***, outputting predictive sets in both regression and classification. Under independence between the missing value pattern and both the features and the response (an assumption justified by our hardness results), these predictive sets are valid conditionally to any pattern of missing values. Moreover, it provides great flexibility in the trade-off between *statistical variability* and *efficiency*. Finally, we experimentally assess the performances of **CP-MDA-Nested*** beyond its scope of theoretical validity, demonstrating promising outcomes in more challenging configurations than independence.

Contents

8.1	Introduction	167
8.1.1	Literature's background	170
8.1.2	Overview of our contributions (and outline)	171
8.2	When is Mask-Conditional-Validity (MCV) a too lofty goal?	173
8.2.1	Fully distribution-free result	174
8.2.2	Restricting the class of admissible missingness distributions	176
8.3	Links between missing covariates and predictive uncertainty	177
8.3.1	Increasing uncertainty with nested masks	177
8.3.2	Guidelines for practitioners: which information through imputation for quantile regression?	183
8.4	Principled unified Missing Data Augmentation (MDA) framework: CP-MDA-Nested*	184
8.4.1	Presentation of CP-MDA-Nested*	184
8.4.2	Theoretical guarantees on CP-MDA-Nested and CP-MDA-Nested* leveraging their connection to leave-one-out CP	189
8.5	A practical glimpse on the impacts of breaking the distribution's assumptions	191
8.5.1	Experiments under $\mathcal{P}_{\text{MCAR}, Y \perp\!\!\!\perp M X}$	191
8.5.2	Beyond MCAR	194
8.5.3	Breaking $Y \perp\!\!\!\perp M X$ Assumption	197
8.A	Hardness results	199
8.B	Link between missing covariates and uncertainty	204
8.C	Leave-one-out predictive sets for randomized algorithms	208
8.D	Theory on CP-MDA-Nested* and CP-MDA-Nested	210

8.1 Introduction

Predictive uncertainty quantification. Over the last decades, major research efforts on statistical and machine learning algorithms have enabled them to leverage large data sets. They are now used to support high-stakes decision-making problems such as medical, energy, or civic applications, to name just a few. To ensure the safe deployment of these models and their adoption by society, it is crucial to acknowledge that these point predictions remain uncertain, and to quantify this uncertainty, communicating the limits of predictive performance. Therefore, uncertainty quantification has received much attention in recent years, particularly in the form of building prediction sets.

Formally, the aim is to build a predictive set for the response $Y \in \mathcal{Y}$, after observing the random vector $X \in \mathcal{X} \subseteq \mathbb{R}^d$ which contains $d \in \mathbb{N}^*$ explanatory variables. Given a *miscoverage level* $\alpha \in [0, 1]$, a *marginally valid* predictive set $\mathcal{C}_\alpha(\cdot)$ is a function satisfying

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(X)) \geq 1 - \alpha. \quad (8.1)$$

The goal is that $\mathcal{C}_\alpha(\cdot)$ is as small as possible while being marginally valid. Distribution-free uncertainty quantification tools are powerful as they require minimal assumptions on the data generation process—typically only access to a sequence of n exchangeable data points—making them usable on a wide range of applications, unlike traditional probabilistic approaches.

Importantly, it has to be noted that Equation (8.1) averages among all probable (X, Y) , and thus might over-cover easy data points (say, e.g., young patients) at the cost of under-covering harder data points (say, e.g., older patients). Therefore, one branch of the literature studied how Equation (8.1) could be turned into a stronger goal. Specifically, [Vovk \(2012\)](#); [Lei and Wasserman \(2014\)](#); [Barber et al. \(2021a\)](#) emphasize trade-offs between theory and practice. They investigate the implications of designing a practical distribution-free method, that is one which outputs sets $\mathcal{C}_\alpha(\cdot)$ such that

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(x) | X = x) \geq 1 - \alpha, \text{ for any } x \in \mathcal{X}. \quad (8.2)$$

Unfortunately, they showed that Equation (8.2) is impossible to achieve in an informative way (i.e., typically $\mathcal{C}_\alpha(\cdot) \equiv \mathcal{Y}$ with high probability) if no assumptions on the data distributions are made. Moreover, finding natural relaxations that are compatible with informative distribution-free predictive sets seems also hard: restrictions to conditioning on $x \in \mathcal{X}$, for an arbitrary mass positive $\mathcal{X} \subseteq \mathcal{X}$, is still hard to achieve informatively ([Barber et al., 2021a](#)).

Missing values. Somewhat paradoxically, as the quantity of data rises, the number of missing data also increases. This phenomenon is modeled through the introduction of a third random variable called the *mask* or *missing pattern*, denoted by $M \in \mathcal{M} \subseteq \{0, 1\}^d$, encoding which variables have not been observed. That is, the mask M is the indicator vector such that for any $j \in \llbracket 1, d \rrbracket$, $M_j = 1$ whenever X_j is missing (not observed), and $M_j = 0$ otherwise. As a consequence, we are working on $\mathcal{P} := \{\text{distributions on } (\mathcal{X}, \mathcal{M}, \mathcal{Y})\}$. For a given pattern $m \in \mathcal{M}$, $X_{\text{obs}(m)}$ is the random vector of observed features, and $X_{\text{mis}(m)}$ is the

random vector of unobserved ones. For example, if we observe $(\text{NA}, 6, 2)$ then $m = (1, 0, 0)$ and $x_{\text{obs}(m)} = (6, 2)$. Notice that the number of different missing patterns, i.e., the size or cardinality of $\mathcal{M} := \#\mathcal{M}$, typically grows exponentially in the dimension (for $\mathcal{M} = \{0, 1\}^d$ there are 2^d different patterns).

The way we deal with those missing values will typically depend on the downstream task at hand. While there is a vast range of studies in the inferential setting (Little, 2019; Josse and Reiter, 2018) with numerous implementations (Mayer et al., 2019), the research effort is scarcer on the prediction framework (Josse et al., 2019; Le Morvan et al., 2020b,a, 2021; Ayme et al., 2022; Van Ness et al., 2022; Ayme et al., 2023; Zaffran et al., 2023; Ayme et al., 2024). It is mostly limited to *point prediction*, except for Zaffran et al. (2023). The literature on both inference and prediction highlights the necessity of taking into account the missingness distribution. Following Rubin (1976), we consider three well-known missingness mechanisms.

Definition 8.1.1 (Missing Completely At Random (MCAR)). The missing pattern distribution is said to be Missing Completely At Random (MCAR) if $M \perp\!\!\!\perp X$. We denote $\mathcal{P}_{\text{MCAR}}$ the corresponding set of distributions, i.e. $\mathcal{P}_{\text{MCAR}} := \{P \in \mathcal{P}, \text{ such that for any } m \in \mathcal{M}, \mathbb{P}_P(M = m|X) = \mathbb{P}_P(M = m), \text{ that is } M \perp\!\!\!\perp X\}$.

Definition 8.1.2 (Missing At Random (MAR)). The missing pattern distribution is said to be Missing At Random (MAR) if M only depends on the observed components of X . We denote \mathcal{P}_{MAR} the corresponding set of distributions, i.e. $\mathcal{P}_{\text{MAR}} := \{P \in \mathcal{P}, \text{ such that for any } m \in \mathcal{M}, \mathbb{P}_P(M = m|X) = \mathbb{P}_P(M = m|X_{\text{obs}(m)})\}$.

Definition 8.1.3 (Missing Non At Random (MNAR)). The missing pattern distribution is said to be Missing Non At Random (MNAR) if M can depend on the observed values of X but also on its missing components. We denote $\mathcal{P}_{\text{MNAR}}$ the corresponding set of distributions, i.e. $\mathcal{P}_{\text{MNAR}} := \mathcal{P}$.

Remark 8.1.1. We thus have $\mathcal{P}_{\text{MCAR}} \subset \mathcal{P}_{\text{MAR}} \subset \mathcal{P}_{\text{MNAR}} = \mathcal{P}$.

Predictive framework with missing covariates. In a predictive framework, the dependence between Y and M plays a key role, maybe even bigger than the relationship between (X, M) . Indeed, in some situations, Y can be a direct function of M : the missingness conveys in itself information about the label. Therefore, these cases are particularly challenging in a predictive framework, as some patterns on the one hand can induce an important label distributional shift, and on the other hand be rarely observed due to the high cardinality of \mathcal{M} . Thus, we focus on settings where there is *not* such a direct dependency, that is Assumption A1. Yet, as we will show in the paper, it remains that the lack of observation of some features influences the uncertainty of the prediction of Y from $X_{\text{obs}(M)}$.

Assumption A1 (M does not explain Y). We say that Y is independent of M given X if $Y \perp\!\!\!\perp M | X$. The associated distribution belongs to $\mathcal{P}_{Y \perp\!\!\!\perp M | X}$.

Definitions 8.1.1 to 8.1.3 and Assumption A1 will be our main assumptions on the joint distribution of (X, M, Y) throughout the manuscript. Our interest is in building predictive sets from n observations $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^n$ on a new test point $(X^{(n+1)}, M^{(n+1)}, Y^{(n+1)})$. We thus also make assumptions on the *links between those samples*: the usual backbone assumption is that we have access to $n+1$ independent and identically distributed (i.i.d.) draws from a distribution Q in a set \mathcal{Q} , with \mathcal{Q} being typically one of $\mathcal{P}_{\text{MCAR}}, \mathcal{P}_{\text{MAR}}, \mathcal{P}$, etc. The data distribution thus belongs to $\{Q^{\otimes(n+1)}, Q \in \mathcal{Q}\}$, which we denote $\mathcal{Q}^{\otimes(n+1)}$. Furthermore, we also consider here a relaxation of i.i.d., namely *exchangeability*, which is often sufficient to obtain guarantees in distribution-free predictive inference.

Assumption A2 (exchangeability). The random variables $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^{n+1}$ are exchangeable, i.e., their distribution does not change when we permute them. We denote $\mathcal{Q}^{\text{exch}(n+1)} = \{Q^{\text{exch}(n+1)}, Q \in \mathcal{Q}\}$ the set of distributions of exchangeable random variables, with marginal distribution in \mathcal{Q} .

An i.i.d. sequence is a fortiori exchangeable, while the reverse is not true (for example, sampling without replacement leads to a sequence that is exchangeable but not i.i.d.).

Remark 8.1.2. We thus have that for any \mathcal{Q} , $\mathcal{Q}^{\otimes(n+1)} \subset \mathcal{Q}^{\text{exch}(n+1)}$.

Predictive uncertainty quantification under missing covariates. When features are missing, Equation (8.1) extends with \mathcal{C}_α a function of two arguments: X and M . Specifically, \mathcal{C}_α is a *marginally valid* predictive set for the test response Y given its corresponding covariates X and the mask M if:

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(X, M)) \geq 1 - \alpha. \quad (\text{MV})$$

However, marginal validity (MV) is not enough from an equity stand point and might result in discriminating between observations depending on their missing pattern (Zaffran et al., 2023). Indeed, missing values create heteroskedasticity in the resulting distribution of Y given $X_{\text{obs}(M)}$. Therefore, they argue that when facing missing values one should aim at *mask-conditional-validity* (MCV) even in the MCAR setting, i.e.:

$$\mathbb{P}(Y \in \mathcal{C}_\alpha(X, M) | M) \geq 1 - \alpha. \quad (\text{MCV})$$

Equation (MCV) is similar in spirit and motivation than Equation (8.2) but on a discrete space. Hence the impossibility results on X -conditional coverage do not hold anymore. However, (MCV) is a challenging goal as it requires the coverage to be controlled on *any* mask $m \in \mathcal{M}$, even those rarely observed at train time.

In the sequel, to highlight the underlying dependencies and randomness, any estimator of $\mathcal{C}_\alpha(\cdot, \cdot)$ fitted on a data set $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^n$ is denoted as $\hat{\mathcal{C}}_{n,\alpha}(\cdot, \cdot)$. We call a *method* a function that, for any $\alpha \in [0, 1]$, takes as input $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^n$ and outputs an estimator $\hat{\mathcal{C}}_{n,\alpha}(\cdot, \cdot)$. Table 8.1 reminds the notations.

Name	Definition	Comment
$\#A$	Cardinal of the set A	
$\mathcal{P}(A)$	Power set of A	
d	Number of features	
\mathcal{X}	Features space	$\mathcal{X} \subseteq \mathbb{R}^d$
\mathcal{Y}	Label space	
\mathcal{M}	Missing values pattern space	$\mathcal{M} \subseteq \{0, 1\}^d$
NA	Not Available (or missing value)	
$\text{obs}(m)$	Indices of the observed components for mask $m \in \mathcal{M}$ (there are $ \text{obs}(m) := \sum_{i=1}^d m_i$ of them)	$\text{obs}(m) \in \mathbb{N}^{ \text{obs}(m) }$
$\text{mis}(m)$	Indices of the missing components for mask $m \in \mathcal{M}$ (there are $ \text{mis}(m) := d - \sum_{i=1}^d m_i$ of them)	$\text{mis}(m) \in \mathbb{N}^{ \text{mis}(m) }$
\mathcal{P}	Set of distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$	
\mathcal{P}_{MAR}	Set of distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$ such that X is Missing At Random	
$\mathcal{P}_{\text{MCAR}}$	Set of distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$ such that X is Missing Completely At Random	
$\mathcal{P}_{\text{YIM} X}$	Set of distributions on $(\mathcal{X}, \mathcal{M}, \mathcal{Y})$ such that $Y \perp\!\!\!\perp M X$	
n	Number of training observations	$n + 1$ is the test index
$P^{\otimes(n+1)}$	Product distribution of P with itself $n + 1$ times (i.e., distribution of $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^{n+1}$ drawn i.i.d. with marginal P)	$P \in \mathcal{P}$
$Q^{\otimes(n+1)}$	$\{Q^{\otimes(n+1)}, Q \in \mathcal{Q}\}$	$Q \subseteq \mathcal{P}$
$P^{\text{exch}(n+1)}$	Exchangeable distribution of $n + 1$ random variables of distribution P	$P \in \mathcal{P}$
$Q^{\text{exch}(n+1)}$	$\{Q^{\text{exch}(n+1)}, Q \in \mathcal{Q}\}$	$Q \subseteq \mathcal{P}$
α	Miscoverage rate	$\alpha \in [0, 1]$
$\mathcal{C}_\alpha(\cdot, \cdot)$	Predictive set function aiming at $1 - \alpha$ coverage	$\mathcal{C}_\alpha : \mathcal{X} \times \mathcal{M} \rightarrow \mathcal{P}(\mathcal{Y})$
$\hat{C}_{n,\alpha}(\cdot, \cdot)$	Estimator for $\mathcal{C}_\alpha(\cdot, \cdot)$ based on $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^n$, through a <i>method</i>	
MV	Marginal validity	
MCV	Mask-conditional-validity	

Table 8.1: Summary of notations.

8.1.1 Literature’s background

Very recent papers have investigated uncertainty quantification with missing values. Both [Gui et al. \(2023a\)](#) and [Shao and Zhang \(2023\)](#) consider the question of distribution-free uncertainty quantification for matrix completion tasks. While the former considers building predictive sets for all of the missing entries, the latter focuses on what they call *matrix prediction* where predictive sets are required only for the last “individual” of the data set. [Seedat et al. \(2023\)](#) addresses the related but distinct problem of missing values in the responses, which is generally known as the semi-supervised setting. They introduce a self-supervised learning approach for incorporating unlabeled training data into the conformalization process. In the same framework, [Lee et al. \(2024\)](#) leverages tools from the causal inference literature to achieve stronger guarantees such as feature and outcome’s missingness conditional coverage, which are, in spirit, close to our focus (yet in a different framework).

Closer to our work of predictive uncertainty quantification under missing covariates is [Zaffran et al. \(2023\)](#), as they focus on the same setting (i.e., to predict Y given X , where X might suffer from missing values both at train time and test time). After showing that *impute-then-predict+conformalization* is marginally valid (MV) for any missing mechanism and imputation, they introduce the harder goal of *mask-conditional-validity* (MCV), motivated by an illustration on the heteroskedasticity generated by the missing values on a Gaussian Linear Model. They present a novel methodology, *Missing Data*

Augmentation (MDA), which combines with conformal prediction (CP, [Vovk et al., 2005](#)) in order to produce MCV sets. CP-MDA includes two algorithms, CP-MDA-Exact and CP-MDA-Nested, the former requiring a strict subsampling step on the training set, while the latter allows to keep the whole training data, which in turns induce large predictive sets. [Zaffran et al. \(2023\)](#) provide theoretical guarantees on the MCV of CP-MDA-Exact and on a technical minor modification of CP-MDA-Nested, under MCAR and $Y \perp\!\!\!\perp M \mid X$ assumptions.

8.1.2 Overview of our contributions (and outline)

In short, our objective is to tackle the following question: **when and how is it possible to achieve MCV?** Notably, we are interested in understanding *i*) what assumptions are necessary to ensure MCV, *ii*) how to design a tailored methodology, and *iii*) what happens when these assumptions are not satisfied.

We start by proving hardness results on distribution-free MCV in Section 8.2. Notably, for a MCV method outputting $\hat{C}_{n,\alpha}(\cdot, \cdot)$ with no assumption except from having access to n i.i.d. draws, we prove that the predictive interval is most often uninformative: for any $m \in \mathcal{M}$ the probability that, say, $\hat{C}_{n,\alpha}(\cdot, m) \equiv \mathcal{Y}$ is higher than $1 - \alpha - \Delta_{m,n}$, where $\Delta_{m,n}$ gets negligible when the mask m is nearly not observed in a sample of size n . In other words, a method that is distribution-free MCV will output uninformative intervals on any mask that does not represent a high enough proportion of the training data. We go further and show that the exact same trade-off still holds for a method that is MCV only on distributions that are MAR, or MCAR, or similarly on distributions such that $Y \perp\!\!\!\perp M \mid X$, i.e., restricting an algorithm to be MCV only when $Y \perp\!\!\!\perp M \mid X$ would still output uninformative sets on rarely observed masks: it is necessary to add another assumption on the dependence between X and M (such as MCAR) to allow for informative MCV on any mask. Importantly, this theoretical analysis brings new insights on the achievability of X -group-conditional validity, beyond MCV¹.

This motivates the investigation of the quantile regression and missing values interplay in Section 8.3, so as to provide guidelines for practical design of probabilistic prediction with missing covariates. This interplay is hard to characterize in general but becomes explicit under missingness assumptions², or a multivariate Gaussian setting or linear model. Our key findings are (*i*—Section 8.3.1) that the uncertainty often increases with more missing values: we analyze different mathematical statements of this main idea (in terms of conditional variance, inter-quantile distance, or predictive interval length) and evaluate theoretically under which distributional assumptions they are satisfied, in particular under MCAR and $Y \perp\!\!\!\perp M \mid X$, motivating our methodological design of Section 8.4; (*ii*—Section 8.3.2) if the goal is to estimate quantiles, it is essential to be able to retrieve the mask to construct intervals, in contrast to classic mean regression where the mask is not as crucial.

¹Precisely, we provide a rigorous quantification of Vladimir Vovk’s comment on X -conditional validity: “of course, the condition that x be a non-atom is essential: if $P_X(x) > 0$, an inductive conformal predictor that ignores all examples with objects different from x will have $1 - \alpha$ object conditional validity and can give narrow predictions if the training set is big enough to contain many examples with x as their object” rephrased from [Vovk \(2012\)](#) to match our notations.

In Section 8.4, we propose a unified framework, **CP-MDA-Nested***, building predictive sets with missing covariates for both regression and classification tasks. Precisely, it bridges the gap between CP-MDA-Exact and CP-MDA-Nested introduced in Zaffran et al. (2023), by encapsulating these two algorithms as well as any in between with more flexible subsampling schemes, allowing to fix the trade-off between coverage variance (CP-MDA-Exact) and overly conservative predictive sets (CP-MDA-Nested). Leveraging the similarity between **CP-MDA-Nested*** and leave-one-out conformal approaches (Vovk, 2015; Barber et al., 2021b; Gupta et al., 2022) we provide theory on the marginal validity of **CP-MDA-Nested*** without subsampling, which holds regardless of the missingness distribution (without any assumption on the dependence between M and X , but also without any assumption on the relationship between M and Y conditionally on X). Moreover, we also establish that **CP-MDA-Nested*** is MCV for a wide range of subsampling schemes under MCAR and $Y \perp\!\!\!\perp M \mid X$.

Finally, in Section 8.5 we conduct synthetic experiments beyond the MCAR and $Y \perp\!\!\!\perp M \mid X$ assumptions. Precisely, we generate missingness that is either MAR (5 different settings), MNAR (11 different settings) or such that $Y \not\perp\!\!\!\perp M \mid X$. **CP-MDA-Nested*** empirically maintains MCV under MAR and MNAR missingness. When $Y \perp\!\!\!\perp M \mid X$ is not satisfied, **CP-MDA-Nested*** does not ensure MCV experimentally, unless the imputation is accurate enough. Overall, these numerical experiments showcase the robustness of **CP-MDA-Nested*** beyond its theoretical scope of validity.

In the following Table 8.2, we summarize and organize our main contributions. We report the theoretical results on the possibility to achieve informative MCV, either positive results (✓) or negative hardness results (✗), along with our more general result on marginal validity. Moreover, we locate experimental results by indicating the figures that align with particular setups. In particular, we distinguish two kinds of experiments: *Numerical extension* of results beyond the conditions where the theory is applicable, which demonstrates promising outcomes in more challenging configurations, and *Numerical confirmation* of results anticipated by theoretical analysis, that is the outcomes of the experiment either *i*) were already expected based on the theory or *ii*) confirm that the theoretical assumptions can not be relaxed to the corresponding distributional setting.

	$\mathcal{P}_{\text{MCAR}}$	\mathcal{P}_{MAR}	$\mathcal{P}_{\text{MNAR}} = \mathcal{P}$	
$\mathcal{P}_{Y \perp\!\!\!\perp M \mid X}$	CP-MDA-Nested*: ✓ <i>Theorem 8.4.2</i>	?	Hardness: ✗ <i>Proposition 8.2.2</i>	<i>Theory</i>
	Figures 8.5a and 8.5b		Figures 8.6a, 8.6b, 8.7a and 8.7b	<i>Num. extension</i>
	Figure 8.4		Remark 8.5.1	<i>Num. confirmation</i>
\mathcal{P}	Hardness: ✗ <i>Proposition 8.2.1</i>	Hardness: ✗ <i>Proposition 8.2.1</i>	Hardness: ✗ <i>Theorem 8.2.1</i> CP-MDA-Nested*: MV <i>Theorem 8.4.1</i>	<i>Theory</i>
	Figure 8.8a			<i>Num. extension</i>
	Figure 8.8b	Remark 8.5.1	Remark 8.5.1	<i>Num. confirmation</i>

Table 8.2: Summary of the main theoretical results.

8.2 When is Mask-Conditional-Validity (MCV) a too lofty goal?

We will show in this section that purely distribution-free MCV guarantees are often uninformative. As a consequence, we will have to impose some non-parametric assumption on the underlying data distribution. We thus have to define the concept of MCV with respect to a class of distributions \mathcal{D} (MCV- \mathcal{D}), and to study the sets \mathcal{D} that allow for informative MCV- \mathcal{D} .

Definition 8.2.1 (MCV- \mathcal{D}). Let \mathcal{D} be a set of distributions on $(\mathcal{X} \times \mathcal{M} \times \mathcal{Y})^{n+1}$. A method outputting $\hat{C}_{n,\alpha}(\cdot, \cdot)$ based on $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^n$ for any $\alpha \in [0, 1]$ is MCV- \mathcal{D} if for any distribution $D \in \mathcal{D}$ and any $\alpha \in [0, 1]$, we have:

$$\mathbb{P}_D \left(Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, M^{(n+1)} \right) \mid M^{(n+1)} \right) \stackrel{a.s.}{\geq} 1 - \alpha,$$

i.e., for any $m \in \mathcal{M}$ such that $\mathbb{P}(M^{(n+1)} = m) > 0$, it holds:

$$\mathbb{P}_D \left(Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m \right) \mid M^{(n+1)} = m \right) \geq 1 - \alpha.$$

If $\mathcal{D} = \mathcal{P}^{\text{exch}(n+1)}$ we recover the holy grail of being MCV for any exchangeable distribution, i.e., the most distribution-free result we could target. If \mathcal{D} is not specified thereon, it will refer to MCV- $\mathcal{P}^{\text{exch}(n+1)}$. An easier goal is to aim at MCV- $\mathcal{P}^{\otimes(n+1)}$, that is MCV on i.i.d. distributions.

Remark 8.2.1. For any sets $\mathcal{D} \subseteq \mathcal{D}'$, a method that is MCV- \mathcal{D}' is also MCV- \mathcal{D} , i.e., MCV- $\mathcal{D}' \Rightarrow$ MCV- \mathcal{D} .

A naive idea to ensure MCV is to split the data set into $\#\mathcal{M}$ sub data sets, one for each mask, and run any marginally valid method on each of the data sets independently. However, as $\#\mathcal{M}$ grows exponentially in the dimension, this is not practical as it will generate small (or even empty) data sets for some masks. In particular, as long as $\mathbb{P}(M = m)$ is low with respect to n for a given $m \in \mathcal{M}$, estimation on the sub data set is hard, and even finite sample method such as conformal prediction (Vovk et al., 2005) will suffer from important statistical variability or uninformativeness. Therefore, in practice, we usually need to go beyond this solution if we aim to achieve MCV for any mask, even those rarely observed at train times. Nevertheless, the task appears challenging without restricting the link between M and (X, Y) , precisely due to the lack of information available in the data set. The question we tackle in this section is the following: **is it possible to achieve distribution-free MCV in an informative way for any mask in \mathcal{M} , even those occurring with low probability?**

Link with group conditional coverage. More generally, the question is that of finding on which subspace of the features it is possible to obtain meaningful conditional guarantees. Thus, the results demonstrated in this section give some answers to the broader question of when is *group-feature-conditional validity* achievable (a relaxation of Equation (8.2)), which has attracted considerable interest lately (see e.g., Romano et al., 2020a; Barber et al., 2021a; Guan, 2022; Jung et al., 2023; Gibbs and Candès, 2023, to name just a few).

Our hardness results shed light on X -group-conditional coverage.

In the rest of this section, M can be interpreted as any additional random variable, that may (or may not) depend on X , on which we aim at achieving distribution-free conditional validity. For example, \mathcal{M} could represent subgroups of \mathcal{X} , eventually overlapping. Specifically, assume $\mathcal{M} = \{0, 1\}^{|\mathcal{G}|}$ for \mathcal{G} a collection of groups on \mathcal{X} , then M is an indicator vector on whether X belongs to each group of \mathcal{G} or not.

A particular case of this generalization is $\mathcal{G} = \left\{ \{X \in \mathcal{X} : X_j \text{ is missing}\}_{j=1}^d \right\}$, recovering our missing covariates setting with M the missing pattern. While our discussion in this section is written towards the missing covariates setting, the interested reader might replace “missing pattern” or “mask” by “groups” whenever it makes sense², and the corresponding result will hold without further restriction or assumptions on the way the groups are designed.

8.2.1 Fully distribution-free result

Our first result, Theorem 8.2.1, confirms the previous intuition: any $\text{MCV-}\mathcal{P}^{\otimes(n+1)}$ method typically does output the whole set \mathcal{Y} with high probability for any distribution, on low probability masks.

Theorem 8.2.1 (Trade-off set size and mask probability). *Suppose that a method outputting $\hat{C}_{n,\alpha}$ is $\text{MCV-}\mathcal{P}^{\otimes(n+1)}$. Then for any $P \in \mathcal{P}$ and any $m \in \mathcal{M}$ such that $P_M(m) > 0$, it holds:*

$$\begin{cases} \text{if } \mathcal{Y} \subseteq \mathbb{R} \text{ (regression)} : \mathbb{P}_{P^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha}(X, m) \right) = \infty \right) \geq 1 - \alpha - \Delta_{m,n}, \\ \text{if } \mathcal{Y} \subseteq \mathbb{N} \text{ (classification)} : \forall y \in \mathcal{Y}, \mathbb{P}_{P^{\otimes(n+1)}} \left(y \in \hat{C}_{n,\alpha}(X, m) \right) \geq 1 - \alpha - \Delta_{m,n}, \end{cases}$$

$$\text{with } \Delta_{m,n} := \sqrt{2 \left(1 - \left(1 - \frac{P_M(m)^2}{2} \right)^{n+1} \right)}.$$

Since for any $x > 0$ and $n \in \mathbb{N}^*$, it holds $1 - (1 - x)^n < nx$, Theorem 8.2.1 implies that:

$$\begin{cases} \text{if } \mathcal{Y} \subseteq \mathbb{R} \text{ (regression)} : \mathbb{P}_{P^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha}(X, m) \right) = \infty \right) \geq 1 - \alpha - P_M(m) \sqrt{(n+1)}, \\ \text{if } \mathcal{Y} \subseteq \mathbb{N} \text{ (classification)} : \forall y \in \mathcal{Y}, \mathbb{P}_{P^{\otimes(n+1)}} \left(y \in \hat{C}_{n,\alpha}(X, m) \right) \geq 1 - \alpha - P_M(m) \sqrt{(n+1)}. \end{cases}$$

Theorem 8.2.1 provides a lower bound on the probability that the predictive set is uninformative for any $m \in \mathcal{M}$ (i.e., typically $\Lambda(\hat{C}_{n,\alpha}(\cdot, m)) = \infty$ or $\#\hat{C}_{n,\alpha}(\cdot, m) \geq \#\mathcal{Y}(1 - \alpha)$).

Remark 8.2.2 ($\text{MCV-}\mathcal{P}^{\otimes(n+1)}$ implies uninformative sets even on simple distributions). Crucially, this lower bound holds for *any* distribution in \mathcal{P} . This implies that the hardness

²The only result that does not extend is Proposition 8.2.1 for \mathcal{P}_{MAR} , as by construction it relies on the missingness structure.

result applies also to smooth, nonpathological, distributions. Particularly, it means that any method that is fully distribution-free MCV (i.e., $\text{MCV-}\mathcal{P}^{\otimes(n+1)}$) will be subject to the lower bound even when applied to data whose actual distribution is as simple as possible (e.g., MCAR and $Y \perp M | X$).

Remark 8.2.3 (Informative sets implies the method is not $\text{MCV-}\mathcal{P}^{\otimes(n+1)}$). Conversely, for a given method constructing predictive sets $\hat{C}_{n,\alpha}$, assume that there exists a distribution $P \in \mathcal{P}$ and a mask m such that $P_M(m) > 0$ and $\Delta_{m,n} < \frac{1-\alpha}{2}$ and under which $\hat{C}_{n,\alpha}$ is consistently of finite measure for different random draws from $P^{\otimes(n+1)}$. Then, this method is not $\text{MCV-}\mathcal{P}^{\otimes(n+1)}$, as otherwise under $P^{\otimes(n+1)}$ the predictive set would be of infinite measure with probability at least 0.25 for $\alpha \leq 0.5$ according to Theorem 8.2.1 (since $1 - \alpha - \Delta_{m,n} \geq \frac{1-\alpha}{2} \geq 0.25$).

Interpretation of the lower bound. Let us now decompose the lower bound. The first term, $1 - \alpha$, is an “irreducible term”. Indeed, the estimator outputting \mathcal{Y} with probability $1 - \alpha$ and the empty set \emptyset with probability α (where the probability corresponds to an exogenous Bernoulli random variable) is valid conditionally on everything, thus a fortiori on M . Hence, the lower bound has to be smaller than $1 - \alpha$ as the set of MCV estimators includes this naive one.

For a given distribution P , the second term, $\Delta_{m,n}$, becomes negligible on any $m \in \mathcal{M}$ such that $P_M(m)$ is small with respect to n , making the lower bound be nearly $1 - \alpha$. This reflects the intuition that it is impossible to achieve informative conditional coverage when conditioning on events whose effective sample size is limited. In other words, the smaller the probability of the event occurring, the larger the training size must be to compensate and make “sure” that enough observations are drawn from that event.

Note that as $\mathcal{P}^{\otimes(n+1)} \subset \mathcal{P}^{\text{exch}(n+1)}$, any $\text{MCV-}\mathcal{P}^{\text{exch}(n+1)}$ estimator is $\text{MCV-}\mathcal{P}^{\otimes(n+1)}$ by Remark 8.2.1. Thus, the conclusion of Theorem 8.2.1 extends to any $\text{MCV-}\mathcal{P}^{\text{exch}(n+1)}$ estimator, on any $P^{\otimes(n+1)}$ with $P \in \mathcal{P}$.³

Proof sketch. For any given distribution $P \in \mathcal{P}$, and a given mask $m \in \mathcal{M}$ such that $P_M(m) > 0$, the idea of the proof is the following. Build another distribution $Q \in \mathcal{P}$, which equals P whenever $M \neq m$, and that “admits” an arbitrary spread on Y when $M = m$ (in short, Q is meant to be pathological yet close to P). By doing so, two statements can be made. First, $Q^{\otimes(n+1)}$ belongs to $\mathcal{P}^{\otimes(n+1)}$, therefore, as $\hat{C}_{n,\alpha}$ is $\text{MCV-}\mathcal{P}^{\otimes(n+1)}$, under $Q^{\otimes(n+1)}$ the probability of $\hat{C}_{n,\alpha}$ being uninformative is $1 - \alpha$ since Y can typically be anywhere. Second, as P and Q are the same everywhere except on $\{M = m\}$, the total variation distance between them is smaller than $P_M(m)$. This leads to the total variation distance between $P^{\otimes(n+1)}$ and $Q^{\otimes(n+1)}$ being smaller than $\Delta_{m,n}$. Combining these two observations, it finally leads to the probability of $\hat{C}_{n,\alpha}$ being uninformative under $P^{\otimes(n+1)}$ which is greater than $1 - \alpha - \Delta_{m,n}$. The complete proof is given in Section 8.A.

A familiar reader will note the similarity with the proofs given by [Lei and Wasserman \(2014\)](#); [Vovk \(2012\)](#). The difference is that, on the one hand, [Vovk \(2012\)](#) proof leverages

³The same is true for the subsequent Proposition 8.2.1 and Proposition 8.2.2.

an “reductio ad absurdum” that does not allow to explicitly build the set on which $P \neq Q$. On the other hand, [Lei and Wasserman \(2014\)](#) is constructive. Nonetheless, it relies on a crucial step that implicitly assumes that conditional-validity holds conditionally on the n data points, leading to an inexact statement: the lower bound obtained becomes 1. As we discussed, as well as [Vovk \(2012\)](#), the lower bound can not be bigger than $1 - \alpha$. We provide an alternate proof to this well-known X -conditional impossibility result that is constructive. Another improvement is that our expression of $\Delta_{m,n}$ comes from a tighter inequality than the ones used in [Lei and Wasserman \(2014\)](#) and [Vovk \(2012\)](#). Indeed, for the original impossibility result, the lower bound does not really matter as we then take its limit when the ball around x_0 shrinks, which is 0. But in our case, this ball is fixed to the event $\{M = m\}$.

8.2.2 Restricting the class of admissible missingness distributions

Interestingly, the proof of Theorem 8.2.1 adapts to $\text{MCV-}\mathcal{P}_{\text{MAR}}^{\otimes(n+1)}$ or $\text{MCV-}\mathcal{P}_{\text{MCAR}}^{\otimes(n+1)}$.

Proposition 8.2.1 (Trade-off set size and mask probability on \mathcal{P}_{MAR} or $\mathcal{P}_{\text{MCAR}}$). *Let \mathcal{Q} be either \mathcal{P}_{MAR} or $\mathcal{P}_{\text{MCAR}}$. Suppose that an estimator $\hat{C}_{n,\alpha}$ is $\text{MCV-}\mathcal{Q}^{\otimes(n+1)}$ at the level α . Then for any $Q \in \mathcal{Q}$ and any $m \in \mathcal{M}$ such that $Q_M(m) > 0$, it holds:*

$$\begin{cases} \text{if } \mathcal{Y} \subseteq \mathbb{R} \text{ (regression)} : \mathbb{P}_{Q^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha}(X, m) \right) = \infty \right) \geq 1 - \alpha - \Delta_{m,n}, \\ \text{if } \mathcal{Y} \subseteq \mathbb{N} \text{ (classification)} : \forall y \in \mathcal{Y}, \mathbb{P}_{Q^{\otimes(n+1)}} \left(y \in \hat{C}_{n,\alpha}(X, m) \right) \geq 1 - \alpha - \Delta_{m,n}, \end{cases}$$

with $\Delta_{m,n}$ given in Theorem 8.2.1.

Remark 8.2.4 (no direct implication between results). Proposition 8.2.1 for $\mathcal{Q} = \mathcal{P}_{\text{MAR}}$ does not imply Proposition 8.2.1 for $\mathcal{Q} = \mathcal{P}_{\text{MCAR}}$, nor the contrary. Indeed, on the one hand, as $\mathcal{P}_{\text{MCAR}}^{\otimes(n+1)} \subseteq \mathcal{P}_{\text{MAR}}^{\otimes(n+1)}$, any method that is $\text{MCV-}\mathcal{P}_{\text{MAR}}^{\otimes(n+1)}$ is $\text{MCV-}\mathcal{P}_{\text{MCAR}}^{\otimes(n+1)}$ (Remark 8.2.1). However, on the other hand, Proposition 8.2.1 (or Theorem 8.2.1) provides a uniform statement over $Q \in \mathcal{Q}$ (Remark 8.2.2): as $\mathcal{P}_{\text{MCAR}}^{\otimes(n+1)} \subseteq \mathcal{P}_{\text{MAR}}^{\otimes(n+1)}$, the final statement holds on more distributions for $\mathcal{Q} = \mathcal{P}_{\text{MAR}}$ than for $\mathcal{Q} = \mathcal{P}_{\text{MCAR}}$. Therefore, Proposition 8.2.1 for $\mathcal{Q} = \mathcal{P}_{\text{MAR}}$ provides a *stronger statement over fewer methods* than Proposition 8.2.1 for $\mathcal{Q} = \mathcal{P}_{\text{MCAR}}$.

For the same reason, Proposition 8.2.1 is not deduced directly from Theorem 8.2.1, but from a careful consideration of the construction in its proof: the adversarial distribution built therein does not make any assumption on the relationship between X and M , which can be as simple as desired.

In fact, the key point for the proof of Theorem 8.2.1 is that the algorithm achieves MCV also on distributions under which Y and M can be dependent even conditionally on X : thus, it allows us to construct an adversarial distribution under which Y is equally likely to be anywhere on the label space for a given $m \in \mathcal{M}$.

In view of this, one could think that in order to break Theorem 8.2.1, and therefore to ensure that MCV is achievable in an informative way even on low probability masks, we

have to *at least* assume $Y \perp\!\!\!\perp M \mid X$ (A1). However, in Proposition 8.2.2, we show that even estimators that are only $\text{MCV-}\mathcal{P}_{\text{YIM} \mid X}^{\otimes(n+1)}$ suffer from the same trade-off on efficiency.

Proposition 8.2.2 (Trade-off set size and mask probability on $\mathcal{P}_{\text{YIM} \mid X}$). *Suppose that an estimator $\hat{C}_{n,\alpha}$ is $\text{MCV-}\mathcal{P}_{\text{YIM} \mid X}^{\otimes(n+1)}$ at the level α . Then for any $P \in \mathcal{P}_{\text{YIM} \mid X}$ and for any $m \in \mathcal{M}$ such that $\frac{1}{\sqrt{2}} \geq P_M(m) > 0$, it holds:*

$$\begin{cases} \text{if } \mathcal{Y} \subseteq \mathbb{R} \text{ (regression)} : \mathbb{P}_{P^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha}(X, m) \right) = \infty \right) \geq 1 - \alpha - \Delta_{m,n}, \\ \text{if } \mathcal{Y} \subseteq \mathbb{N} \text{ (classification)} : \forall y \in \mathcal{Y}, \mathbb{P}_{P^{\otimes(n+1)}} \left(y \in \hat{C}_{n,\alpha}(X, m) \right) \geq 1 - \alpha - \Delta_{m,n}, \end{cases}$$

with $\Delta_{m,n} := \sqrt{2 \left(1 - (1 - 2P_M(m)^2)^{n+1} \right)}$.

All in all, Proposition 8.2.2 demonstrates that even the simplest relationship between Y and M does not allow informative predictive sets. This reveals that to ensure that it is possible to obtain informative sets even on low probability masks (or events), one has to design a method that will be conditionally valid *only* on distributions with a constrained structure of dependence between Y and M given X , but also between M and X . In particular, trying to ensure $\text{MCV-}\mathcal{P}_{\text{MCAR}, \text{YIM} \mid X}^{\otimes(n+1)}$ (where $\mathcal{P}_{\text{MCAR}, \text{YIM} \mid X}^{\otimes(n+1)} := \mathcal{P}_{\text{MCAR}}^{\otimes(n+1)} \cap \mathcal{P}_{\text{YIM} \mid X}^{\otimes(n+1)}$) as done in Zaffran et al. (2023) appears as a natural way to approach the minimal set of assumptions.

Remark 8.2.5. In Figure 8.4, we illustrate that, on a distribution $P \in \mathcal{P}_{\text{MCAR}, \text{YIM} \mid X}^{\otimes(n+1)}$, a provably $\text{MCV-}\mathcal{P}_{\text{MCAR}, \text{YIM} \mid X}^{\otimes(n+1)}$ method (introduced in Section 8.4) consistently outputs finite length predictive intervals (regression case). Therefore, we can conclude that obtaining a hardness result on $\mathcal{P}_{\text{MCAR}, \text{YIM} \mid X}^{\otimes(n+1)}$ appears impossible, as such it would induce Remark 8.2.3 (with $\mathcal{P}_{\text{MCAR}, \text{YIM} \mid X}^{\otimes(n+1)}$ instead of $\mathcal{P}^{\otimes(n+1)}$).

8.3 Links between missing covariates and predictive uncertainty

In light of the previous section, MCV appears hard to achieve. Thus, the problem that we aim to address now is to **find ways to model properly the missing covariates' influence on predictive uncertainty**. To understand the relationship between missing values and predictive uncertainty, this section explores simplified distributions on (X, M, Y) —such as MCAR and $Y \perp\!\!\!\perp M \mid X$ —and/or on (X, Y) —such as linearity, Gaussianity. We consider the regression case with $\mathcal{Y} = \mathbb{R}$. This exploration aims to facilitate the development of suitable frameworks for probabilistic inference when covariates are missing—i.e., models that are as close as possible to achieving MCV.

8.3.1 Increasing uncertainty with nested masks

The hardness results of Section 8.2 induce that MCV cannot be (efficiently) achieved without structural assumptions on the links between the predictive distributions conditional

on each missing pattern. In this subsection, we gain insights into the underlying reasons for this phenomenon: the predictive uncertainty depends on the missing pattern, a form of *heteroskedasticity*. In summary, we explore the following idea, which is a natural modelization attempt in that direction:

Idea: *The predictive uncertainty increases when less covariates are observed.*

In technical words, the aforementioned heteroskedasticity takes the form of an *isotonicity* (monotony) with respect to the mask, with the inclusion order given by Definition 8.3.1 below. In short: the more missing values, the more uncertainty there is.

Definition 8.3.1 (Included masks). Let $(m, m') \in \mathcal{M}^2$, $m \subset m'$ if for any $j \in \llbracket 1, d \rrbracket$ such that $m_j = 1$ then $m'_j = 1$, i.e., m' includes at least the same missing values than m .

Hereafter, we formally quantify such a statement, in particular in terms of conditional variance, inter-quantile distance, and predictive interval length. We demonstrate that some of those statements are valid, to different extent, under distributional assumptions, either generic or on specific model or examples. To that end, we introduce several properties, that can be considered as non-parametric assumptions on the underlying distributions. We put together some results of this section in the following Table 8.3, that can be used as a reading guide throughout the section.

8.3.1.1 CONDITIONAL VARIANCE ISOTONY W.R.T. THE MISSING DATA PATTERNS

We start by focusing on the link between M and the *conditional variance* of $Y|X_{\text{obs}(M)}$, that constitutes a natural proxy on the predictive uncertainty. Denote $V(X_{\text{obs}(M)}, M) := \text{Var}(Y|X_{\text{obs}(M)}, M)$ the conditional variance of Y given $(X_{\text{obs}(M)}, M)$. We introduce two properties regarding its ordering with respect to M : (Var-1) and (Var-2).

$$V(X_{\text{obs}(m)}, m) \stackrel{a.s.}{\leq} V(X_{\text{obs}(m')}, m') \quad \text{for any } m \subset m', \quad (\text{Var-1})$$

$$\mathbb{E}[V(X_{\text{obs}(M)}, M)|M = m] \leq \mathbb{E}[V(X_{\text{obs}(M)}, M)|M = m'] \quad \text{for any } m \subset m'. \quad (\text{Var-2})$$

Property Var-1 is stronger than Property Var-2 as it is an almost sure result w.r.t. the covariates X . The following proposition ensures that (Var-2) is satisfied under $\mathcal{P}_{\text{MCAR}, \text{YIM} | X}$ (that is, assumptions for which no hardness result can exist).

Proposition 8.3.1. *Under $\mathcal{P}_{\text{MCAR}, \text{YIM} | X}$, (Var-2) is valid.*

Property \ Setup	Setup			$\mathcal{P}_{\text{MCAR}, \text{YIM} X}$
	Model 8.3.2	Model 8.3.1		
Variance	Var-1	Var-1 Var-2		Var-2
Inter-quantile	IQ-1	IQ-2		
Length of Oracle PI	Len-1	Len-2		Len-2

Table 8.3: Summary of the results from Section 8.3.1.

The proof of this result is given in Section 8.B.1. This is a first significant result: under general assumptions—i.e., strong assumption on the relation between the mask and both the response and the features, but no assumptions on their distribution—, the averaged variance is always smaller on smaller masks. This establishes the existence of a link between the uncertainties *on patterns that can be compared*, that is patterns that are nested in one another. Note that the order given by Definition 8.3.1 is only a partial order: the average variance ordering is only enforced w.r.t. that partial order.

It is possible that the predictive uncertainty increases on average with the mask (Equation (Var-2)) but not almost surely on X (Equation (Var-1)), as illustrated by Model 8.3.1 below:

Model 8.3.1 (Unidimensional heteroskedasticity). Consider the following one-dimensional model:

- $X \sim \mathcal{N}(0, \sigma^2)$, $\sigma \in \mathbb{R}_+$;
- $\xi \sim \mathcal{N}(0, \tau^2)$, $\tau \in \mathbb{R}_+$, such that $\xi \perp X$;
- $Y = \beta X + X\xi$, with $\beta \in \mathbb{R}$;
- $M \sim \mathcal{B}(\rho)$, with $\rho \in [0, 1]$, and $M \perp (X, Y)$.

Under this model, we obtain that $M \perp X$ (MCAR) and $Y \perp M | X$, and

$$\begin{cases} \text{Var}(Y|X, M=0) = \tau^2 X^2 \\ \text{Var}(Y|M=1) = (\beta^2 + \tau^2)\sigma^2 \end{cases} \Rightarrow \begin{cases} \mathbb{E}[\text{Var}(Y|X, M=0)] = \tau^2 \sigma^2 \\ \mathbb{E}[\text{Var}(Y|M=1)] = (\beta^2 + \tau^2)\sigma^2 \end{cases}.$$

Thus Equation (Var-2) is verified but Equation (Var-1) is only satisfied for X such that $X^2 \leq \left(1 + \frac{\beta^2}{\tau^2}\right)\sigma^2$. This is illustrated in Figure 8.1. The first subplot represents Y depending on X , while the third subplot displays $Y - \beta X$ depending on X , that is an illustration of the uncertainty of the distribution of $Y|X$. For any X outside the vertical dashed lines (corresponding to $\pm(1 + \beta^2/\tau^2)\sigma^2$), the conditional variance of Y given X is larger than the overall variance when X is missing. Yet, the average variance of Y when X is missing is indeed higher than the average variance of Y when X is observed: this can be seen on the two histograms on subplots 2 and 4.

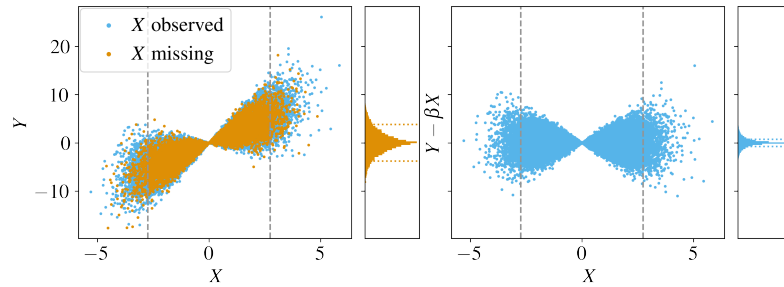


Figure 8.1: Visualisation of a random draw from the data distribution of Model 8.3.1, with 100000 i.i.d. samples, $\rho = 0.2$, $\sigma^2 = 1.5$, $\tau^2 = 1$ and $\beta = 2$. The colors indicate whether X is observed or missing. The first subplot represents Y depending on X , while the third subplot displays $Y - \beta X$ depending on X only for observed X , that is an illustration of the uncertainty of $Y|X$. The second subplot is an histogram of Y when X is missing, while the fourth subplot is an histogram of $Y - \beta X$ when X is observed, i.e., they represent the predictive distribution of Y depending on whether X is observed or missing.

Finally, while Model 8.3.1 shows that (Var-1) is not always true, even under the assumptions of Proposition 8.3.1, we now show that it can be achieved in the following Gaussian linear model, a particular case of Gaussian pattern mixture model.

Model 8.3.2 (Gaussian linear model (GLM)). The data is generated according to a linear model and the covariates are Gaussian conditionally to the pattern:

- $Y = \beta^T X + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \perp (X, M)$, $\beta \in \mathbb{R}^d$.
- for all $m \in \mathcal{M}$, there exist $\mu^m \in \mathbb{R}^d$ and $\Sigma^m \in \mathbb{R}^{d \times d}$ such that $X|(M = m) \sim \mathcal{N}(\mu^m, \Sigma^m)$.

Such a model results in a MCAR distribution when $\Sigma^m \equiv \Sigma$. Indeed under Model 8.3.2 the resulting predictive distribution is given by $Y|(X_{\text{obs}(m)}, M = m) \sim \mathcal{N}(\tilde{\mu}^m, \tilde{\sigma}^m)$ for any $m \in \mathcal{M}$, with:

$$\begin{aligned}\tilde{\mu}^m &= \beta_{\text{obs}(m)}^T X_{\text{obs}(m)} + \beta_{\text{mis}(m)}^T \mu_{\text{mis}|\text{obs}}^m, \\ \tilde{\sigma}^m &= \beta_{\text{mis}(m)}^T \Sigma_{\text{mis}|\text{obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2,\end{aligned}$$

with $\mu_{\text{mis}|\text{obs}}^m$ and $\Sigma_{\text{mis}|\text{obs}}^m$ defined in Section 8.B.1.2 (Le Morvan et al., 2020b; Ayme et al., 2022; Zaffran et al., 2023). Crucially, $\tilde{\sigma}^m$ depends on m in a non-linear fashion, even in MCAR. That is, even in MCAR and a homoskedastic model for $Y|X$, the predictive distribution of $Y|X_{\text{obs}(M)}$ is heteroskedastic: basically, the distribution of Y is a mixture of various distributions with the mask being the latent variable. This simple example already illustrates that missing values generate strong heteroskedasticity: in Proposition 8.3.2, we show that under this Model 8.3.2 and $\mathcal{P}_{\text{MCAR}}$, the variance of the conditional distribution of Y increases when the missing pattern increases (in the sense of Definition 8.3.1).

Proposition 8.3.2 (Conditional variance increases with the mask under MCAR GLM). *Under Model 8.3.2 and $\mathcal{P}_{\text{MCAR}}$, if the covariance matrix Σ is positive definite, Equation (Var-1) is satisfied.*

To prove that the variance increases with the pattern, we prove that for any $m \subset m'$, $\Sigma_{\text{mis}|\text{obs}}^{m'} \succcurlyeq \begin{pmatrix} \Sigma_{\text{mis}|\text{obs}}^m & 0 \\ 0 & \mathbf{0} \end{pmatrix}$. This is proved in Section 8.B.1.2.

Next, in order to go beyond variances, we focus on inter-quantile distances as a measure of uncertainty, and establish a general result on the expected length of oracle predictive intervals.

8.3.1.2 CONDITIONAL INTER-QUANTILE ISOTONY W.R.T. THE MISSING DATA PATTERNS

Ideally, we would like to access the oracle predictive interval (the interval satisfying Equation (MCV) with minimal expected length). Thus, in this section we are interested in characterizing its behavior with respect to M , in order to be able to mimic it. We denote this interval $\mathcal{C}_\alpha^{*,P}$, that is formally defined for any $m \in \mathcal{M}$ as:

$$\mathcal{C}_\alpha^{*,P}(\cdot, m) := \arg \min_{\substack{\mathcal{C}_\alpha: \mathcal{X} \times \mathcal{M} \rightarrow \mathcal{P}(\mathbb{R}) \\ \text{s.t. } \mathbb{P}_P(Y \in \mathcal{C}_\alpha(X, m) | M = m) \geq 1 - \alpha,}} \mathbb{E}_P[\Lambda(\mathcal{C}_\alpha(X_{\text{obs}(m)}, m)) | M = m].$$

In fact, under Model 8.3.2, the oracle predictive interval is uniquely defined by the quantiles $\alpha/2$ and $1 - \alpha/2$ of the $\mathcal{N}(\tilde{\mu}^m, \tilde{\sigma}^m)$. More importantly, this oracle interval even achieves X -conditional coverage. Proposition 8.3.2 shows that under $\mathcal{P}_{\text{MCAR}}$ and Model 8.3.2, increasing the number of missing values (in a nested way) induces an increase in the predictive uncertainty of Y : the oracle intervals, that are given by inter-quantiles intervals, are nested. Notably, this is true almost surely on X_{obs} and not only marginally.

To generalize this property beyond the Gaussian case, we introduce the inter-quantile distance, that encodes the uncertainty for conditional predictive distribution. For all $\beta \leq \frac{1}{2} \leq \gamma$, we define the inter-quantile space for quantile distributions:

$$\text{IQ}_{\beta,\gamma}(X_{\text{obs}(M)}, M) = q_{\gamma}(\mathbb{P}_{Y|X_{\text{obs}(M)}, M}) - q_{\beta}(\mathbb{P}_{Y|X_{\text{obs}(M)}, M}).$$

And the following two assumptions, that are similar in spirit to (Var-1) and (Var-2)

$$\text{IQ}_{\beta,\gamma}(X_{\text{obs}(m)}, m) \stackrel{\text{a.s.}}{\leq} \text{IQ}_{\beta,\gamma}(X_{\text{obs}(m')}, m') \quad \text{for any } m \subset m', \quad (\text{IQ-1})$$

$$\mathbb{E} [\text{IQ}_{\beta,\gamma}(X_{\text{obs}(M)}, M) | M = m] \leq \mathbb{E} [\text{IQ}_{\beta,\gamma}(X_{\text{obs}(M)}, M) | M = m'] \quad \text{for any } m \subset m'. \quad (\text{IQ-2})$$

The assumptions on the quantiles and the variance are equivalent for Gaussian (conditional) distributions. As a consequence, (IQ-2) is satisfied under Model 8.3.2 and $\mathcal{P}_{\text{MCAR}}$ as well as under Model 8.3.1, while (IQ-1) is satisfied only under Model 8.3.2 and $\mathcal{P}_{\text{MCAR}}$. Inter-quantile assumptions are related to predictive intervals: for any distribution P such that $P_{Y|X_{\text{obs}(M)}, M}$ is a.s. unimodal, the oracle predictive interval $\mathcal{C}_{\alpha}^{*,P}$ writes as an inter-quantile interval almost surely, that is there exist functions $\beta, \gamma : \mathcal{X} \times \mathcal{M} \rightarrow [0, 1]$ such that

$$\begin{aligned} \mathcal{C}_{\alpha}^{*,P}(X_{\text{obs}(M)}, M) &\stackrel{\text{a.s.}}{=} \left[q_{\beta(X_{\text{obs}(M)}, M)}(P_{Y|X_{\text{obs}(M)}, M}); q_{\gamma(X_{\text{obs}(M)}, M)}(P_{Y|X_{\text{obs}(M)}, M}) \right] \\ \mathbb{E}_P[\gamma(X_{\text{obs}(M)}, M) - \beta(X_{\text{obs}(M)}, M) | M] &\stackrel{\text{a.s.}}{=} 1 - \alpha. \end{aligned}$$

Indeed, to minimize the average length, the best oracle solution consists in minimizing the length conditionally to $(X_{\text{obs}(M)}, M)$, which is achieved by an inter-quantile interval, under the unimodality assumption. The quantity $\gamma(X_{\text{obs}(M)}, M) - \beta(X_{\text{obs}(M)}, M)$ corresponds to the $(X_{\text{obs}(M)}, M)$ -conditional coverage, that is on average, conditionally to $M = m$, the targeted $1 - \alpha$.

Yet, in practice, the constructed intervals are not the oracle ones. Therefore, a natural question is whether (IQ-2) extends to a non-oracle \mathcal{C}_{α} . As generally \mathcal{C}_{α} is not based on the underlying true conditional quantiles, we focus on \mathcal{C}_{α} length instead, a quantity similar in spirit to the inter-quantile. We consider the two following assumptions:

$$\Lambda(\mathcal{C}_{\alpha}(X_{\text{obs}(m)}, m)) \stackrel{\text{a.s.}}{\leq} \Lambda(\mathcal{C}_{\alpha}(X_{\text{obs}(m')}, m')) \quad \text{for any } m \subset m', \quad (\text{Len-1})$$

$$\mathbb{E} [\Lambda(\mathcal{C}_{\alpha}(X_{\text{obs}(M)}, M)) | M = m] \leq \mathbb{E} [\Lambda(\mathcal{C}_{\alpha}(X_{\text{obs}(M)}, M)) | M = m'] \quad \text{for any } m \subset m'. \quad (\text{Len-2})$$

We have the following Theorem 8.3.1 on isotonicity (Len-2) under $\mathcal{P}_{\text{MCAR}, \mathbb{Y} \parallel \mathbf{X}}$.

Theorem 8.3.1. *Let \mathcal{C}_α be an $\text{MCV-}\mathcal{P}_{\text{MCAR}, \text{YIM} | \text{X}}$ predictive interval. There exists a predictive interval $\widetilde{\mathcal{C}}_\alpha$ which*

- i) is $\text{MCV-}\mathcal{P}_{\text{MCAR}, \text{YIM} | \text{X}}$,*
- ii) has conditional length smaller or equal to that of \mathcal{C}_α on each pattern,*
- iii) is averaged-length-isotonic w.r.t. the patterns, i.e., satisfies (Len-2).*

The proof of Theorem 8.3.1 exploits the fact that under $\mathcal{P}_{\text{MCAR}, \text{YIM} | \text{X}}$, a strategy to ensure conditional coverage w.r.t. a pattern m , is to transform $(X_{\text{obs}}(m), m)$ into $(X_{\text{obs}}(m'), m')$ by additionally masking some entries, and using the predictive interval given on pattern m' . For $m \subset m'$, we denote $X_{\text{obs}(\max(m, m'))}$ the point in which we additionally mask elements of m' in X . We have that under $\mathcal{P}_{\text{MCAR}, \text{YIM} | \text{X}}$, the distribution of the data *post-masking* is equal to that of the data with more missing entries: $\mathbb{P}_{Y|X_{\text{obs}(\max(M, m'))}, \max(M, m')} = \mathbb{P}_{Y|X_{\text{obs}}(m'), M=m'}$. We can leverage this observation to build intervals: if the averaged length of the predictive interval conditionally to a pattern $m \subset m'$ is larger than that conditionally to a pattern $m \subset m'$, we can replace $\mathcal{C}_\alpha(X_{\text{obs}}(m), m)$ by $\mathcal{C}_\alpha(X_{\text{obs}}(m'), m')$, ensuring both that new interval length is smaller and that we satisfy (Len-2). Formally, we proceed by induction: starting from the pattern $m' = (1, \dots, 1)$ (no data observed), we first consider all patterns $m = (1, \dots, 1, 0, 1, \dots)$ with a single observed value, and define $\widetilde{\mathcal{C}}_\alpha(X_{\text{obs}}(M), M)$, conditionally to $M = m$, as either $\mathcal{C}_\alpha(X_{\text{obs}}(M), M)$ or $\mathcal{C}_\alpha(X_{\text{obs}(\max(M, m'))}, \max(M, m'))$ (depending on which expected length is smaller). We then repeat the same reasoning inductively. For a pattern m , we pick for $\widetilde{\mathcal{C}}_\alpha$ either $\mathcal{C}_\alpha(\cdot, m)$ or the minimal-length interval among all $\mathcal{C}_\alpha(\cdot, m')$ for all patterns m' that have one more missing data than m , and artificially mask on of the components of $X_{\text{obs}}(m)$ when predicting.

Interpretation: we leverage towards predictive interval construction the fact that we can transform an observed point, by removing some covariates, and recover the same distribution as the one with more missing entries. This idea will be one of the key techniques leveraged in Section 8.4.

As consequence of Theorem 8.3.1 is the following corollary, that is obtained by a minimality argument for the oracle interval (i.e., knowing that applying the aforementioned transformation to the oracle does not change it, as it already has minimal-expected length on each pattern):

Corollary 8.3.1. *Let $P \in \mathcal{P}_{\text{MCAR}, \text{YIM} | \text{X}}$. Then the oracle interval $\mathcal{C}_\alpha^{*,P}$ is averaged-length-isotonic w.r.t. the patterns, i.e., satisfies (Len-2).*

Overall, (Len-2) is thus satisfied by our target sets under $\mathcal{P}_{\text{MCAR}, \text{YIM} | \text{X}}$, and thus appears as a reasonable constraint to impose on our predictive sets. Indeed, it seems to be close to the minimal set of assumptions required in order to ensure that no hardness result exists (Section 8.2) while inducing a leverageable structure between patterns that can be compared (Theorem 8.3.1).

8.3.2 Guidelines for practitioners: which information through imputation for quantile regression?

In this section, we highlight specificities of predictive uncertainty quantification under missing covariates with respect to mean regression, and provide generic guidelines usable in practice.

Impute-then-predict. Most predictive algorithms can not cope directly with missing covariates. To bypass this, the most common approach is to impute the incomplete data via an imputation function Φ , that maps observed values to themselves and missing values to a function of the observed values. Using notations from [Le Morvan et al. \(2021\)](#) we note $\varphi^m : \mathbb{R}^{|\text{obs}(m)|} \rightarrow \mathbb{R}^{|\text{mis}(m)|}$ the imputation function which, given a mask $m \in \mathcal{M}$, takes as input observed values and outputs imputed values, i.e., plausible values. Then, the overall imputation function Φ belongs to $\mathcal{F}^I := \left\{ \Phi : \mathcal{X} \times \mathcal{M} \rightarrow \mathcal{X} : \forall j \in \llbracket 1, d \rrbracket, (\Phi(X, M))_j = X_j \mathbb{1}_{M_j=0} + (\varphi^M(X_{\text{obs}(M)}))_j \mathbb{1}_{M_j=1} \right\}$. The imputed data set becomes the n random variables $(\Phi(X, M), M, Y)$. In practice, Φ is the result of an algorithm \mathcal{I} trained on $\{(X^{(k)}, M^{(k)})\}_{k=1}^{n+1}$. The impact of imputation has been studied for mean regression tasks (in particular in [Le Morvan et al., 2021](#); [Ayme et al., 2023, 2024](#)).

How to account for the missingness when imputing? Given the impact of missing covariates on the shape of prediction uncertainty discussed in Section 8.3.1, impute-then-predict raises a specific concern: is there a way to impute which incorporates the necessary information on the missing values?

Hereafter, we show that the answer is significantly different if we restrict ourselves to mean regression. Specifically, we show that incorporating the mask (e.g., by concatenating the mask to the features) is more critical for quantile regression. To that end, we provide in Proposition 8.3.3 simple models showcasing that unbiased imputation choices are sufficient to obtain an optimal model for regression but fail for quantile regression. For mean regression, the efficiency of such imputation methods have been established in practice (see e.g., [Josse et al., 2019](#); [Le Morvan et al., 2021](#)) and Proposition 8.3.3 support those findings.

Proposition 8.3.3. Assume $\mathcal{P}_{\text{MCAR}, \text{YIM} | \mathcal{X}}$ and $Y = \beta^{*T} X + \varepsilon$ with ε s.t. $\mathbb{E}[\varepsilon | X_{\text{obs}(M)}, M] = 0$.

i) *Mean regression*

- if the covariates $(X_j)_{j=1}^d$ are independent, then the optimal linear model taking $\Phi_{\text{mean}}(X, M)$ as input is Bayes optimal, with Φ_{mean} the imputation by the mean;
- the optimal linear model taking $\Phi_{\text{conditional mean}}(X, M)$ as input is Bayes optimal, with $\Phi_{\text{conditional mean}}$ the imputation by the conditional mean;

ii) *Any quantile linear model taking unbiased imputed data as input (i.e., $\mathbb{E}[\Phi(X, M) | M] \stackrel{\text{a.s.}}{=} \mathbb{E}[X]$) leads to intervals of constant expected length across patterns, thus is not Bayes optimal when $Y \not\propto X$.*

Point *i)* of Proposition 8.3.3 illustrates that if the learner was able to retrieve the true underlying regression coefficients and the data were imputed by their mean or conditional mean, then the learned model would be the best possible at the task of predicting the conditional expectation, i.e., all necessary information is preserved by using only the imputed data set and not leveraging the associated mask. Although the non-necessity of using the mask in the conditional expectation estimation and MCAR framework does not systematically extend when the data is more complex than linear, it is insightful as even in the linear setting, the same does not hold for quantile regression.

Indeed, point *ii)* of the same Proposition 8.3.3 highlights that on the contrary a learner accessing the true underlying regression coefficients with the very same unbiased imputed data would not lead to an optimal model, as a method whose resulting predictive interval have constant lengths across the missing patterns does not retrieve the underlying heteroskedasticity induced by the missing values (Section 8.3.1), and thereby cannot be MCV. Precisely, the assumption on the imputation for this result corresponds for example to imputing by the feature's expectation (i.e., Φ_{mean}), the feature's conditional expectation (i.e., $\Phi_{\text{conditional mean}}$), or a random draw from a distribution whose expectation is the feature's expectation, under $\mathcal{P}_{\text{MCAR}}$. This includes MICE (van Buuren and Groothuis-Oudshoorn, 2011), which consists in imputing by random draws from the conditional distribution hence the imputed data have the same expectation than the features themselves.

Overall, Proposition 8.3.3 tells that *i)* the state-of-the-art imputation method MICE is not the best choice for predictive uncertainty quantification, *ii)* by contrast to mean regression, in the linear case imputing by the expectation or the conditional expectation is detrimental. In fact, data-independent constant imputation would result in more adaptive intervals. This is because quantile regression needs to retrieve the information on the patterns to adapt its structure to it. Therefore, when using such imputations, **a natural idea is to give the information of the mask to the model by concatenating the mask to the features.**

8.4 Principled unified Missing Data Augmentation (MDA) framework: CP-MDA-Nested*

In this section, we go beyond generic guidelines and we introduce a general framework, coined CP-MDA-Nested*, to generate predictive sets that achieve MCV under $\mathcal{P}_{\text{MCAR}, \text{YIM} | \mathbf{X}}$. Our approach is applicable to both classification and regression tasks, by building upon any conformal score function (Vovk et al., 2005). It combines over-masking ideas introduced in Section 8.3, with subsampling techniques, and similar machinery than leave-one-out conformal prediction methods (Barber et al., 2021b; Gupta et al., 2022).

8.4.1 Presentation of CP-MDA-Nested*

We start by reminding the necessary concepts of split Conformal Prediction (CP) in the complete case, without missing values, before diving into the details of our unified framework CP-MDA-Nested*.

8.4.1.1 BACKGROUND ON SPLIT CP

Introduced in [Papadopoulos et al. \(2002\)](#); [Vovk et al. \(2005\)](#); [Lei et al. \(2018\)](#), split CP builds predictive regions by first splitting the n points of the training set into two disjoint sets $\text{Tr}, \text{Cal} \subset \llbracket 1, n \rrbracket$, to create a *proper training set*, Tr , and a *calibration set*, Cal , of sizes $\#\text{Tr} = (1 - \rho)n$ and $\#\text{Cal} = \rho n$ with $\rho \in]0, 1]$. On the proper training set, a model \hat{f} (chosen by the user) is fitted, and then used to predict on the calibration set. *Conformity scores* $\mathcal{S} = \left\{ \left(s \left(X^{(k)}, Y^{(k)}; \hat{f} \right) \right)_{k \in \text{Cal}} \right\} \cup \{+\infty\}$ are computed to assess how well the fitted model \hat{f} predicts the response values of the calibration points. In regression, usually the absolute value of the residuals is used, i.e. $s(x, y; \hat{\mu}) = |\hat{\mu}(x) - y|$. In classification, the simplest score is $s(x, y; \hat{p}) = 1 - \hat{p}(x)_y$ (where $\hat{p} : \mathcal{X} \mapsto [0, 1]^{\mathcal{Y}}$ outputs a vector of estimated probabilities for each class). Finally, the $(1 - \alpha)$ -th quantile of these scores $q_{1-\alpha}(\mathcal{S})$ (i.e., their $\lceil (1 - \alpha)(\#\text{Cal} + 1) \rceil$ smallest value) is computed to define the predictive region: $\hat{C}_{n,\alpha}(x) := \{y \in \mathcal{Y} \text{ such that } s(x, y; \hat{f}) \leq q_{1-\alpha}(\mathcal{S})\}$. In regression with the absolute values of the residual score, this reduces to $\hat{C}_{n,\alpha}(x) := [\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$.

This procedure satisfies Equation (8.1) for any \hat{f} , any (finite) sample size n , as long as the data points are exchangeable.⁴ For more details on split CP, we refer to [Angelopoulos and Bates \(2023\)](#); [Vovk et al. \(2005\)](#), as well as to [Manokhin \(2022\)](#).

8.4.1.2 CP-MDA-NESTED*

From an high level perspective, the idea is to apply split CP on top of an impute-then-predict pipeline (of imputation function Φ), and to modify the calibration step in order to ensure MCV. This is called CP-MDA, for *conformal prediction with missing data augmentation*. Generally, for a given test point $(X^{(n+1)}, M^{(n+1)})$, CP-MDA works by artificially masking covariates in the calibration set so as to match *at least* the mask of the test point, by creating a new mask $\widetilde{M}^{(k)} = \max(M^{(k)}, M^{(n+1)})$ for each $k \in \text{Cal}$. In other words, it corresponds to discarding from the calibration set the covariates that are missing in the test point. This leads to $M^{(n+1)} \subseteq \widetilde{M}^{(k)}$, i.e., all over-masked (or *augmented*) points $(X^{(k)}, \widetilde{M}^{(k)}, Y^{(k)})_{k \in \text{Cal}}$ have at least the missing entries of $(X^{(n+1)}, M^{(n+1)})$. The points such that $\widetilde{M}^{(k)} = M^{(n+1)}$ can be used directly as under distributional assumptions $(\mathcal{P}_{\text{MCAR}, \text{YIM} | \mathcal{X}}^{\otimes(n+1)})$, they now have the same mask and distribution as the test point. Yet for many calibration points it remains that $\widetilde{M}^{(k)} \neq M^{(n+1)}$ (precisely, for all the $k \in \text{Cal}$ such that $M^{(k)} \not\subseteq M^{(n+1)}$). This means that those over-masked points follow another conditional distribution than the one of the test point, and MCV can not be directly ensured.

An idea is to subsample the calibration set so that the effective calibration set contains only $k \in \text{Cal}$ such that $M^{(k)} \not\subseteq M^{(n+1)}$ (i.e., $\widetilde{M}^{(k)} = M^{(n+1)}$) (this is the approach followed in CP-MDA-Exact, [Zaffran et al., 2023](#)). However, this can lead to overly small calibration set size in high dimension, resulting in a large variance (on the coverage level and thus set size). Therefore, two questions naturally arise:

- How to build the calibration set?

⁴Only the calibration and test data points need to be exchangeable.

- How to leverage the test point so as to account for the different distributions present in the over-masked calibration set—and with many of them not matching the test mask conditional distribution—when constructing the predictive set?

The answers we suggest define our generalized framework CP-MDA-Nested*, whose pseudocode is available in Algorithm 16, and are illustrated in Figure 8.2.

Construction of the calibration set. CP-MDA-Nested* includes a subsampling step: it calibrates on the set of indices $\widetilde{\text{Cal}} \subseteq \text{Cal}$ provided by the user, where $\widetilde{\text{Cal}}$ can be obtained with any subsampling strategy, that might even be stochastic, as long as the randomness is independent of the covariates and outputs, $(X^{(k)}, Y^{(k)})_{k \in \text{Cal} \cup \{n+1\}}$ (it can still depend on the masks). The following strategies work if the data distribution belongs to $\mathcal{P}_{\text{MCAR}, \text{YIM} | \mathbf{X}}^{\otimes(n+1)}$ (which is an assumption we make anyway when using CP-MDA-Nested* since, as we show precisely in Theorem 8.4.2, CP-MDA-Nested* is typically MCV- $\mathcal{P}_{\text{MCAR}, \text{YIM} | \mathbf{X}}^{\otimes(n+1)}$):

- subsampling only the indices $\{k \in \text{Cal} : M^{(k)} \subseteq M^{(n+1)}\} := \widetilde{\text{Cal}}$ (this is the strategy of CP-MDA-Exact, Zaffran et al., 2023);
 - no subsampling, $\widetilde{\text{Cal}} := \text{Cal}$ (this is the path taken by CP-MDA-Nested, Zaffran et al., 2023);
 - subsampling only the indices $\{k \in \text{Cal} : M^{(k)} \subseteq m'\} := \widetilde{\text{Cal}}$, for some $m' \supseteq M^{(n+1)}$;
 - obtain $\widetilde{\text{Cal}}$ by subsampling from the indices $\{k \in \text{Cal} : M^{(k)} \subseteq m'\}$, for some $m' \supseteq M^{(n+1)}$, using a mixture distribution, whose weights only depend on $(M^{(k)})_{k \in \text{Cal} \cup \{n+1\}}$.
- Then, for any $k \in \widetilde{\text{Cal}}$, the over-mask is constructed, defining $\widetilde{M}^{(k)} = \max(M^{(k)}, M^{(n+1)})$. This is schematized in Figure 8.2.

Leveraging temporary test points. After the subsampling step aforementioned, the over-masked calibration points and the test point do not necessarily have the same conditional distribution conditionally to the mask, as $M^{(n+1)} \subseteq \widetilde{M}^{(k)}$ without equality in general. In order to match those distributions, an idea is to create **temporary test points** (one for each calibration point) and to apply $\widetilde{M}^{(k)}$ to it. This is illustrated in **green** in Figure 8.2. CP-MDA-Nested* evaluates the number of over-masked calibration points that have a conformity score smaller than that of the **corresponding over-masked test point** for a given $y \in \mathcal{Y}$. Then, the predictive set includes only the $y \in \mathcal{Y}$ such that this number is small enough (a threshold that depends on α and the effective calibration size). This careful treatment of the test point allows to compare scores obtained from identical distributions conditionally on their associated mask.

8.4.1.3 KEY COMMENTS ON CP-MDA-NESTED*

In summary, CP-MDA-Nested* bridges the gap between CP-MDA-Exact and CP-MDA-Nested by proposing a tighter generalized framework. On the one hand, CP-MDA-Exact comes with a potentially small calibration set, thus with increased variability. On the other hand, by leveraging all calibration points, including those with very few observed covariates, the average interval length produced by CP-MDA-Nested is typically larger than

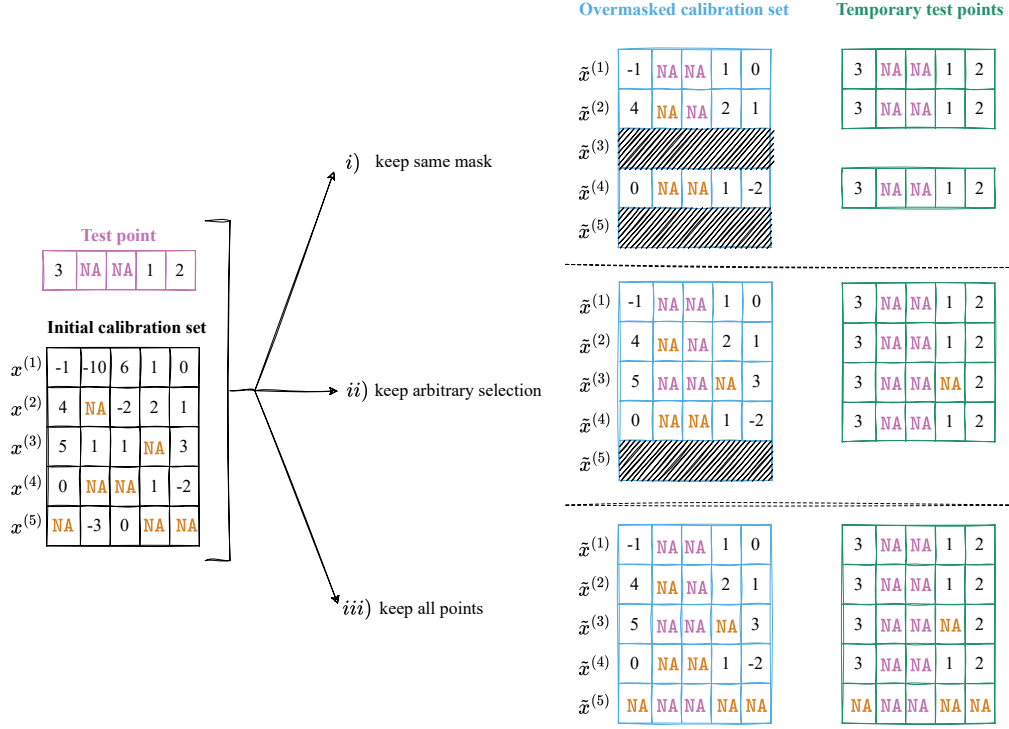


Figure 8.2: CP-MDA-Nested* illustration. Different subsampling strategies are shown, with their associated over-masked calibration set and temporary test points according to one test point.

Algorithm 16 CP-MDA-Nested*

Input: Training set $\{(X^{(k)}, M^{(k)}, Y^{(k)})\}_{k=1}^n$, imputation algorithm \mathcal{I} , learning algorithm \mathcal{A} taking its values in $\mathcal{F} := \mathcal{Y}^{\mathcal{X} \times \mathcal{M}}$, calibration proportion $\rho \in]0, 1]$, $\{\text{Tr}, \text{Cal}, \Phi, \hat{A}\}$ the output of the splitting Algorithm 17 ran on $\left\{\{(X^{(k)}, M^{(k)}, Y^{(k)})\}_{k=1}^n, \mathcal{I}, \mathcal{A}, \rho\right\}$, conformity score function $s(\cdot, \cdot; f)$ for $f \in \mathcal{F}$, significance level α , test point $(X^{(n+1)}, M^{(n+1)})$, subsampled set of calibration indices $\widetilde{\text{Cal}} \subseteq \text{Cal}$

Output: Prediction set $\widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*}(X^{(n+1)}, M^{(n+1)})$

// Generate an over-masked calibration set:

1: **for** $k \in \widetilde{\text{Cal}}$ **do** Additional nested masking

2: $\widetilde{M}^{(k)} = \max(M^{(k)}, M^{(n+1)})$

3: **end for** Over-masked calibration set generated. //

4: $\widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*}(X^{(n+1)}, M^{(n+1)}) := \left\{y \in \mathcal{Y} : (1 - \alpha)(1 + \#\widetilde{\text{Cal}}) >$

$$\sum_{k \in \text{Cal}} \mathbb{1} \left\{ s \left(\left(X^{(k)}, \widetilde{M}^{(k)} \right), Y^{(k)}; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) < s \left(\left(X^{(n+1)}, \widetilde{M}^{(k)} \right), y; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) \right\} \right\}$$

Algorithm 17 Split and train

Input: Imputation algorithm \mathcal{I} , learning algorithm \mathcal{A} taking its values in $\mathcal{F} := \mathcal{Y}^{\mathcal{X} \times \mathcal{M}}$, training set $\{(X^{(k)}, M^{(k)}, Y^{(k)})\}_{k=1}^n$, calibration proportion $\rho \in]0, 1]$

Output: Splitted sets of indices Tr and Cal, imputation function Φ , fitted predictor \hat{A}

- 1: Randomly split $\{1, \dots, n\}$ into 2 disjoint sets Tr & Cal of sizes $\#\text{Tr} = (1 - \rho)n$ and $\#\text{Cal} = \rho n$
- 2: Fit the imputation function: $\Phi(\cdot, \cdot) \leftarrow \mathcal{I}(\{(X^{(k)}, M^{(k)}), k \in \text{Tr}\})$
- 3: Fit the learning algorithm \mathcal{A} : $\hat{A}(\cdot, \cdot) \leftarrow \mathcal{A}(\{(\Phi(X^{(k)}, M^{(k)}), M^{(k)}), k \in \text{Tr}\})$

that of CP-MDA-Exact (cf. (Len-2)). Furthermore, CP-MDA-Nested is less generic than CP in the sense that it is specific to predictive *intervals* (unlike CP-MDA-Exact which is as generic as CP and can be plugged with any score function, including classification). Overall, CP-MDA-Nested* unifies this framework for any score function and provides high flexibility in the trade-offs between *efficiency* and *variability*:

- At the extreme of no subsampling at all, we obtain a generalization of CP-MDA-Nested which encapsulates the classification setting;
- This generalization provides tighter sets than that of CP-MDA-Nested in the particular case of interval-based scores (see Remark 8.4.1);
- At the other extreme of the strictest subsampling procedure, we retrieve CP-MDA-Exact;
- Any other less restrictive subsampling (possibly with a random selection between various augmented mask) belongs to this framework, providing more flexibility in the trade-offs between exact validity and statistical variability.

This overview is summarized in Table 8.4.

In the case where the nested predictive sets are intervals and $\widetilde{\text{Cal}} = \text{Cal}$, then the final predictive sets obtained through CP-MDA-Nested* are included in the ones of CP-MDA-Nested.

Remark 8.4.1. When $\widetilde{\text{Cal}} = \text{Cal}$, and using absolute value of the residuals scores or conformalized quantile regression scores (Romano et al., 2019), or any score such that $\{y \in \mathcal{Y} \text{ such that } s(x, y; \hat{f}) \leq b\}$ for some b is an interval, then $\widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*}(\cdot) \subseteq \widehat{C}_{n,\alpha}^{\text{MDA-Nested}}(\cdot)$ (see Section 8.D).

This unification allows us to provide theoretical guarantees, stated in Section 8.4.2, leveraging the deep connections between CP-MDA-Nested* and leave-one-out conformal

Method	CP-MDA-Exact	CP-MDA-Nested* (new)	CP-MDA-Nested
Size of actual calibration set	# points in Cal with $M \subseteq M^{(n+1)}$	Any	#Cal
Mask of the points used for calibration	exactly $M^{(n+1)}$		all, leading to \widetilde{M} s.t. $M^{(n+1)} \subseteq \widetilde{M}$
Overall behavior	Too few Cal points \rightarrow high coverage variance	Flexible	Too large intervals (cf. (Len-2))
Applies to classification	✓	✓(new)	✗
Outputs non-interval sets	✓	✓(new)	✗
Marginal guarantee (MV)	✓	✓(new)	✓(new)
Conditional guarantee (MCV)	✓	✓(new)	✓(new)

Table 8.4: Summary of the high-level characteristics of MDA algorithms, coming from the literature, as well as our novel contributions indicated by “(new)”. Characteristics are given for a test point $(X^{(n+1)}, Y^{(n+1)}, M^{(n+1)})$. Details regarding guarantees are given in Table 8.5.

Guarantees	MV	MCV
CP-MDA-Exact i.e., CP-MDA-Nested* with subsampling only $k \in \text{Cal}$ s.t. $M^{(k)} \subseteq M^{(n+1)}$	$\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} X}^{\otimes(n+1)}$, level α , with upper bound, from Zaffran et al. (2023)	$\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} X}^{\otimes(n+1)}$, level α , with upper bound, from Zaffran et al. (2023)
CP-MDA-Nested*	$\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} X}^{\otimes(n+1)}$, level 2α	$\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} X}^{\otimes(n+1)}$, level 2α
CP-MDA-Nested* without subsampling	$\mathcal{P}^{\text{exch}(n+1)}$, level 2α	$\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} X}^{\otimes(n+1)}$, level 2α

Table 8.5: Theoretical guarantees of CP-MDA-Nested* depending on the subsampling choice.

methods (such as Barber et al., 2021b; Gupta et al., 2022). Indeed, the rationale for predicting on masked test points, using the augmented calibration masked, is that we want to treat the test and calibration points in a symmetric way. We summarize them in the following Table 8.5.

8.4.2 Theoretical guarantees on CP-MDA-Nested and CP-MDA-Nested* leveraging their connection to leave-one-out CP

Hereafter, we give our theoretical results on the coverage of our CP-MDA-Nested* algorithm.

Theorem 8.4.1 (Marginal validity of CP-MDA-Nested*). *CP-MDA-Nested* with $\widetilde{\text{Cal}} = \text{Cal}$ and (and thus CP-MDA-Nested) is MV- $\mathcal{P}^{\text{exch}(n+1)}$ at the level $1 - 2\alpha$.*

Theorem 8.4.1 provides a lower bound on CP-MDA-Nested* and CP-MDA-Nested coverage as $1 - 2\alpha$. This result is important as it equips CP-MDA-Nested* with $\widetilde{\text{Cal}} = \text{Cal}$ and CP-MDA-Nested with controlled coverage on any exchangeable distribution: they are marginally valid even on MNAR distributions or when $Y \not\perp M | X$. It means that despite modifying the data set independently from X and Y and breaking the structure of (X, M, Y) , the obtained estimator makes reliable predictions including when X, M , and Y are strongly dependent. This originates from the fact that the whole data set has been treated equally, including the test point.

Theorem 8.4.2 (Conditional validity of CP-MDA-Nested*). *CP-MDA-Nested* with $\widetilde{\text{Cal}}$ independent of the data set $(X^{(k)}, Y^{(k)})_{k \in \text{Cal} \cup \{n+1\}}$ (and thus CP-MDA-Nested) is MCV- $\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} | X}^{\otimes(n+1)}$ at the level $1 - 2\alpha$.*

The proofs of Theorems 8.4.1 and 8.4.2 are deferred to Section 8.D.1 and Section 8.D.2 respectively. They are heavily based on the deep connections between CP-MDA-Nested* with Jackknife+ and general leave-one-out or k -fold CP (Barber et al., 2021b; Vovk, 2015; Gupta et al., 2022). Indeed, one can observe that, for each $k \in \text{Cal}$, we evaluate the conformity score of the test point $(X^{(n+1)}, M^{(n+1)}, Y^{(n+1)})$ using the k -th augmented mask, as the equivalent of evaluating the conformity score of the test point with the fitted model that has left-out the k -th calibration point. This connection between CP-MDA-Nested* and leave-one-out conformal approaches directly stems from the same core motivations: *i*) both enforce a design that use all the observations of the training or calibration sets to handle

small sample sizes, *ii*) both need to avoid invalid designs that arise naturally when keeping all these points, such as comparing scores obtained with different predictors.

On the factor 2 and link with empirical quantile aggregation. Despite the coverage guarantee being of $1 - 2\alpha$ instead of the desired $1 - \alpha$, in practice, our experiments in Section 8.5 show that CP-MDA-Nested* without subsampling (i.e., CP-MDA-Nested) tends to over-cover. This aligns with Figure 2 of Barber et al. (2021b), that illustrates the fact that leave-one-out conformal methods achieve empirically exactly $1 - \alpha$ coverage, while K -fold conformal approaches over-cover. The reason behind this phenomenon is still unclear in the community, and is likely to be the same than the reason for CP-MDA-Nested* over-coverage, as one can see CP-MDA-Nested* as having access to many folds of calibration points, since each augmented calibration mask typically appears many times in the calibration set. In particular, Zaffran et al. (2023) provide $\text{MCV-}\mathcal{P}_{\text{MCAR}, \text{YIM} | \mathbf{X}}^{\otimes(n+1)}$ guarantees at the level $1 - \alpha$ on a modified version of CP-MDA-Nested which leverages this folding point of view by calibrating only on one (arbitrarily) chosen such fold. Similarly than for K -fold and leave-one-out conformal methods, we can look at CP-MDA-Nested* as a way to aggregate many valid empirical quantiles or p -values, one for each fold, i.e., one for each augmented mask. Due to the strong dependencies between these random variables, such an aggregation does not lead to a valid aggregated quantile or p -value, and induces a loss of coverage.

Theorem 8.4.2 proof approach: coupling our algorithm with a leave-one-out conformal method on a virtual complete data set. We work conditionally to the mask of the test point, $M^{(n+1)}$. Then, we introduce a randomized predictor, for which “training” consists in randomly picking one individual predictor among a bag of individual predictors, each of them corresponding to an augmented calibration mask. This bag contains exactly $2^{|\text{obs}(M^{(n+1)})|}$ possible individual predictors, where $|\text{obs}(M^{(n+1)})|$ is the number of 1s in $M^{(n+1)}$, i.e., the number of observed features in the test point. Each individual predictor in the bag is thus parametrized by a *super/over-mask* of $M^{(n+1)}$. We call such a predictor a mixture-predictor, as it basically consists in picking randomly one individual predictor in a mixture of individual predictors. That sampling has to be made independently of the data the mixture predictor is applied to, but non necessarily uniformly. Furthermore, we ensure that the individual predictor indexed by a mask M only relies on the covariates $X_{\text{obs}(M)}$ for the prediction, in order for this algorithm to be applicable in practice (e.g., an invalid design would be individual predictors that require the knowledge of some of the $X_{\text{mis}(M)}$, unobserved in practice, in order to make predictions).

We then show that our algorithm CP-MDA-Nested*, applied to the data set with missing entries $\left(X_{\text{obs}(M^{(k)})}^{(k)}, Y^{(k)}, M^{(k)}\right)_{k=1}^{n+1}$, has the same guarantees in expectation as the leave-one-out conformal that uses the mixture predictor, applied onto a virtual complete data set $(X^{(k)}, Y^{(k)})_{k=1}^{n+1}$, if we make some assumptions on the missingness distribution. More specifically, we show that *there exists a coupling between the two algorithms*, that ensures that the output (and thus coverage) have the same distribution. This ultimately enables us to reuse existing guarantees for leave-one-out conformal estimators.

8.5 A practical glimpse on the impacts of breaking the distribution's assumptions

In this concluding section, we investigate the numerical performances of **CP-MDA-Nested*** mainly outside its theoretical set of assumptions. Experiments under $\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} | X}$ are provided in Section 8.5.1, then Section 8.5.2 presents numerical results when the data distribution either belongs to \mathcal{P}_{MAR} or $\mathcal{P}_{\text{MNAR}}$, and finally Section 8.5.3 reports empirical performances when $Y \not\perp M | X$.

In all experiments, the data are imputed using iterative regression (**iterative ridge** implemented in Scikit-learn, [Pedregosa et al. \(2011\)](#)). The predictive models are fitted on the imputed data concatenated with the mask. The prediction algorithm is a neural network, fitted to minimize the pinball loss ([Sesia and Romano, 2021](#)). For the vanilla QR, we use both the training and calibration sets for training. The training set contains 500 data points, and the calibration set 250 data points. To evaluate the marginal coverage, a test set is generated with missing values following the same distribution as on the training and calibration sets. Then, to estimate mask-conditional coverage (i.e., $\mathbb{P}(Y \in \hat{C}_{n,\alpha}(X, m) | M = m)$ for each $m \in \mathcal{M}$), we generate another test set by imposing that the number of observations per pattern is fixed to 100, in order to ensure that the variability is not impacted by $\mathbb{P}(M = m)$. Each experiment is repeated 100 times (unless stated otherwise).

8.5.1 Experiments under $\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} | X}$

Data generation. The data is generated with $d = 10$ according to Model 8.3.2 (regression), $Y = \beta^T X + \varepsilon$ with $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, \dots, 1)^T$ and $\Sigma = \varphi(1, \dots, 1)^T(1, \dots, 1) + (1 - \varphi)I_d$, $\varphi \in \{0, 0.8\}$ depending on the experiment, Gaussian noise $\varepsilon \sim \mathcal{N}(0, 1) \perp (X, M)$ and the following regression coefficients $\beta = (1, 2, -1, 3, -0.5, -1, 0.3, 1.7, 0.4, -0.3)^T$. Each of these 10 features is missing with probability 0.2 independently from anything else.

8.5.1.1 CP-MDA-NESTED* PROVIDES FLEXIBILITY

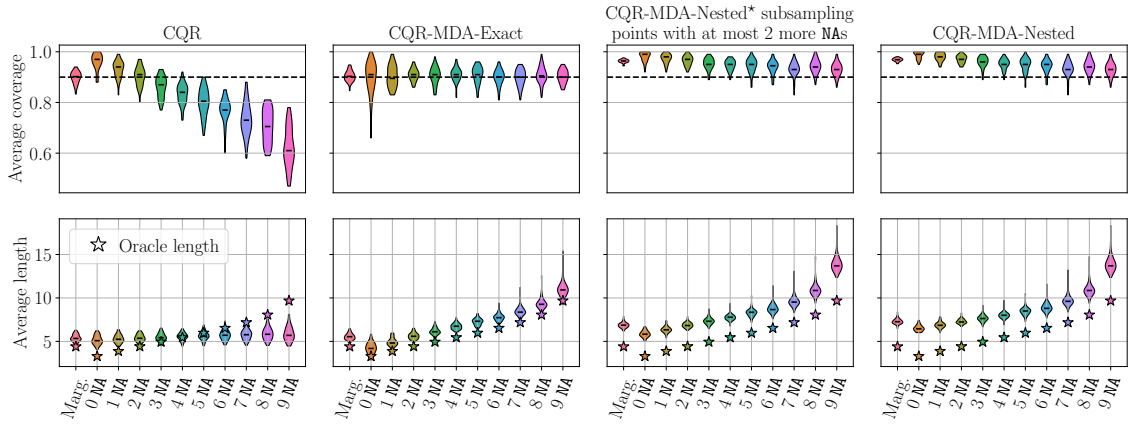
In our first experiments, we compare CQR to CP-MDA-Exact and CP-MDA-Nested, as well as **CP-MDA-Nested*** where we subsample all the calibration points that have at most two features that are missing among the observed features of the test point. As $d = 10$, there are 1024 different patterns, avoiding to display the performances of the algorithms on each of the patterns. Therefore, instead, we represent the coverage and the length of the predictive intervals depending on the mask size, a proxy for mask-conditional coverage. For each pattern size, 100 observations are drawn according to the distribution of $M | \text{size}(M)$ in the test set. In this subsection only, the number of repetition is of 50.

Figure 8.3a displays the results of this experiment. As noticed in [Zaffran et al. \(2023\)](#), CQR is not $\text{MCV-}\mathcal{P}_{\text{MCAR}, \mathbb{Y} \perp \mathbb{M} | X}^{\otimes(n+1)}$ as its intervals undercover or overcover depending on the number of missing values. The three versions of **CP-MDA-Nested*** ensure that the coverage is at least $1 - \alpha$ for any pattern size, as supported by our theory (Section 8.4.2)⁵ Comparing CP-MDA-Exact and CP-MDA-Nested, we observe that CP-MDA-Exact is more

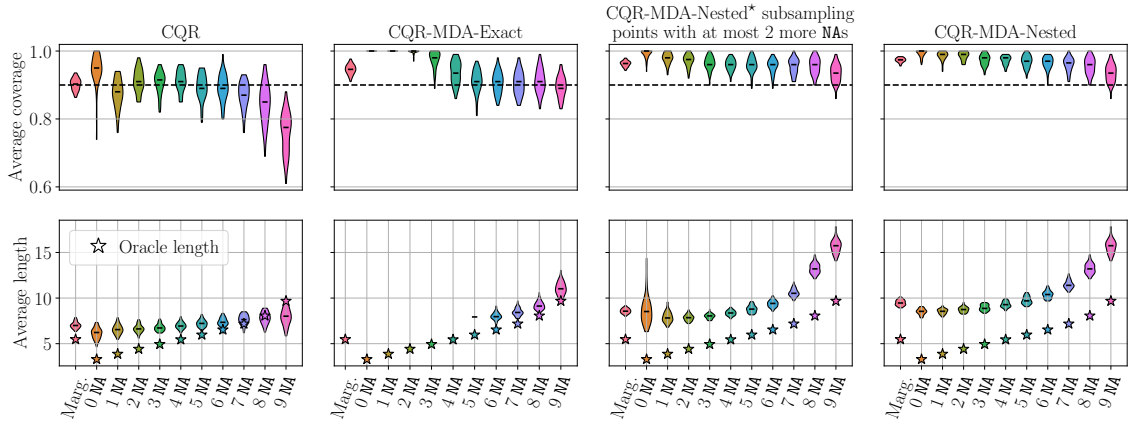
⁵Note that MCV implies validity on any mask size, but not the contrary.

efficient as it produces smaller intervals and its coverage is exactly of $1 - \alpha$ on average, while suffering for more variability than CP-MDA-Nested. The intermediate version of **CP-MDA-Nested*** allows to reduce CP-MDA-Exact variability while improving the efficiency of the intervals by 5.5% marginally (the comparison consists in computing the difference between **CP-MDA-Nested*** and CP-MDA-Nested intervals' median length, and normalize it by CP-MDA-Nested intervals' median length), reaching nearly 10% of improvement on the test points that have no missing values. For 7 to 9 missing values, this **CP-MDA-Nested*** is equivalent to CP-MDA-Nested as the subsampling scheme of **CP-MDA-Nested*** boils down to keeping all the calibration points on these patterns.

CP-MDA-Nested reveals all its interest over CP-MDA-Exact in settings where the exact subsampled calibration set contains really few points for some masks (e.g., in high dimension or when the probability of missing values is high). In Figure 8.3b, the probability of each features being missing is increased to 0.4. We observe that CP-MDA-Exact outputs infinite intervals more than half of the time on the marginal test, as well as on the test sets



(a) Each features is missing with probability 0.2.



(b) Each features is missing with probability 0.4.

Figure 8.3: Validity and efficiency with **MCAR missing values** on dependent Gaussian features, with $\varphi = 0.8$, and such that $\mathbf{Y} \perp \mathbf{M} \mid \mathbf{X}$. Average coverage (top) and length (bottom) as a function of the missing pattern sizes. The black horizontal line in each violin plot corresponds to the median. The first violin plot shows the marginal coverage. The marginal test set includes 2000 observations. The mask-conditional test set includes 100 individuals for each missing data pattern size.

containing between 0 and 4 missing values. This is particularly unpractical. On the contrary, CP-MDA-Nested produces finite length intervals on any test point, at the cost of being overly conservative. The improvements brought by CP-MDA-Nested* with subsampling only the calibration points with at most 2 additional missing values are more stringent. In particular, the efficiency is improved by nearly 9.5% marginally, and is in between 8.5% and 10% on test points that have between 1 and 6 missing values.

Note that this is only one example of CP-MDA-Nested* for a given subsampling strategy, and that in practice the choice of strategy is highly dependent on the settings and could lead to even better performances. From now on, we restrict the subsequent experiments with CP-MDA-Nested* to the two extremes—CP-MDA-Exact and CP-MDA-Nested—as their main goal is to investigate the robustness beyond $\mathcal{P}_{\text{MCAR}, \mathbf{Y} \perp \mathbf{M} | \mathbf{X}}$. For the same reason, we do not want to restrict ourselves to the mask-size conditional coverage, as it does not imply mask conditional coverage. Therefore, we use another visualization approach that was introduced in Zaffran et al. (2023). As an appetizer, Figure 8.4 presents the results under $\mathcal{P}_{\text{MCAR}, \mathbf{Y} \perp \mathbf{M} | \mathbf{X}}^{\otimes(n+1)}$ for QR, CQR, CP-MDA-Exact and CP-MDA-Nested, using this visualization. The x -axis represents the average coverage and the average length is in the y -axis. The marker colors are associated to the different methods. A method is MCV if all the markers of its color are at the right of the vertical dotted line (90%). The design of Figure 8.4, and the following figures, requires a cautious interpretation. For each method we report, for the pattern having the highest (or lowest) coverage, its length and coverage. However, as this pattern may depend on the method, the length for the highest/lowest should not be directly compared between methods.

This Figure 8.4 illustrates that CP-MDA-Nested* is MCV- $\mathcal{P}_{\text{MCAR}, \mathbf{Y} \perp \mathbf{M} | \mathbf{X}}^{\otimes(n+1)}$. Our hardness results of Section 8.2 provide a new perspective on these results:

Remark 8.5.1. If CP-MDA-Nested* was MCV on a broader class of distributions than $\mathcal{P}_{\text{MCAR}, \mathbf{Y} \perp \mathbf{M} | \mathbf{X}}^{\otimes(n+1)}$ for which a hardness result exists, then it would produce uninformative intervals on any distribution within this class, including $\mathcal{P}_{\text{MCAR}, \mathbf{Y} \perp \mathbf{M} | \mathbf{X}}^{\otimes(n+1)}$. Therefore, the fact

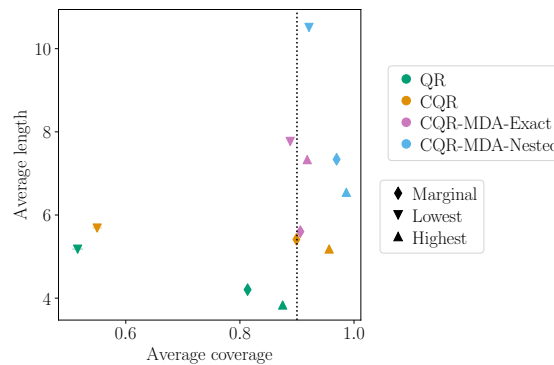


Figure 8.4: Validity and efficiency with **MCAR missing values** on dependent Gaussian features, with $\varphi = 0.8$, and such that $\mathbf{Y} \perp \mathbf{M} | \mathbf{X}$. Colors represent the methods. Diamonds (\blacklozenge) represent marginal coverage while the patterns giving the lowest and highest coverage are represented with triangles (\blacktriangledown and \blacktriangle). Vertical dotted lines represent the target coverage of 90%. Experimental details: $\#\text{Tr} = 500$; $\#\text{Cal} = 250$; the marginal test set includes 2000 observations; the mask-conditional test set includes 100 individuals for each missing data pattern.

that **CP-MDA-Nested*** obtain finite length intervals in this experiment (Figure 8.4) tends to confirm (with high probability) that the theory on the **CP-MDA-Nested*** MCV can not be extended to $\mathcal{P}_{\text{YIM}|X}^{\otimes(n+1)}$ or $\mathcal{P}_{\text{MAR}}^{\otimes(n+1)}$ nor $\mathcal{P}_{\text{MCAR}}^{\otimes(n+1)}$. This analysis is included in Table 8.2, as a numerical confirmation on **CP-MDA-Nested*** theory.

8.5.2 Beyond MCAR

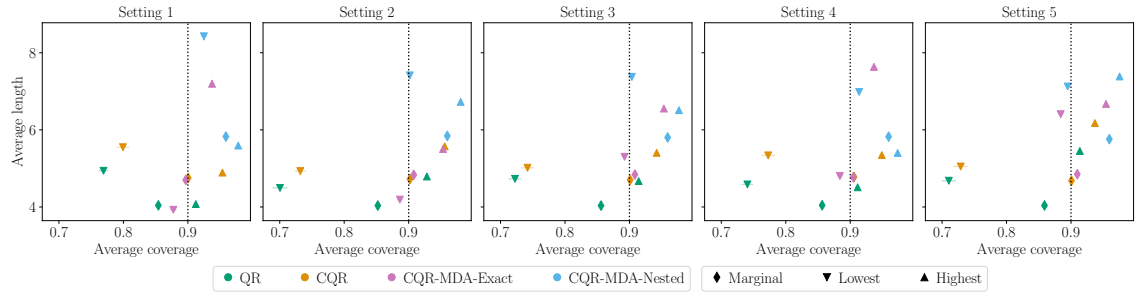
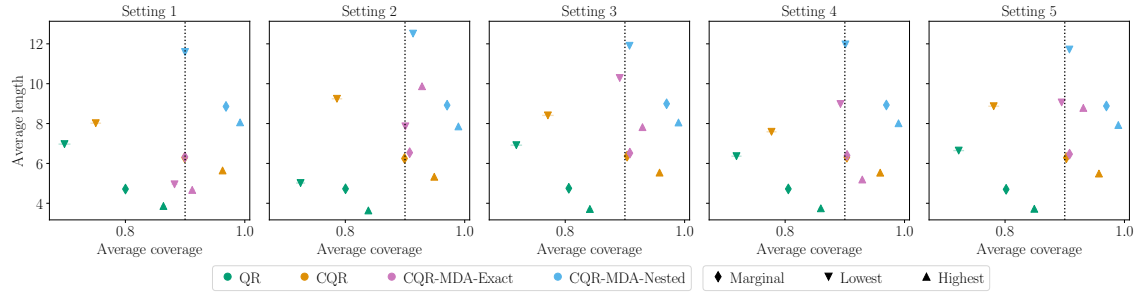
Beyond MCAR experiments. To generate missing values under MAR or MNAR distribution, we select 6 variables (denote this set X_{missing}) out of 10 that can be missing (the 4 others form the set X_{observed}). Especially, $X_{\text{missing}} = \{X_1, X_2, X_3, X_5, X_8, X_9\}$ in order to include different range of associated regression coefficients. We used the GitHub repository associated to [Muzellec et al. \(2020\)](#) in order to introduce missing values in X_{missing} according to the following mechanisms, fixing the proportion of missing entries to be 20%. For each of these following settings, we run two sets of experiments: one in which the correlation between the features is high ($\varphi = 0.8$) and therefore imputing through iterative regression allows to recover quite accurately the missing values, and one in which the features are independent ($\varphi = 0$) leading to an imputation that can not be better than the marginal expectation of the features.

- MAR experiments (Figure 8.5). Missing values in X_{missing} are introduced under a MAR mechanism. To do so, a logistic model of arguments X_{observed} determines the probability of the variables in X_{missing} to be missing. This setting is declined 5 times, with different weights for the logistic model. Within each one, the experiments are repeated 100 times to assess for the variability.

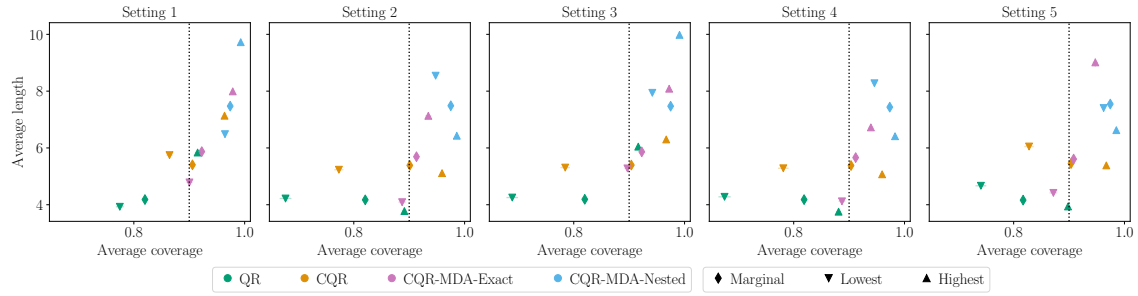
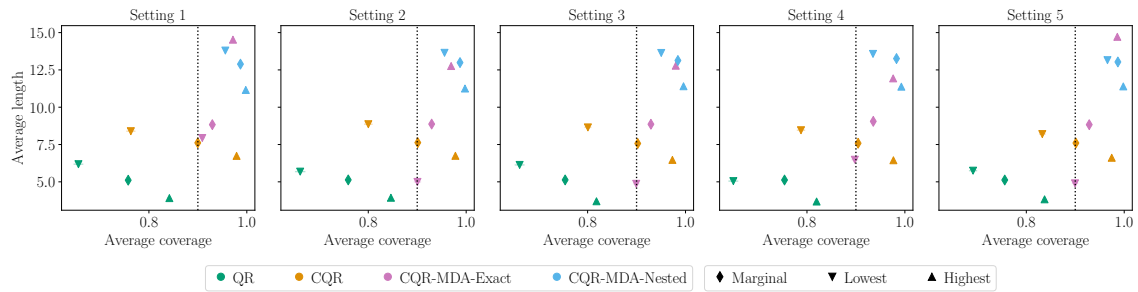
- MNAR self-masked (Figure 8.6). Missing values in X_{missing} are introduced under a MNAR self masked mechanism. To do so, a logistic model determines the probability of each variable in X_{missing} to be missing by taking as input the exact same variable. This setting is declined 5 times, with different weights for the logistic model. Within each one, the experiments are repeated 100 times to assess for the variability.

- MNAR quantile censorship (Figure 8.7). Missing values in X_{missing} are introduced under a quantile censorship MNAR mechanism. In particular, missing values are introduced at random in each q -quantile of the variables in X_{missing} . q varies between 0.5, 0.75, 0.8, 0.85, 0.9 and 0.95 and this way we obtain 6 different settings. Within each one, the experiments are repeated 100 times to assess for the variability.

These experiments show that impute-then-CQR is marginally valid even under \mathcal{P}_{MAR} and $\mathcal{P}_{\text{MNAR}}$. This is expected due to Proposition 3.3 of [Zaffran et al. \(2023\)](#), that demonstrates that vanilla impute-then-SplitCP is marginally valid for any missing mechanism as long as the initial data set is exchangeable. However, it is not MCV, which is also expected for the same reason that the fact that it is not MCV under $\mathcal{P}_{\text{MCAR}, \text{YIM}|X}$. Most importantly, **CP-MDA-Nested***, through **CP-MDA-Exact** and **CP-MDA-Nested**, is both marginally valid and MCV, despite the MCAR assumption not being satisfied, even when the imputation can not retrieve more information than the features expectation (i.e., when $\varphi = 0$). This is a positive empirical result that hints robustness of **CP-MDA-Nested*** on more complex relationships between X and M than independence.

(a) Dependent Gaussian features, with $\varphi = 0.8$.

(b) Independent Gaussian features.

Figure 8.5: Same caption than Figure 8.4, for **MAR missing values**, each panel representing a different setting (set of parameters) for the missingness distribution.(a) Dependent Gaussian features, with $\varphi = 0.8$.

(b) Independent Gaussian features.

Figure 8.6: Same caption than Figure 8.5, for **MNAR self masked missing values**.

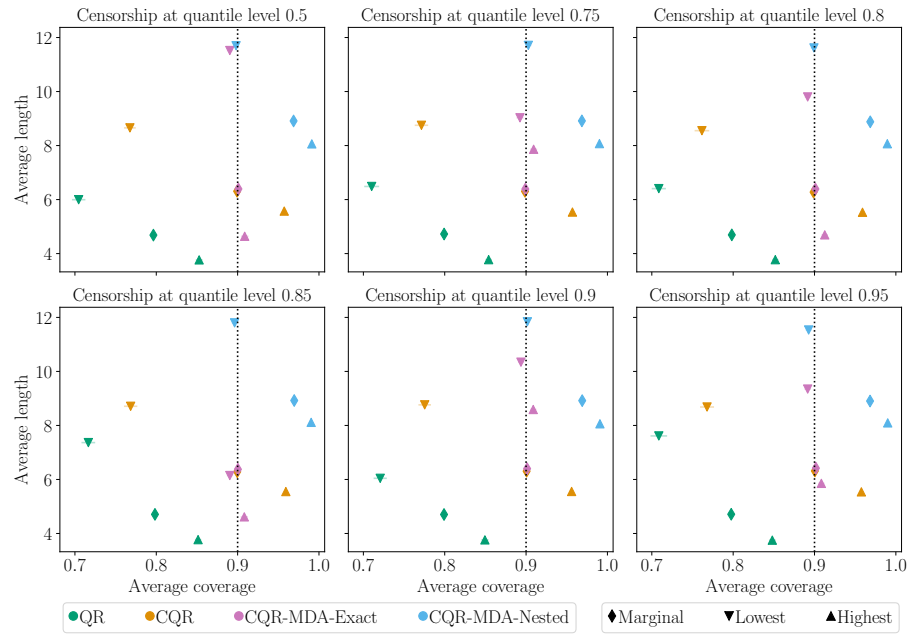
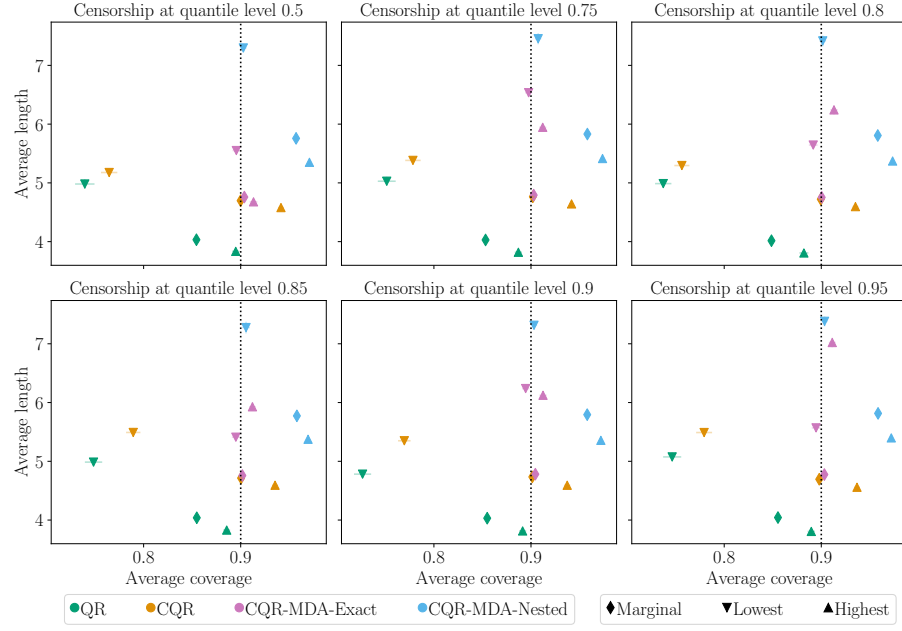
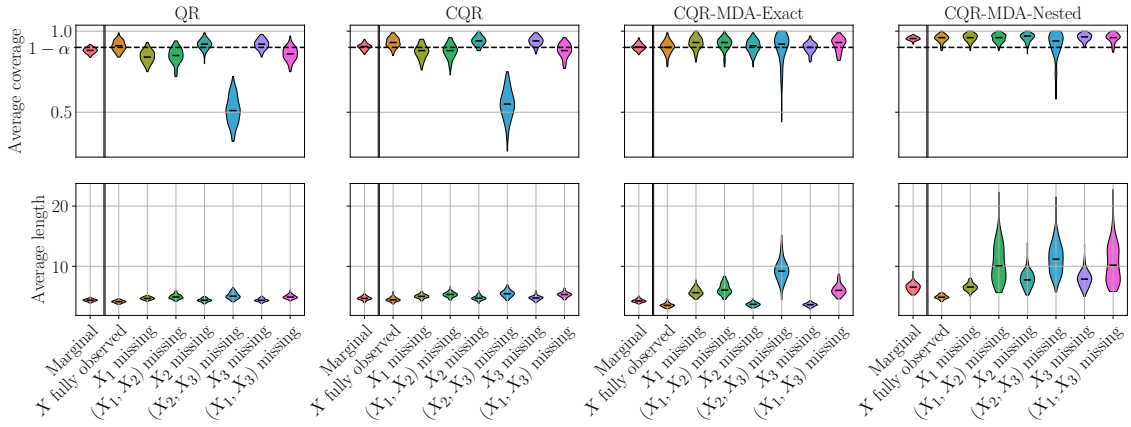
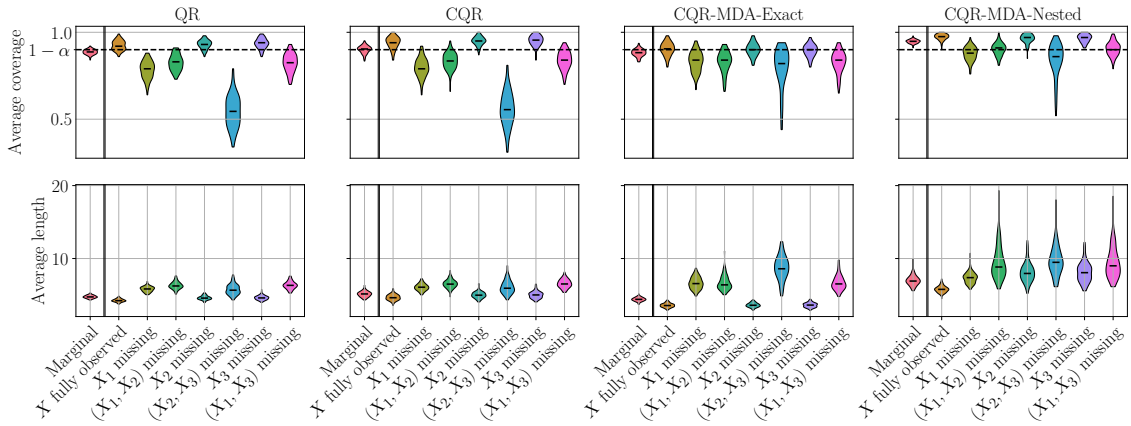


Figure 8.7: Same caption than Figure 8.5, for MNAR quantile censorship missing values.

8.5.3 Breaking $Y \perp\!\!\!\perp M \mid X$ Assumption

Our last set experiments aim at breaking the $Y \perp\!\!\!\perp M \mid X$ assumption. We focus on $d = 3$ to be able to display all of the patterns and thus better illustrate the phenomenon. We generate data with $\varepsilon \sim \mathcal{N}(0, 1) \perp\!\!\!\perp (X, M)$, $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 1)^T$, $\Sigma = \varphi(1, 1, 1)^T(1, 1, 1) + (1 - \varphi)I_d$, $\varphi \in \{0, 0.8\}$ depending on the experiment, and $M_i \sim \mathcal{B}(0.2)$ for any $i \in \llbracket 1, 3 \rrbracket$, independently from X and ε . Finally: $Y = X_1 \mathbb{1}\{M_1 = 0\} + 2X_1 \mathbb{1}\{M_1 = 1\} + 3X_2 \mathbb{1}\{M_2 = 1, M_3 = 1\} + \varepsilon$. Note that according to this data generation process, the masks for which at least X_1 is missing, and the mask where X_2 and X_3 are missing, have important predictive power. As there are only 3 features that can be missing in this setting, Figures 8.8a and 8.8b represent the 7 different missing patterns.

These figures highlight that in the easiest setting where the conditional expectation imputation is able to reconstruct the missing values quite accurately ($\varphi = 0.8$, Figure 8.8a) CP-MDA-Nested* manages to maintain MCV. However, in the hardest case of uncorrelated features ($\varphi = 0$, Figure 8.8b), it does not achieve MCV as it undercovers on the masks that have predictive power. Yet, CP-MDA-Nested* still improves upon vanilla impute-then-predict+CQR, and in particular CP-MDA-Nested is slightly more robust than CP-MDA-Exact.

(a) Dependent Gaussian features, with $\varphi = 0.8$.

(b) Independent Gaussian features.

Figure 8.8: Y and M are not independent given X , and the features are Gaussian dependent with $\varphi = 0.8$. Average coverage (top) and length (bottom) as a function of the missing patterns. The first violin plot shows the marginal coverage. The marginal test set includes 2000 observations. The mask-conditional test set includes 100 individuals for each missing data pattern.

Appendix to Predictive Uncertainty Quantification with Missing Covariates

The appendices are organized as follows.

Section 8.A provides a the proofs for the hardness results presented in Section 8.2.

Section 8.B contains the proofs of the Section 8.3 results.

Section 8.C reminds the proof of leave-one-out CP in the case of randomized algorithms.

Section 8.D derives CP-MDA-Nested* theoretical validities proofs, marginal and conditional.

8.A Hardness results

8.A.1 Most general distribution-free result: Theorem 8.2.1

Proof. Let $n \in \mathbb{N}^*$ the total training size (proper training and calibration).

Let $\alpha \in]0, 1[$.

Let $\hat{C}_{n,\alpha}$ be MCV, as defined in Definition 8.2.1.

Let P a distribution on $\mathcal{X} \times \mathcal{M} \times \mathcal{Y}$.

Let $m_0 \in \mathcal{M}$.

Denote by $\rho := P_M(\{m_0\})$.

\hookrightarrow Regression case.

Let $D > 0$.

Define Q another distribution on $\mathcal{X} \times \mathcal{M} \times \mathcal{Y}$ such that for any $A \subseteq \mathcal{X}$, for any $L \subseteq \mathcal{M}$ and for any $B \subseteq \mathcal{Y}$:

$$Q(A \times L \times B) := P(A \times L \setminus \{m_0\} \times B) + P_{(X,M)}(A \times \{m_0\}) R(B),$$

with R defined on \mathcal{Y} , uniform on $[-D; D]$.

Recall that the total variation distance between two probability distributions on \mathcal{Z} , say P and Q , is defined as: $TV(P, Q) := \sup_{Z \in \mathcal{Z}} |P(Z) - Q(Z)|$.

On the one hand, by construction, $TV(P, Q) \leq P_M(\{m_0\}) = \rho$. Hence, using Lemma 8.A.1: $TV(P^{\otimes(n+1)}, Q^{\otimes(n+1)}) \leq \sqrt{2 \left(1 - \left(1 - \frac{\rho^2}{2}\right)^{n+1}\right)}$. Therefore, for any

$A \subseteq \mathcal{X}$, for any $L \subseteq \mathcal{M}$ and for any $B \subseteq \mathcal{Y}$:

$$P^{\otimes(n+1)}(A \times L \times B) \geq Q^{\otimes(n+1)}(A \times L \times B) - \sqrt{2 \left(1 - \left(1 - \frac{\rho^2}{2}\right)^{n+1}\right)}. \quad (8.3)$$

On the other hand, as $\hat{C}_{n,\alpha}$ is MCV, it satisfies:

$$\begin{aligned} 1 - \alpha &\leq \mathbb{P}_{Q^{\otimes(n+1)}} \left(Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \mid M^{(n+1)} = m_0 \right) \\ &= \mathbb{E}_{Q^{\otimes(n+1)}} \left[\mathbb{1} \left\{ Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right\} \mid M^{(n+1)} = m_0 \right] \\ &= \mathbb{E}_{Q^{\otimes(n)}} \left[\mathbb{E}_Q \left[\mathbb{1} \left\{ Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right\} \right. \right. \\ &\quad \left. \left. \mid M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)} \right)_{k=1}^n \right] \right] \\ &= \mathbb{E}_{Q^{\otimes(n)}} \left[\mathbb{E}_Q \left[\mathbb{1} \left\{ Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right\} \right. \right. \\ &\quad \left. \left. \mid X^{(n+1)}, M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)} \right)_{k=1}^n \right] \right. \\ &\quad \left. \mid M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)} \right)_{k=1}^n \right] \\ &= \mathbb{E}_{Q^{\otimes(n)}} \left[\mathbb{E}_Q \left[\int_{\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right)} q \left(y \mid X^{(n+1)}, m_0 \right) dy \right. \right. \\ &\quad \left. \left. \mid M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)} \right)_{k=1}^n \right] \right] \\ &= \mathbb{E}_{Q^{\otimes(n)}} \left[\mathbb{E}_Q \left[\Lambda \left(\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \cap [-D; D] \right) \times \frac{1}{2D} \right. \right. \\ &\quad \left. \left. \mid M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)} \right)_{k=1}^n \right] \right] \\ &= \mathbb{E}_{Q^{\otimes(n+1)}} \left[\Lambda \left(\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \cap [-D; D] \right) \times \frac{1}{2D} \mid M^{(n+1)} = m_0 \right] \end{aligned}$$

Note that $\Lambda \left(\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \cap [-D; D] \right) \times \frac{1}{2D} \leq 1$ almost surely. Therefore, using Lemma 8.A.2, for any $t > 0$:

$$\begin{aligned} \mathbb{P}_{Q^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \cap [-D; D] \right) \times \frac{1}{2D} \geq 1 - t \right) &\geq 1 - \frac{\alpha}{t} \\ \mathbb{P}_{Q^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \cap [-D; D] \right) \geq (1 - t)2D \right) &\geq 1 - \frac{\alpha}{t} \\ \Rightarrow \mathbb{P}_{Q^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right) \geq (1 - t)2D \right) &\geq 1 - \frac{\alpha}{t}. \end{aligned}$$

Let $t = 1 - \frac{1}{\sqrt{D}}$ and obtain $\mathbb{P}_{Q^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right) \geq 2\sqrt{D} \right) \geq 1 - \frac{\alpha}{1 - \frac{1}{\sqrt{D}}}$.

Combining with Equation (8.3), we finally get:

$$\mathbb{P}_{P^{\otimes(n+1)}} \left(\Lambda \left(\hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right) \geq 2\sqrt{D} \right) \geq 1 - \frac{\alpha}{1 - \frac{1}{\sqrt{D}}} - \sqrt{2 \left(1 - \left(1 - \frac{\rho^2}{2}\right)^{n+1}\right)}.$$

Letting $D \rightarrow +\infty$, the result is proven.

\hookrightarrow Classification case.

Let $y \in \mathcal{Y}$.

Define Q another distribution on $\mathcal{X} \times \mathcal{M} \times \mathcal{Y}$ such that for any $A \subseteq \mathcal{X}$, for any $L \subseteq \mathcal{M}$ and for any $B \subseteq \mathcal{Y}$:

$$Q(A \times L \times B) := P(A \times L \setminus \{m_0\} \times B) + P_{(X,M)}(A \times \{m_0\}) S(B),$$

with S defined on \mathcal{Y} , being null everywhere except on y (a dirac in y).

On the one hand, exactly as in the regression case, by construction, $TV(P, Q) \leq P_X(E) \leq P_M(m_0) = \rho$. $TV(P^{\otimes(n+1)}, Q^{\otimes(n+1)}) \leq \sqrt{2 \left(1 - \left(1 - \frac{\rho^2}{2}\right)^{n+1}\right)}$. Therefore, for any $A \subseteq \mathcal{X}$, for any $L \subseteq \mathcal{M}$ and for any $B \subseteq \mathcal{Y}$:

$$P^{\otimes(n+1)}(A \times L \times B) \geq Q^{\otimes(n+1)}(A \times L \times B) - \sqrt{2 \left(1 - \left(1 - \frac{\rho^2}{2}\right)^{n+1}\right)}. \quad (8.3)$$

On the other hand, as $\hat{C}_{n,\alpha}$ is MCV, it satisfies:

$$\begin{aligned} 1 - \alpha &\leq \mathbb{P}_{Q^{\otimes(n+1)}} \left(Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \mid M^{(n+1)} = m_0 \right) \\ &= \mathbb{E}_{Q^{\otimes(n)}} \left[\mathbb{E}_Q \left[\mathbb{1} \left\{ Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right\} \mid M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)} \right)_{k=1}^n \right] \right] \\ &= \mathbb{E}_{Q^{\otimes(n)}} \left[\mathbb{E}_Q \left[\mathbb{1} \left\{ y \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right\} \mid M^{(n+1)} = m_0, \left(X^{(k)}, M^{(k)}, Y^{(k)} \right)_{k=1}^n \right] \right] \\ &= \mathbb{E}_{Q^{\otimes(n+1)}} \left[\mathbb{1} \left\{ y \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right\} \right] \\ &= \mathbb{P}_{Q^{\otimes(n+1)}} \left(y \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right). \end{aligned}$$

Combining with Equation (8.3), we finally get:

$$\mathbb{P}_{P^{\otimes(n+1)}} \left(y \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \right) \geq 1 - \alpha - \sqrt{2 \left(1 - \left(1 - \frac{\rho^2}{2}\right)^{n+1}\right)}$$

which concludes the proof for the classification case. □

The proof of Theorem 8.2.1 relied on the following Lemmas 8.A.1 and 8.A.2.

Lemma 8.A.1. *For P and Q two probability distributions, and $n \in \mathbb{N}^*$, it holds:*

$$TV(P^n, Q^n) \leq \sqrt{2 \left(1 - \left(1 - \frac{TV(P, Q)^2}{2}\right)^n\right)}.$$

Proof. The proof of this lemma is based on the relationship between the total variation distance and the Hellinger distance between two probability distributions denoted by $H(\cdot, \cdot)$ (see Tsybakov, 2009).

Let $n \in \mathbb{N}^*$ and let P and Q be two probability distributions.

On the one hand, note that:

$$TV(P, Q) \leq H(P, Q). \quad (8.4)$$

On the other hand, observe that:

$$H^2(P^n, Q^n) = 2 \left(1 - \left(1 - \frac{H^2(P, Q)}{2} \right)^n \right). \quad (8.5)$$

Therefore, by combining Equations (8.4) and (8.5) (that can be found in [Tsybakov, 2009](#)), we obtain the desired result. \square

Lemma 8.A.2. *Let W be a random variable such that $0 \leq W \leq 1$ and $\mathbb{E}[W] \geq \beta$ with $\beta \in [0, 1]$.*

Then, for any $t > 0$, it holds $\mathbb{P}(W \geq 1 - t) \geq 1 - \frac{1-\beta}{t}$.

Proof. Let $t > 0$.

As $W \leq 1$, $1 - W \geq 0$. Therefore, using Markov's inequality:

$$\mathbb{P}(1 - W \geq t) \leq \frac{\mathbb{E}[1 - W]}{t} = \frac{1 - \mathbb{E}[W]}{t} \leq \frac{1 - \beta}{t}$$

Noting that:

$$\mathbb{P}(1 - W \geq t) = \mathbb{P}(W \leq 1 - t) = 1 - \mathbb{P}(W \geq 1 - t),$$

we finally get $\mathbb{P}(W \geq 1 - t) \geq 1 - \frac{1-\beta}{t}$. \square

8.A.2 Restricting to $\mathcal{P}_{\text{YIM}|\mathcal{X}}$: Proposition 8.2.2

Proof. The skeleton of the proof is the exactly the same than the one of Theorem 8.2.1, with a careful attention required in the construction of the adversarial distribution Q .

Let $n \in \mathbb{N}^*$ the total training size (proper training and calibration).

Let $\alpha \in]0, 1[$.

Let $\hat{C}_{n,\alpha}$ be MCV- $\mathcal{P}_{\text{YIM}|\mathcal{X}}^{\otimes(n+1)}$.

Let $P \in \mathcal{P}_{\text{YIM}|\mathcal{X}}$.

Let $(X, M, Y) \sim P$.

Let $m_0 \in \mathcal{M}$ such that $\rho := P_M(\{m_0\}) > 0$.

\hookrightarrow Regression case.

Let $D > 0$.

We will now define Q another distribution on $\mathcal{X} \times \mathcal{M} \times \mathcal{Y}$ which is:

- (i) close in total variation to P with respect to ρ ;
- (ii) such that Assumption A1 holds (to ensure that $\hat{C}_{n,\alpha}$ is also MCV under Q);
- (iii) such that there exists some subset of \mathcal{X} , say F_0 , which determines the event of drawing mask m_0 under Q . This allows to remark that

$$\begin{aligned} & \mathbb{P}_{Q^{\otimes(n+1)}} \left(Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \mid M^{(n+1)} = m_0 \right) \\ &= \mathbb{P}_{Q^{\otimes(n+1)}} \left(Y^{(n+1)} \in \hat{C}_{n,\alpha} \left(X^{(n+1)}, m_0 \right) \mid X^{(n+1)} \in F_0 \right). \end{aligned}$$

Let $(\tilde{X}, \tilde{M}, \tilde{Y}) \sim Q$. Q is built in the following way.

Let $F_0 \subseteq \mathcal{X}$ such that $P_X(F_0) = \rho$.

$$\begin{cases} \text{if } X \notin F_0 \text{ and } M \neq m_0 : (\tilde{X}, \tilde{M}, \tilde{Y}) = (X, M, Y), \\ \text{if } X \in F_0 \text{ or } M = m_0 : (\tilde{X}, \tilde{M}, \tilde{Y}) \sim \mathcal{U}(F_0) \times \delta_{m_0} \times \mathcal{U}([-D, D]). \end{cases}$$

Using this construction, the proof will follow as in Theorem 8.2.1. The only “tricky points” to check are (i), (ii), and (iii).

By construction, (iii) is directly satisfied.

Remark that by construction $\mathbb{P}\left((X, M, Y) \neq (\tilde{X}, \tilde{M}, \tilde{Y})\right) \leq 2\delta$ (the worst case scenario being if F_0 has been chosen such that $\mathbb{1}\{X \in F_0\} \mathbb{1}\{M = m_0\} \stackrel{a.s.}{=} 0$, leading to an equality in the previous equation). Therefore, using Lemma 8.A.3, we get that $TV(P, Q) \leq 2\delta$, therefore verifying (i).

The remaining task is to show that (ii) is satisfied. Let $B \in \mathcal{Y}$. We have:

$$\begin{aligned} \mathbb{P}\left(\tilde{Y} \in B | \tilde{X}, \tilde{M}\right) &= \begin{cases} \mathbb{P}(Y \in B | X, M) & \text{if } \tilde{X} \in F_0 \\ \Lambda(B \cap [-D; D]) \frac{1}{2D} & \text{if } \tilde{X} \notin F_0 \end{cases} \\ &= \begin{cases} \mathbb{P}(Y \in B | X) & \text{if } \tilde{X} \in F_0 \text{ as } P \text{ satisfies Assumption A1} \\ \Lambda(B \cap [-D; D]) \frac{1}{2D} & \text{if } \tilde{X} \notin F_0 \end{cases} \\ &= \mathbb{P}\left(\tilde{Y} \in B | \tilde{X}\right). \end{aligned}$$

\hookrightarrow Classification case.

The idea is as previously, except that, as in the other hardness results, we replace the uniform distribution by a Dirac. In particular, let $y \in \mathcal{Y}$.

Let $(\tilde{X}, \tilde{M}, \tilde{Y}) \sim Q$. Q is built in the following way.

Let $F_0 \subseteq \mathcal{X}$ such that $P_X(F_0) = \rho$.

$$\begin{cases} \text{if } X \notin F_0 \text{ and } M \neq m_0 : (\tilde{X}, \tilde{M}, \tilde{Y}) = (X, M, Y), \\ \text{if } X \in F_0 \text{ or } M = m_0 : (\tilde{X}, \tilde{M}, \tilde{Y}) \sim \mathcal{U}(F_0) \times \delta_{m_0} \times \delta_y. \end{cases}$$

The conclusion follows as in Theorem 8.2.1, since, as shown in the regression case above, Q is such that: (i) $TV(P, Q) \leq 2\rho$, (ii) Assumption A1 and (iii) holds by construction. \square

Lemma 8.A.3. Let \mathbb{P}_Z and $\mathbb{P}_{Z'}$ be two distributions for the random variables X and X' taking their value in \mathcal{Z} . $TV(\mathbb{P}_Z, \mathbb{P}_{Z'}) \leq \mathbb{P}(Z \neq Z')$.

Proof.

$$\begin{aligned}
TV(\mathbb{P}_Z, \mathbb{P}_{Z'}) &= \sup_{A \subseteq \mathcal{Z}} |\mathbb{P}_Z(A) - \mathbb{P}_{Z'}(A)| \\
&= \sup_{A \subseteq \mathcal{Z}} |\mathbb{E}[\mathbf{1}\{Z \in A\}] - \mathbb{E}[\mathbf{1}\{Z' \in A\}]| \\
&\leq \sup_{A \subseteq \mathcal{Z}} \mathbb{E}[|\mathbf{1}\{Z \in A\} - \mathbf{1}\{Z' \in A\}|] \\
&= \sup_{A \subseteq \mathcal{Z}} \mathbb{E}[|\mathbf{1}\{Z \in A\} - \mathbf{1}\{Z' \in A\}| \mathbf{1}\{Z \neq Z'\}] \\
&\leq \sup_{A \subseteq \mathcal{Z}} \mathbb{E}[\mathbf{1}\{Z \neq Z'\}] \\
&= \sup_{A \subseteq \mathcal{Z}} \mathbb{P}(Z \neq Z')
\end{aligned}$$

□

8.B Link between missing covariates and uncertainty

8.B.1 Proofs for Conditional Variance results

8.B.1.1 RESULTS UNDER $\mathcal{P}_{\text{MCAR}, \text{YIM}|\text{X}}$ (PROPOSITION 8.3.1)

Proof. Under the assumptions, $M \perp (Y, X)$, and thus for any m :

$$\begin{aligned}
\mathbb{E}[V(X_{\text{obs}(M)}, M) | M = m] &= \mathbb{E}[V(X_{\text{obs}(m)}, m) | M = m] \\
&= \mathbb{E}[V(X_{\text{obs}(m)}, m)] \\
&= \mathbb{E}[\text{Var}(Y | X_{\text{obs}(m)})]
\end{aligned}$$

Moreover, for any $m \subset m'$,

$$\begin{aligned}
\text{Var}(Y | X_{\text{obs}(m')}) &= \mathbb{E}[\text{Var}(Y | X_{\text{obs}(m)}) | X_{\text{obs}(m')}] + \text{Var}(\mathbb{E}[Y | X_{\text{obs}(m)}] | X_{\text{obs}(m')}) \\
&\geq \mathbb{E}[\text{Var}(Y | X_{\text{obs}(m)}) | X_{\text{obs}(m')}]
\end{aligned}$$

Thus $\mathbb{E}[\text{Var}(Y | X_{\text{obs}(m')})] \geq \mathbb{E}[\text{Var}(Y | X_{\text{obs}(m)})]$. And finally:

$$\mathbb{E}[V(X_{\text{obs}(M)}, M) | M = m'] \geq \mathbb{E}[V(X_{\text{obs}(M)}, M) | M = m].$$

□

8.B.1.2 RESULTS UNDER GAUSSIAN LINEAR MODEL AND $\mathcal{P}_{\text{MCAR}}$

Previous works (Le Morvan et al., 2020b; Ayme et al., 2022; Zaffran et al., 2023) have shown that under Model 8.3.2, $Y | (X_{\text{obs}(m)}, M = m) \sim \mathcal{N}(\tilde{\mu}^m, \tilde{\sigma}^m)$ for any $m \in \mathcal{M}$, with:

$$\begin{aligned}
\tilde{\mu}^m &= \beta_{\text{obs}(m)}^T X_{\text{obs}(m)} + \beta_{\text{mis}(m)}^T \mu_{\text{mis}|\text{obs}}^m \\
\mu_{\text{mis}|\text{obs}}^m &= \mu_{\text{mis}(m)}^m + \Sigma_{\text{mis}(m), \text{obs}(m)}^m (\Sigma_{\text{obs}(m), \text{obs}(m)}^m)^{-1} (X_{\text{obs}(m)} - \mu_{\text{obs}(m)}^m), \\
\tilde{\sigma}^m &= \beta_{\text{mis}(m)}^T \Sigma_{\text{mis}|\text{obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2 \\
\Sigma_{\text{mis}|\text{obs}}^m &= \Sigma_{\text{mis}(m), \text{mis}(m)}^m - \Sigma_{\text{mis}(m), \text{obs}(m)}^m (\Sigma_{\text{obs}(m), \text{obs}(m)}^m)^{-1} \Sigma_{\text{obs}(m), \text{mis}(m)}^m.
\end{aligned}$$

We now provide the proof of Proposition 8.3.2.

Proof. Consider Model 8.3.2 and assume additionally that the missing mechanism is MCAR. Therefore, for any $m \in \mathcal{M}$, $\Sigma^m = \Sigma$. Hence, for any $m \in \mathcal{M}$:

$$\text{Var}(Y|X_{\text{obs}(m)}, M = m) = \beta_{\text{mis}(m)}^T \Sigma_{\text{mis}|\text{obs}}^m \beta_{\text{mis}(m)} + \sigma_\varepsilon^2,$$

with $\Sigma_{\text{mis}|\text{obs}}^m = \Sigma_{\text{mis}(m), \text{mis}(m)} - \Sigma_{\text{mis}(m), \text{obs}(m)} (\Sigma_{\text{obs}(m), \text{obs}(m)})^{-1} \Sigma_{\text{obs}(m), \text{mis}(m)}$.

Let $(m, m') \in \mathcal{M}^2$ such that $m \subseteq m'$. Our goal is to show that:

$$\begin{aligned} \text{Var}(Y|X_{\text{obs}(m')}, M = m') - \text{Var}(Y|X_{\text{obs}(m)}, M = m) &\geq 0 \\ \beta_{\text{mis}(m')}^T \Sigma_{\text{mis}|\text{obs}}^{m'} \beta_{\text{mis}(m')} + \sigma_\varepsilon^2 - \beta_{\text{mis}(m)}^T \Sigma_{\text{mis}|\text{obs}}^m \beta_{\text{mis}(m)} - \sigma_\varepsilon^2 &\geq 0 \\ \beta_{\text{mis}(m')}^T \Sigma_{\text{mis}|\text{obs}}^{m'} \beta_{\text{mis}(m')} - \beta_{\text{mis}(m)}^T \Sigma_{\text{mis}|\text{obs}}^m \beta_{\text{mis}(m)} &\geq 0 \\ \beta_{\text{mis}(m')}^T \Sigma_{\text{mis}|\text{obs}}^{m'} \beta_{\text{mis}(m')} - \beta_{\text{mis}(m')}^T \begin{pmatrix} \Sigma_{\text{mis}|\text{obs}}^m & 0 \\ 0 & \mathbf{0} \end{pmatrix} \beta_{\text{mis}(m')} &\geq 0 \\ \beta_{\text{mis}(m')}^T \left(\Sigma_{\text{mis}|\text{obs}}^{m'} - \begin{pmatrix} \Sigma_{\text{mis}|\text{obs}}^m & 0 \\ 0 & \mathbf{0} \end{pmatrix} \right) \beta_{\text{mis}(m')} &\geq 0, \end{aligned}$$

holds for any β . Therefore, we have to show that $\Sigma_{\text{mis}|\text{obs}}^{m'} - \begin{pmatrix} \Sigma_{\text{mis}|\text{obs}}^m & 0 \\ 0 & \mathbf{0} \end{pmatrix}$ is semi-definite positive.

The marginal covariance matrix Σ can be rewritten by blocks in the following way:

$$\Sigma = \begin{pmatrix} A & B & C \\ B^T & D & E \\ C^T & E^T & F \end{pmatrix},$$

where:

$$\begin{cases} A = \Sigma_{\text{mis}(m), \text{mis}(m)}, \\ \begin{pmatrix} D & E \\ E^T & F \end{pmatrix} = \Sigma_{\text{obs}(m), \text{obs}(m)}, \\ \begin{pmatrix} A & B \\ B^T & D \end{pmatrix} = \Sigma_{\text{mis}(m'), \text{mis}(m')}, \\ F = \Sigma_{\text{obs}(m'), \text{obs}(m')}. \end{cases}$$

Additionally, assume that $\Sigma > 0$ (that is, Σ is definite positive)

Therefore, $D > 0, F > 0$. Thus F is invertible, of inverse $F^{-1} > 0$. Furthermore, $G := D - EF^{-1}E^T$ is also positive definite, as it is the sum of $D > 0$ and $EF^{-1}E^T \geq 0$, and thus G is invertible.

$\Sigma_{\text{mis}|\text{obs}}^m$ and $\Sigma_{\text{mis}|\text{obs}}^{m'}$ can be rewritten using the previous decomposition.

On the one hand, for m it gives:

$$\begin{aligned} \Sigma_{\text{mis}|\text{obs}}^m &= A - \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} D & E \\ E^T & F \end{pmatrix}^{-1} \begin{pmatrix} B^T \\ C^T \end{pmatrix} \\ &= A - \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} G^{-1} & -G^{-1}EF^{-1} \\ -F^{-1}E^TG^{-1} & F^{-1} + F^{-1}E^TG^{-1}EF^T \end{pmatrix} \begin{pmatrix} B^T \\ C^T \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
&= A - \begin{pmatrix} B & C \end{pmatrix} \begin{pmatrix} G^{-1}B^T - G^{-1}EF^{-1}C^T \\ -F^{-1}E^TG^{-1}B^T + F^{-1}C^T + F^{-1}E^TG^{-1}EF^TC^T \end{pmatrix} \\
&= A - BG^{-1}B^T + BG^{-1}EF^{-1}C^T \\
&\quad + CF^{-1}E^TG^{-1}B^T - CF^{-1}C^T - CF^{-1}E^TG^{-1}EF^TC^T \\
&\text{(rearranging)} = A - CF^{-1}C^T - BG^{-1}B^T + BG^{-1}EF^{-1}C^T \\
&\quad + CF^{-1}E^TG^{-1}B^T - CF^{-1}E^TG^{-1}EF^TC^T \\
&= A - CF^{-1}C^T - BG^{-1}(B^T - EF^{-1}C^T) \\
&\quad + CF^{-1}E^TG^{-1}(B^T - EF^TC^T) \\
&= A - CF^{-1}C^T - (B - CF^{-1}E^T)G^{-1}(B^T - EF^{-1}C^T),
\end{aligned}$$

and by denoting $H := B - CF^{-1}E^T$, we finally obtain (as F is symmetric):

$$\Sigma_{\text{mis}|\text{obs}}^m = A - CF^{-1}C^T - HG^{-1}H^T.$$

On the other hand, for m' :

$$\begin{aligned}
\Sigma_{\text{mis}|\text{obs}}^{m'} &= \begin{pmatrix} A & B \\ B^T & D \end{pmatrix} - \begin{pmatrix} C \\ E \end{pmatrix} F^{-1} \begin{pmatrix} C^T & E^T \end{pmatrix} \\
&= \begin{pmatrix} A & B \\ B^T & D \end{pmatrix} - \begin{pmatrix} CF^{-1}C^T & CF^{-1}E^T \\ EF^{-1}C^T & EF^{-1}E^T \end{pmatrix} \\
&= \begin{pmatrix} A - CF^{-1}C^T & B - CF^{-1}E^T \\ B^T - EF^{-1}C^T & D - EF^{-1}E^T \end{pmatrix} \\
&= \begin{pmatrix} A - CF^{-1}C^T & B - CF^{-1}E^T \\ B^T - EF^{-1}C^T & G \end{pmatrix} \\
\Sigma_{\text{mis}|\text{obs}}^{m'} &= \begin{pmatrix} A - CF^{-1}C^T & H \\ H^T & G \end{pmatrix}
\end{aligned}$$

Therefore, combining the two terms and rewriting together, we obtain:

$$\begin{aligned}
\Sigma_{\text{mis}|\text{obs}}^{m'} - \begin{pmatrix} \Sigma_{\text{mis}|\text{obs}}^m & 0 \\ 0 & \mathbf{0} \end{pmatrix} &= \begin{pmatrix} A - CF^{-1}C^T & H \\ H^T & G \end{pmatrix} - \begin{pmatrix} A - CF^{-1}C^T - HG^{-1}H^T & 0 \\ 0 & \mathbf{0} \end{pmatrix} \\
&= \begin{pmatrix} A - CF^{-1}C^T - A + CF^{-1}C^T + HG^{-1}H^T & H \\ H^T & G \end{pmatrix} \\
\Sigma_{\text{mis}|\text{obs}}^{m'} - \begin{pmatrix} \Sigma_{\text{mis}|\text{obs}}^m & 0 \\ 0 & \mathbf{0} \end{pmatrix} &= \begin{pmatrix} HG^{-1}H^T & H \\ H^T & G \end{pmatrix}.
\end{aligned}$$

Hence, our objective is to show that $\begin{pmatrix} HG^{-1}H^T & H \\ H^T & G \end{pmatrix}$ is semi-definite positive.

Let $z = \begin{pmatrix} x & y \end{pmatrix} \in \mathbb{R}^{1 \times (\#m + (\#m' - \#m))}$.

$$z \begin{pmatrix} HG^{-1}H^T & H \\ H^T & G \end{pmatrix} z^T = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} HG^{-1}H^T & H \\ H^T & G \end{pmatrix} \begin{pmatrix} x^T \\ y^T \end{pmatrix}$$

$$\begin{aligned}
&= xHG^{-1}H^Tx^T + xHy^T + yH^Tx^T + yGy^T \\
&= xHG^{-1}GG^{-1}H^Tx^T + xHG^{-1}Gy^T + yGG^{-1}H^Tx^T + yGy^T \\
&= xHG^{-1}G(G^{-1}H^Tx^T + y^T) + yG(G^{-1}H^Tx^T + y^T) \\
&= (xHG^{-1} + y)G(G^{-1}H^Tx^T + y^T) \\
&= (xHG^{-1} + y)G(xHG^{-1} + y)^T \\
&\geq 0 \text{ as } G \text{ is positive definite.}
\end{aligned}$$

□

8.B.2 Impact of the imputation under a linear quantile regression model (Proposition 8.3.3)

To prove Item i) of Proposition 8.3.3, we prove the following Lemma 8.B.1.

Lemma 8.B.1. Assume $\mathcal{P}_{\text{MCAR}}$, and $Y = \beta^{*T}X + \varepsilon$ with ε s.t. $\mathbb{E}[\varepsilon|X_{\text{obs}(M)}, M] = 0$.

Then $\mathbb{E}[Y|X_{\text{obs}(M)}, M] = \beta^{*T}\Phi_{\text{conditional mean}}(X, M)$, with $\Phi_{\text{conditional mean}}$ the imputation by the conditional mean. Furthermore, if the covariates are independent, then $\mathbb{E}[Y|X_{\text{obs}(M)}, M] = \beta^{*T}\Phi_{\text{mean}}(X, M)$, with Φ_{mean} the imputation by the mean.

Proof.

$$\begin{aligned}
\mathbb{E}[Y|X_{\text{obs}(M)}, M] &= \mathbb{E}[\beta^{*T}X|X_{\text{obs}(M)}, M] = \sum_{i=1}^d \beta_i^* \mathbb{E}[X_i|X_{\text{obs}(M)}, M] \\
&= \sum_{i=1}^d \beta_i^* (X_i \mathbb{1}\{i \in \text{obs}(M)\} + \mathbb{E}[X_i|X_{\text{obs}(M)}, M] \mathbb{1}\{i \notin \text{obs}(M)\}) \\
\mathcal{P}_{\text{MCAR}} \rightarrow &= \sum_{i=1}^d \beta_i^* (X_i \mathbb{1}\{i \in \text{obs}(M)\} + \mathbb{E}[X_i|X_{\text{obs}(M)}] \mathbb{1}\{i \notin \text{obs}(M)\}) \\
&= \sum_{i=1}^d \beta_i^* (\Phi_{\text{conditional mean}}(X, M))_i \\
\text{if } (X_i)_{i=1}^d \perp\!\!\!\perp, \mathbb{E}[X_i|X_{\text{obs}(M)}] &= \mathbb{E}[X_i] \rightarrow = \sum_{i=1}^d \beta_i^* (\Phi_{\text{mean}}(X, M))_i
\end{aligned}$$

□

To prove Item ii) of Proposition 8.3.3, we prove the following Proposition 8.B.1. Indeed, the oracle predictive intervals vary at least once in length we respect to the patterns, as, on the one hand, under $\mathcal{P}_{\text{MCAR}, \text{YIM}|X}$ Equation (Len-2) holds and, on the other hand, when $Y \not\perp X$ the variance of Y given X is different than the overall variance of Y .

Proposition 8.B.1 (Non-adaptivity of the linear quantile regression). Assume that:

- i) the quantile regression is learned within the class of linear models;
- ii) the (random) values used to impute have the same expectation than the feature itself, i.e., $\mathbb{E}[\Phi(X, m)|M = m] = \mathbb{E}[X]$ for any $m \in \mathcal{M}$ such that $\mathbb{P}(M = m) > 0$.

Then the expectation of the predictive intervals length is independent of the missing pattern.

Proof. Since the quantile regression is learned within the class of linear models, the fitted quantile functions (upper and lower) can be written as $\hat{q}_\delta(z) = \beta_\delta^T z + \beta_\delta^0$, with $\beta \in \mathbb{R}^d$ and $\beta^0 \in \mathbb{R}$. Therefore, the length of the resulting interval L_α at some—imputed—point $\Phi(X_{\text{obs}(M)}, M)$ will be:

$$\begin{aligned} L_\alpha(\Phi(X_{\text{obs}(M)}, M)) &:= \hat{q}_{\delta_{(u)}}(\Phi(X_{\text{obs}(M)}, M)) - \hat{q}_{\delta_{(l)}}(\Phi(X_{\text{obs}(M)}, M)) \\ &= \left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T \right) \Phi(X_{\text{obs}(M)}, M) + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0, \end{aligned}$$

with $\delta_{(l)}$ and $\delta_{(u)}$ chosen by the user or fixed by the algorithm such that $\delta_{(u)} - \delta_{(l)} = 1 - \alpha$. Thus:

$$\begin{aligned} \mathbb{E} [L_\alpha(\Phi(X_{\text{obs}(M)}, M))] &= \mathbb{E} \left[\left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T \right) \Phi(X_{\text{obs}(M)}, M) + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0 \right] \\ &= \left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T \right) \mathbb{E} [\Phi(X_{\text{obs}(M)}, M)] + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0. \end{aligned}$$

Let $m \in \mathcal{M}$ such that $\mathbb{P}(M = m) > 0$. Conditioning by m :

$$\mathbb{E} [L_\alpha(\Phi(X_{\text{obs}(M)}, M)) | M = m] = \left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T \right) \mathbb{E} [\Phi(X_{\text{obs}(M)}, M) | M = m] + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0.$$

Given the assumption that $\mathbb{E} [\Phi(X_{\text{obs}(M)}, M) | M = m] = \mathbb{E} [X]$, one can conclude that:

$$\mathbb{E} [L_\alpha(\Phi(X_{\text{obs}(M)}, M)) | M = m] = \sum_{j=1}^d \left(\beta_{\delta_{(u)}}^T - \beta_{\delta_{(l)}}^T \right)_j \mathbb{E} [X] + \beta_{\delta_{(u)}}^0 - \beta_{\delta_{(l)}}^0 \perp\!\!\!\perp M.$$

□

8.C Leave-one-out predictive sets for randomized algorithms

We provide in this section a more detailed proof of leave-one-out or k -fold cross-conformal (Vovk, 2015) and jackknife+ (Barber et al., 2021b) methods which allows us to highlight where exactly the arguments of data exchangeability and symmetrical algorithm play a role. In particular, by emphasizing these precise influences, we can understand how to include a non-deterministic symmetrical algorithm (such as Random Forest or Stochastic Gradient Descent).

8.C.1 On the definition of randomized symmetric algorithms

Definition 8.C.1 (Randomized learning algorithm). A randomized learning algorithm is defined as:

$$\mathcal{A} : \left(\bigcup_{n \geq 0} (\mathcal{X} \times \mathcal{Y})^n \right) \times [0, 1] \mapsto \mathcal{Y}^{\mathcal{X}} \\ \left(X^{(k)}, Y^{(k)} \right)_{k=1}^n \times \xi \mapsto \hat{A}(\cdot)$$

where ξ encodes the randomness of \mathcal{A} .

Definition 8.C.2 (Randomized symmetric algorithm (Kim and Barber, 2023)). A randomized learning algorithm \mathcal{A} is symmetric if for any data set $(X^{(k)}, Y^{(k)})_{k=1}^n$, for any permutation σ on $\llbracket 1, n \rrbracket$, there exists a coupling that maps $\xi \sim \mathcal{U}([0, 1])$ to $\xi' \sim \mathcal{U}([0, 1])$, which depends only on σ , s.t.:

$$\mathcal{A}\left(\left(X^{(k)}, Y^{(k)}\right)_{k=1}^n; \xi\right) = \mathcal{A}\left(\left(X^{(\sigma(k))}, Y^{(\sigma(k))}\right)_{k=1}^n; \xi'\right).$$

8.C.2 Detailing leave-one-out conformal predictors validity proof

Let $(X^{(k)}, Y^{(k)})_{k=1}^{n+1}$ be exchangeable, and \mathcal{A} a (possible randomized) symmetric algorithm.

Let s be a conformity score function. For $i \in \llbracket 1, n \rrbracket$, denote $\hat{A}_{-i}(\cdot) := \mathcal{A}\left(\left(X^{(k)}, Y^{(k)}\right)_{\substack{k=1 \\ k \neq i}}^n\right)$, that is the fitted left-one-out algorithm, removing data point i .

Consider the leave-one-out conformal estimator defined as:

$$\hat{C}_{n,\alpha}^{\text{LOO}}(x) := \left\{ y \in \mathcal{Y} : \sum_{k=1}^n \mathbb{1} \left\{ s\left(X^{(k)}, Y^{(k)}; \hat{A}_{-k}\right) < s\left(x, y; \hat{A}_{-k}\right) \right\} < (1 - \alpha)(n + 1) \right\}$$

Previous works (Barber et al., 2021b; Gupta et al., 2022) have proven that under exchangeability of $(X^{(k)}, Y^{(k)})_{k=1}^{n+1}$ and symmetry of \mathcal{A} , $\mathbb{P}\left(Y^{(n+1)} \in \hat{C}_{n,\alpha}^{\text{LOO}}(X^{(n+1)})\right) \geq 1 - 2\alpha$. We recall below the key proof's steps, detailing the last one which uses the exchangeability and symmetry arguments.

Step 1. Remark that:

$$\begin{aligned} & \left\{ Y^{(n+1)} \notin \hat{C}_{n,\alpha}^{\text{LOO}}(X^{(n+1)}) \right\} \\ &= \left\{ \sum_{k=1}^n \mathbb{1} \left\{ s\left(X^{(k)}, Y^{(k)}; \hat{A}_{-k}\right) < s\left(X^{(n+1)}, Y^{(n+1)}; \hat{A}_{-k}\right) \right\} \geq (1 - \alpha)(n + 1) \right\} \\ &:= \left\{ \sum_{k=1}^n \mathbb{1} \left\{ S^{(k),n+1} < S^{(n+1),k} \right\} \geq (1 - \alpha)(n + 1) \right\} \\ &:= \left\{ \sum_{k=1}^n \mathcal{C}_{n+1,k} \geq (1 - \alpha)(n + 1) \right\}. \end{aligned}$$

with $S^{(i),j} := s\left(X^{(i)}, Y^{(i)}; \hat{A}_{-(i,j)}\right)$ the score on data point i of the predictor that has been fitted without seeing nor data point i nor data point j , for $(i, j) \in \llbracket 1, n + 1 \rrbracket^2$ and extending \hat{A}_{-i} to $\hat{A}_{-(i,j)} := \mathcal{A}\left(\left(X^{(k)}, Y^{(k)}\right)_{\substack{k=1 \\ k \notin \{i,j\}}}^{n+1}\right)$, where the $n + 1$ data point is added.

Denote by $\mathcal{C}_{\mathcal{A}}$ the function building the comparison matrix $\mathcal{C} \in \{0, 1\}^{(n+1) \times (n+1)}$: $\mathcal{C}_{\mathcal{A}}\left(\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}\right)_{i,j} = \mathbb{1} \left\{ S^{(i),j} > S^{(j),i} \right\} = \mathcal{C}_{i,j}$.

Step 2. Deterministically, Barber et al. (2021b) shows that $\#\{i \in \llbracket 1, n + 1 \rrbracket : \sum_{j=1}^{n+1} \mathcal{C}_{i,j} \geq (1 - \alpha)(n + 1)\} \leq 2\alpha(n + 1)$. This is shown for *any* comparison matrix.

Step 3. The last (and crucial) step of leave-one-out conformal predictors is to show that for any permutation σ on $\llbracket 1, n+1 \rrbracket$ it holds: $(\mathcal{C}_{\sigma(i), \sigma(j)})_{i,j} \stackrel{d}{=} (\mathcal{C}_{i,j})_{i,j}$.

$$\begin{aligned}
\mathcal{C}_{\sigma(i), \sigma(j)} &= \mathcal{C}_{\mathcal{A}} \left(\left(X^{(k)}, Y^{(k)} \right)_{k=1}^{n+1} \right)_{\sigma(i), \sigma(j)} \\
&= \mathbb{1} \left\{ s \left(Y^{(\sigma(i))}, X^{(\sigma(i))}, \mathcal{A} \left(\left(X^{(k)}, Y^{(k)} \right)_{k=1, k \notin \{\sigma(i), \sigma(j)\}}^{n+1}; \xi \right) \right) \right. \\
&\quad \left. > s \left(Y^{(\sigma(j))}, X^{(\sigma(j))}, \mathcal{A} \left(\left(X^{(k)}, Y^{(k)} \right)_{k=1, k \notin \{\sigma(i), \sigma(j)\}}^{n+1}; \xi \right) \right) \right\} \\
&= \mathbb{1} \left\{ s \left(Y^{(\sigma(i))}, X^{(\sigma(i))}, \mathcal{A} \left(\left(X^{(\sigma(k))}, Y^{(\sigma(k))} \right)_{k=1, k \notin \{i, j\}}^{n+1}; \xi'_{\sigma} \right) \right) \right. \\
&\quad \left. > s \left(Y^{(\sigma(j))}, X^{(\sigma(j))}, \mathcal{A} \left(\left(X^{(\sigma(k))}, Y^{(\sigma(k))} \right)_{k=1, k \notin \{i, j\}}^{n+1}; \xi'_{\sigma} \right) \right) \right\} \quad \mathcal{A} \text{ is symmetric} \\
&= \mathcal{C}_{\mathcal{A}} \left(\left(X^{(\sigma(k))}, Y^{(\sigma(k))} \right)_{k=1}^{n+1} \right)_{i,j}
\end{aligned}$$

Thus, leveraging the fact that $\xi'_{\sigma} \perp (X^{(k)}, Y^{(k)})_{k=1}^{n+1}$ and that $(X^{(k)}, Y^{(k)})_{k=1}^{n+1}$ are exchangeable, we obtain that:

$$(\mathcal{C}_{\sigma(i), \sigma(j)})_{i,j \in \llbracket 1, n+1 \rrbracket^2} \stackrel{d}{=} \mathcal{C}_{\mathcal{A}} \left(\left(X^{(k)}, Y^{(k)} \right)_{k=1}^{n+1} \right) = (\mathcal{C}_{i,j})_{i,j \in \llbracket 1, n+1 \rrbracket^2}.$$

Hence, for any permutation σ on $\llbracket 1, n+1 \rrbracket$ it holds that $\Pi_{\sigma}^T \mathcal{C} \Pi_{\sigma} \stackrel{d}{=} \mathcal{C}$, concluding the proof as then each element of $\llbracket 1, n+1 \rrbracket$ is equally likely to belong to $\{i \in \llbracket 1, n+1 \rrbracket : \sum_{j=1}^{n+1} \mathcal{C}_{i,j} \geq (1-\alpha)(n+1)\}$.

8.D Theory on CP-MDA-Nested* and CP-MDA-Nested

Let us first remark that $\widehat{\mathcal{C}}_{n,\alpha}^{\text{MDA-Nested}^*}(\cdot) \subseteq \widehat{\mathcal{C}}_{n,\alpha}^{\text{MDA-Nested}}(\cdot)$ when the conformity score function outputs intervals and $\widetilde{\text{Cal}} = \text{Cal}$ (Remark 8.4.1).

Proof.

$$\begin{aligned}
&\left\{ Y^{(n+1)} \notin \widehat{\mathcal{C}}_{n,\alpha}^{\text{MDA-Nested}} \left(X^{(n+1)}, M^{(n+1)} \right) \right\} \\
&= \left\{ Y^{(n+1)} > \widehat{Q}_{1-\alpha} \left(\mathcal{U}_{\alpha} \left(X^{(n+1)} \right) \right) \right. \\
&\quad \left. \text{or } Y^{(n+1)} < \widehat{Q}_{\alpha} \left(\mathcal{L}_{\alpha} \left(X^{(n+1)} \right) \right) \right\} \\
&= \left\{ (1-\alpha)(\#\text{Cal} + 1) \leq \sum_{k=1}^n \mathbb{1} \left\{ Y^{(n+1)} > u_{\alpha}^{(k)} \left(X^{(n+1)} \right) \right\} \right. \\
&\quad \left. \text{or } (1-\alpha)(\#\text{Cal} + 1) \leq \sum_{k=1}^n \mathbb{1} \left\{ Y^{(n+1)} < \ell_{\alpha}^{(k)} \left(X^{(n+1)} \right) \right\} \right\}
\end{aligned}$$

$$\begin{aligned}
& \subset \left\{ (1 - \alpha)(\#\text{Cal} + 1) \leq \sum_{k=1}^n \mathbb{1} \left\{ Y^{(n+1)} > u_{\alpha}^{(k)} \left(X^{(n+1)} \right) \right. \right. \\
& \quad \left. \left. \text{or } Y^{(n+1)} < \ell_{\alpha}^{(k)} \left(X^{(n+1)} \right) \right\} \right\} \\
& = \left\{ (1 - \alpha)(\#\text{Cal} + 1) \right. \\
& \quad \leq \sum_{k=1}^n \mathbb{1} \left\{ s \left(\left(X^{(n+1)}, \widetilde{M}^{(k)} \right), Y^{(n+1)}; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) \right. \\
& \quad \quad \left. \left. > s \left(\left(X^{(k)}, \widetilde{M}^{(k)} \right), Y^{(k)}; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) \right\} \right\} \\
& = \left\{ Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*} \left(X^{(n+1)}, M^{(n+1)} \right) \right\}
\end{aligned}$$

□

Therefore, any upper bound on the miscoverage of CP-MDA-Nested* extends to CP-MDA-Nested.

8.D.1 Marginal validity of CP-MDA-Nested*.

The proof of Theorem 8.4.1 is highly inspired by the leave-one-out conformal predictors proof, from Barber et al. (2021b) and detailed previously in Section 8.C.

Proof. One can see this proof as analogous of the one of leave-one-out conformal predictors, where “predicting on point i with point j left out” corresponds to “predicting on point i when additionally masking it with the mask of point j ”.

Step 1.

$$\begin{aligned}
& \left\{ Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*} \left(X^{(n+1)}, M^{(n+1)} \right) \right\} \\
& = \left\{ (1 - \alpha)(\#\text{Cal} + 1) \right. \\
& \quad \leq \sum_{k \in \text{Cal}} \mathbb{1} \left\{ s \left(\left(X^{(n+1)}, \widetilde{M}^{(k)} \right), Y^{(n+1)}; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) \right. \\
& \quad \quad \left. \left. > s \left(\left(X^{(k)}, \widetilde{M}^{(k)} \right), Y^{(k)}; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) \right\} \right\} \\
& := \left\{ (1 - \alpha)(\#\text{Cal} + 1) \leq \sum_{k \in \text{Cal}} \mathbb{1} \left\{ S^{(n+1),k} > S^{(k),n+1} \right\} \right\},
\end{aligned}$$

where we defined $S^{(i),j} := s \left(\left(X^{(i)}, \max(M^{(i)}, M^{(j)}) \right), Y^{(i)}; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right)$, that is the score of the point i when the mask of the point j is applied to it, on top of its own mask $M^{(i)}$.

Step 2. Define the comparison matrix $\mathcal{C} \in \{0, 1\}^{(\#\text{Cal}+1) \times (\#\text{Cal}+1)}$, s.t. for $(i, j) \in (\text{Cal} \cup \{n+1\})^2$: $\mathcal{C}_{i,j} = \mathbb{1} \{ S^{(i),j} > S^{(j),i} \}$. Hence, we now have (since by definition $\mathcal{C}_{n+1,n+1} = 0$):

$$\left\{ Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*} \left(X^{(n+1)}, M^{(n+1)} \right) \right\} = \left\{ \sum_{k \in \text{Cal} \cup \{n+1\}} \mathcal{C}_{n+1,k} \geq (1 - \alpha)(\#\text{Cal} + 1) \right\}.$$

Denote $W(\mathcal{C}) = \{i \in \text{Cal} \cup \{n+1\} : \sum_{k \in \text{Cal} \cup \{n+1\}} \mathcal{C}_{i,k} \geq (1-\alpha)(\#\text{Cal}+1)\}$. We can re-write:

$$\left\{ Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*} \left(X^{(n+1)}, M^{(n+1)} \right) \right\} = \{n+1 \in W(\mathcal{C})\}.$$

Therefore $\mathbb{P} \left\{ Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*} \left(X^{(n+1)}, M^{(n+1)} \right) \right\} = \mathbb{P} \{n+1 \in W(\mathcal{C})\}$. Thus, we will now bound $\mathbb{P} \{n+1 \in W(\mathcal{C})\}$.

Again, $\#W(\mathcal{C}) \leq 2\alpha(\#\text{Cal}+1)$ deterministically (Barber et al., 2021b).

Step 3. To conclude the proof, observe that the matrix \mathcal{C} can be viewed as the output of a deterministic function \mathcal{C} of the exchangeable (by A2) sequence $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^{n+1}$: $\mathcal{C} = \mathcal{C} \left((X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^{n+1} \right)$.

Thus, for any permutation σ on $\text{Cal} \cup \{n+1\}$, it holds:

$$\mathcal{C} \left((X^{(k)}, M^{(k)}, Y^{(k)})_{k \in \text{Cal} \cup \{n+1\}} \right) \stackrel{d}{=} \mathcal{C} \left((X^{(\sigma(k))}, M^{(\sigma(k))}, Y^{(\sigma(k))})_{k \in \text{Cal} \cup \{n+1\}} \right) := \mathcal{C}^\sigma.$$

It follows that for any $k \in \text{Cal} \cup \{n+1\}$, $\mathbb{P}\{k \in W(\mathcal{C})\} = \mathbb{P}\{k \in W(\mathcal{C}^\sigma)\}$ for any permutation σ on $\text{Cal} \cup \{n+1\}$. Therefore $\mathbb{P}\{k \in W(\mathcal{C})\}$ does not depend on k . Finally:

$$\begin{aligned} \mathbb{P} \left\{ Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*} \left(X^{(n+1)}, M^{(n+1)} \right) \right\} &= \mathbb{P}\{n+1 \in W(\mathcal{C})\} \\ &= \frac{1}{\#\text{Cal}+1} \sum_{k \in \text{Cal} \cup \{n+1\}} \mathbb{P}\{k \in W(\mathcal{C})\} \\ &= \frac{1}{\#\text{Cal}+1} \mathbb{E}[\#W(\mathcal{C})] \\ &\leq \frac{1}{\#\text{Cal}+1} 2\alpha(\#\text{Cal}+1) = 2\alpha. \end{aligned}$$

□

8.D.2 MCV of CP-MDA-Nested*

To prove that CP-MDA-Nested* and CP-MDA-Nested are MCV- $\mathcal{P}_{\text{MCAR}, \text{YIM}}^{\otimes(n+1)} | \mathbf{X}$, we leverage again the parallel with leave-one-out conformal predictors, but this time seeing the missing pattern as exogenous randomness, which is possible when working with distributions in $\mathcal{P}_{\text{MCAR}, \text{YIM}} | \mathbf{X}$.

Proof. Under $\mathcal{P}_{\text{MCAR}, \text{YIM}}^{\otimes(n+1)} | \mathbf{X}$, it holds that $M^{(n+1)} \perp\!\!\!\perp \left((X^{(k)}, Y^{(k)})_{k \in \text{Cal}}, (X^{(n+1)}, Y^{(n+1)}) \right)$. Thus the sequence $\left\{ (X^{(k)}, M^{(n+1)}, Y^{(k)})_{k \in \text{Cal}}, (X^{(n+1)}, M^{(n+1)}, Y^{(n+1)}) \right\}$ is exchangeable conditionally to $M^{(n+1)}$.

Remark now that for any $(X, M, Y) \in \mathcal{X} \times \mathcal{M} \times \mathcal{Y}$, we can rewrite the score on this point with augmented mask $\widetilde{M} := \max(M, M^{(n+1)})$ as:

$$s \left((X, \widetilde{M}), Y; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) := s \left((X, M^{(n+1)}), Y; \widetilde{A} \left(\widetilde{\Phi}(\cdot, \cdot; M), \cdot; M \right) \right),$$

where, for an additional mask $M' \in \mathcal{M}$, $\widetilde{\Phi}(X, M; M') := \Phi(X, \max(M, M'))$ and similarly $\widetilde{A}(X, M; M') := \hat{A}(X, \max(M, M'))$.

Thus, we can re-write CP-MDA-Nested* as:

$$\begin{aligned}
& \left\{ Y^{(n+1)} \notin \widehat{C}_{n,\alpha}^{\text{MDA-Nested}^*} \left(X^{(n+1)}, M^{(n+1)} \right) \right\} \\
&= \left\{ (1 - \alpha)(\#\text{Cal} + 1) \right. \\
&\quad \leq \sum_{k \in \text{Cal}} \mathbb{1} \left\{ s \left(\left(X^{(n+1)}, \widetilde{M}^{(k)} \right), Y^{(n+1)}; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) \right. \\
&\quad \quad \left. \left. > s \left(\left(X^{(k)}, \widetilde{M}^{(k)} \right), Y^{(k)}; \hat{A}(\Phi(\cdot, \cdot), \cdot) \right) \right\} \right\} \\
&= \left\{ (1 - \alpha)(\#\text{Cal} + 1) \right. \\
&\quad \leq \sum_{k \in \text{Cal}} \mathbb{1} \left\{ s \left(\left(X^{(n+1)}, M^{(n+1)} \right), Y^{(n+1)}; \widetilde{A} \left(\widetilde{\Phi} \left(\cdot, \cdot; M^{(k)} \right), \cdot; M^{(k)} \right) \right) \right. \\
&\quad \quad \left. \left. > s \left(\left(X^{(k)}, M^{(n+1)} \right), Y^{(k)}; \widetilde{A} \left(\widetilde{\Phi} \left(\cdot, \cdot; M^{(k)} \right), \cdot; M^{(k)} \right) \right) \right\} \right\}.
\end{aligned}$$

Therefore, an equivalent rewriting of CP-MDA-Nested* is a specific instance of what is presented in Algorithm 18, where the differences with CP-MDA-Nested* (Algorithm 16) are highlighted through green text.

Algorithm 18 MDA based on random masks

Input: Imputation function Φ , fitted predictor \hat{A} , conformity score function $s(\cdot, \cdot; f)$ for $f \in \mathcal{F} := \mathcal{Y}^{\mathcal{X} \times \mathcal{M}}$, level α , calibration set $\{(X^{(k)}, M^{(k)}, Y^{(k)})\}_{k \in \widetilde{\text{Cal}}}$, test point $(X^{(n+1)}, M^{(n+1)})$

Output: Prediction set $\widehat{C}_{n,\alpha}^{\text{MDA-RandomMask}}(X^{(n+1)}, M^{(n+1)})$

- 1: Define $\mathcal{G}(\nu) := \widetilde{A}(\widetilde{\Phi}(\cdot, \cdot; \nu); \nu)$ for some $\nu \in \mathcal{M}$
 - 2: **for** $k \in \widetilde{\text{Cal}}$ **do** Additional nested masking
 - 3: Randomly draw ν_k , independently from $(X^{(k)}, Y^{(k)}, X^{(n+1)}, Y^{(n+1)})$
 - 4: Fit $\hat{g}_k := \mathcal{G}(\nu_k) = \widetilde{A}(\widetilde{\Phi}(\cdot, \cdot; \nu_k); \nu_k)$
 - 5: **end for**
 - 6: $\widehat{C}_{n,\alpha}^{\text{MDA-RandomMask}}(X^{(n+1)}, M^{(n+1)})$
 $:= \left\{ y \in \mathcal{Y} : (1 - \alpha)(1 + \#\text{Cal}) > \sum_{k \in \widetilde{\text{Cal}}} \mathbb{1} \left\{ s \left(\left(X^{(k)}, M^{(k)} \right), Y^{(k)}; \hat{g}_k \right) < s \left(\left(X^{(n+1)}, M^{(n+1)} \right), y; \hat{g}_k \right) \right\} \right\}$
-

Indeed, conditionally on $M^{(n+1)}$, we can apply Algorithm 18 to the modified data set $(X^{(k)}, M^{(n+1)}, Y^{(k)})_{k \in \widetilde{\text{Cal}}}$, by using the $(M^{(k)})_{k \in \widetilde{\text{Cal}}}$ as random draw for $(\nu_k)_{k \in \widetilde{\text{Cal}}}$ in line 3. This is legit only when the distribution of $(X^{(k)}, M^{(n+1)}, Y^{(k)})_{k \in \widetilde{\text{Cal}} \cup \{n+1\}}$ belongs to $\mathcal{P}_{\text{MCAR}, \text{YIM}}^{\otimes(\#\widetilde{\text{Cal}}+1)} | \mathcal{X}$, as then for any $k \in \widetilde{\text{Cal}}$, it holds that $M^{(k)} \perp\!\!\!\perp (X^{(k)}, Y^{(k)}, X^{(n+1)}, Y^{(n+1)})$.

This Algorithm 18 is a special case of leave-one-out CP presented in Section 8.C, with a randomized algorithm that only returns a pre-determined function associated with a parameter value, without fitting anything on the $n - 1$ data points. Therefore, the validity result of leave-one-out CP extends to Algorithm 18.

In particular, under $\mathcal{P}_{\text{MCAR}, \mathbb{Y} \parallel \mathbb{M} | \mathbb{X}}^{\otimes(n+1)}$, CP-MDA-Nested* corresponds to applying Algorithm 18 to the data set $(X^{(k)}, M^{(n+1)}, Y^{(k)})_{k \in \text{Cal}}$ which is exchangeable conditionally on $M^{(n+1)}$, and by using in line 3 the $(M^{(k)})_{k \in \text{Cal}}$ as random draw for $(\nu_k)_{k \in \text{Cal}}$. Therefore, CP-MDA-Nested* is MCV- $\mathcal{P}_{\text{MCAR}, \mathbb{Y} \parallel \mathbb{M} | \mathbb{X}}^{\otimes(n+1)}$ at the level $1 - 2\alpha$. \square

The idea in this re-writing is to see that, conditionally on $M^{(n+1)}$, CP-MDA-Nested* predicting on the test point $(X^{(n+1)}, M^{(n+1)})$ given the data set $(X^{(k)}, M^{(k)}, Y^{(k)})_{k=1}^n$, is in fact another run of CP-MDA-Nested* which predicts on a complete test point $\check{X}^{(n+1)} \in \check{\mathcal{X}}$, where $\check{\mathcal{X}}$ is the set of dimension $|\text{obs}(M^{(n+1)})|$ containing only the observed dimensions of \mathcal{X} according to $M^{(n+1)}$, given the cropped data set $(\check{X}^{(k)}, \check{M}^{(k)}, Y^{(k)})_{k=1}^n$, with $\check{M}^{(k)} \in \check{\mathcal{M}}$ that, similarly to \check{X} , is the set of dimension $|\text{obs}(M^{(n+1)})|$ containing only the observed dimensions of \mathcal{M} according to $M^{(n+1)}$.

Part IV

Conclusion and perspectives



Conclusion

In this thesis, we have studied several aspects of post-hoc predictive uncertainty quantification approaches, and in particular conformal methods, motivated by the goal of forecasting electricity prices. However, the methods that we developed are generic enough to be applied in any sensitive field.

Our first contribution provides an extension of Conformal Prediction (CP) to time series forecasting, a challenging context as time series are not exchangeable, the only assumption of conformal prediction. Namely, we start by studying theoretically the efficiency of Adaptive Conformal Inference (ACI, [Gibbs and Candès, 2021](#)) depending on its learning rate γ using Markov Chain theory. The results emphasize that *i*) on exchangeable residuals, ACI's efficiency worsens linearly in γ with respect to standard CP, and *ii*) when the residuals are auto-regressive, there exists an optimal $\gamma^* > 0$ that depends on the auto-regressive coefficient in a non-monotonic fashion. Therefore, we propose an adaptive algorithm, coined **AgACI**, that wraps around ACI using online aggregation under expert advice to avoid having to choose γ . Finally, we perform extensive synthetic experiments with benchmarks methods that underline the benefits brought by ACI with a well-chosen γ and by our proposed algorithm. We conclude with an application to French electricity prices forecasting in 2019, leading to the same conclusion.

We deepen this application in our second contribution, which focuses on probabilistic forecasting of French electricity prices in the demanding years 2020 and 2021. Our goal is to understand to which extent it is possible to adaptively post-process existing probabilistic forecasts so as to be more robust to sudden important non-stationarity. First, we construct a novel explanatory variable that demonstrates great interest empirically: the nuclear availability. Then, we conduct extensive numerical experiments that emphasizes *i*) the need for more adaptivity as none of them achieves nominal coverage, and *ii*) the difficulty of choosing one given model. By adding either a layer of conformalization—in an appropriate way that respect the temporal structure, such as our new proposal **OSSCP-horizon** or **AgACI**—or a layer of online aggregation, the coverage is improved, even in late 2021, while preserving informativeness. We highlight that aggregating various **AgACI**, each one of them being based on different individual forecasters, provides enhanced performances, and simultaneously reveals key aspects of the markets.

Our third contribution moves away from time series to focus on predictive uncertainty quantification with missing values. We consider impute-then-predict strategies, topped with CP. We first show that this plugged-in approach ensures marginal validity for any missingness distribution and almost all imputation function. However, by examining a Gaussian linear model, we find out that missing values induce heteroskedasticity, that is not taken into account by CP. This leads to uneven coverage depending on the missing pattern. Therefore, we suggest two algorithms, relying on the core idea of Missing Data Augmentation (MDA), and prove that they are valid conditionally to the patterns of missing values, despite their exponential number, under independence assumptions. We then show that a universally consistent quantile regression algorithm trained on the imputed data is Bayes optimal for the pinball risk, thus achieving valid coverage conditionally to any given

data point. Finally, synthetic experiments along with real critical care data application support our theory and reflect improved performance of our MDA methods.

Our last contribution constitutes a deep delve into when and how it is possible to build predictive sets that are valid conditionally on the missing pattern. We start by proving hardness results that justify the independence assumptions made by MDA's algorithms: without these assumptions, any method that is valid conditionally on the missing pattern outputs predictive sets that includes almost all the label space. Then, we characterize the interplay between missing values and predictive uncertainty quantification in (Gaussian) linear models or under non-parametric assumptions on the data distribution. Precisely, we illustrate in some cases that the predictive uncertainty increases with more missing values by providing various formal quantifications of this statement, and that even when the features are independent from the missing pattern it is crucial to allow the predictive model to know the missing pattern. In a third part, we bridge the gap between the two algorithms of MDA, providing a wide range of MDA methods in **CP-MDA-Nested***, and extending them to classification. Leveraging the unified framework, we are able to obtain stronger theoretical guarantees on its validity. Lastly, we test the robustness of MDA on synthetic experiments breaking the independence assumptions. This emphasizes that an important dependence between the missing patterns and the covariates does not undermine MDA's mask-conditional-validity, yet this is not true for the link between the response and the missing pattern.

Open directions

Following these works, several exciting perspectives are raised, beyond the ones mentioned in conclusion of each individual chapter.

Multidimensional predictive uncertainty quantification (*ongoing work*). All of the methods discussed in this manuscript produce a one dimensional predictive set. A natural question is: do they extend to a multidimensional response? For instance, we could wish to forecast the electricity prices of different market and in different countries simultaenously. As long as we design a score function that maps any point that belongs to the multidimensional \mathcal{Y} onto a unidimensional quantity ([Feldman et al., 2023](#); [Cauchois et al., 2021](#)), the theory presented will follow. However, such an approach does not take into account correlation and dependences between the uncertainty themselves, as it models the predictive uncertainty as a scalar quantity.

Therefore, an informative design' choice would be to define a score function that takes its value in a multidimensional space too. But then, to leverage CP framework, we would need to compute the empirical quantile of these multidimensional scores. However, defining multivariate quantiles is demanding as there is no canonical ordering on multivariate spaces. One historical solution could be to resort to Tukey's depth ([Tukey, 1975](#)) but it requires making distributional assumptions. To overcome this limitation, leveraging tools from the optimal transport literature, specifically Monge-Kantorovich ranks ([Chernozhukov et al., 2017](#); [Hallin et al., 2021](#)), seems promising.

Missing values. Our MDA approach achieves Mask-Conditional-Validity (MCV) by assuming independence between the mask M and the covariates X , as well as between the mask and the response Y given the covariates. In Chapter 8, we showed that we can not hope to achieve MCV without constraining the link between M and X and the link between M and Y . Relaxing the assumption of $Y \perp M | X$ seems to be particularly tricky as even the task of point-prediction (without uncertainty quantification) can be very challenging in this situation (Ayme et al., 2022). However, aiming at MCV on MAR and $Y \perp M | X$ distribution appears more within our reach. Indeed, an idea would be to build on causal inference tools (J. M. Robins, 1994; Hirano et al., 2003), such as inverse propensity weights. The underpinning idea is that while the conformity scores are still exchangeable with missing values, they are not exchangeable conditional on the mask, and under MAR mechanism we could learn the weights allowing to obtain weighted exchangeability conditional on the mask.

Another attractive path is to leverage isotonic regression (Barlow et al., 1972) to design an uncertainty quantification model conveying the core idea that more NAs induce more uncertainty, relying on the key observation that we do not need an ordering on the whole \mathcal{M} but only between nested patterns.

Broader point of view on Part III – Missing Values. The fundamental idea in both Chapters 7 and 8 is that even though the predictive distributions vary with the missing pattern, we are able to improve our predictive uncertainty quantification on one of these distributions thanks to the other ones. In fact, this idea echoes with domain adaptation questions, and it would be interesting to see how to broaden our analysis. Especially, as discussed in Chapter 8, our theoretical results – on the hardness part as well as on the MCV of our methodology – do not take any advantage of the specificity of missingness and they extend directly to any features’ group. However, our algorithms’ design relies on the core idea that we can modify the historical data to match the test domain. This is easy with missing values, but appears trickier beyond. Finding concrete other applications that would be compatible with our framework is appealing.

On leave-one-out CP approaches. To prove the MCV of our generalized MDA framework, we relied on the deep similarities with leave-one-out CP approaches based on some randomized algorithm \mathcal{A} . This uncloaked an interesting basics question on such approaches. As we do not require anymore the assumption of stochastic domination of the quantiles, it remains unclear as to why MDA-Nested overcovers. Our preliminary investigations highlight that leave-one-out CP approaches also suffer from over-cover when plugged in with an algorithm that is a mixture of deterministic predictors. Especially, assume that fitting \mathcal{A} corresponds to randomly choosing (by drawing from a Bernoulli of parameter ρ) between two pre-determined estimated regressors. Then, when ρ equals either 0 or 1, we retrieve Split CP, achieving (nearly) exactly $1 - \alpha$ coverage. When $\rho \in]0, 1[$, experimentally we observe over-coverage. The question now is: what drives the coverage behavior in between these two extremes? The answer seems unclear for now, and is in fact related to the more general question of why K -fold CP over-covers but not leave-one-out

CP? (see the experiments and especially Figure 2 in [Barber et al., 2021b](#)) This is coherent with our finding as *i*) **MDA-Nested** is in fact close to K -fold CP since several occurrence of the same augmented mask are present in the calibration set, and *ii*) our experiments on mixture of deterministic predictors also corresponds to multiple repetition of the same predictor which is what happen in K -fold CP. Therefore, this over-covering phenomenon appears to impact randomized algorithms beyond MDA, and it is thus crucial to understand.

On the implications of the theoretical properties. A broader and more fundamental perspective of post-hoc finite sample distribution-free uncertainty quantification is how the different theoretical properties (marginal and different notion of conditional validity, efficiency) intertwine. While all of these properties appear to be rooted into practice, the link between them is not well understood. For example:

- i) We can show that optimizing one can be detrimental to another: for some distributions the smallest prediction set is only marginal, as achieving conditional coverage would then increase the predictive set size. How efficiency and features-conditionality interact?
- ii) In Chapter 8, our hardness results (but seen with the lens of general groups instead of missing pattern) make one step in the direction of understanding how the efficiency depends on the calibration size. However, they only characterize the probability of uninformative sets, and the rest of the distribution remain uncharacterized. Can we derive theoretical results on the expected length depending on the calibration size? This would shed light on the practicality of binning the calibration set in order to achieve approximate conditional coverage.
- iii) Efficiency is substantially used to assess the performances of predictive sets. In Chapter 5, we have discussed extensively on the impact of infinite intervals and their impact on how to qualify a method as informative, and we ended up relying on the empirical median length instead of the empirical average length. Why should we assess efficiency through the mean and not the median?

Characterizing theoretically the interplays between all metrics is necessary to guide practice and design informed decision-making pipelines based on predictive uncertainty quantification. Indeed, identifying what can be deduced from a property and then used by external agents, depends on how this property connects to other practical requirements, i.e., other metrics.

Résumé long en français

Prévision des prix spot français de l'électricité

Cette thèse a été réalisée dans le cadre d'une convention individuelle de formation par la recherche avec EDF (Electricité de France).

Transition énergétique et électrique

“Qui aurait pu prédire la crise climatique?”

Inutile de rappeler ici que selon l'IPBES (plateforme intergouvernementale scientifique et politique sur la biodiversité et les services écosystémiques), en 150 ans, 83% de la biomasse des espèces sauvages et 41,5% de la biomasse végétale ont disparu à cause des activités humaines ; que le GIEC (groupe d'experts intergouvernemental sur l'évolution du climat) a été créé il y a plus de 35 ans pour tirer la sonnette d'alarme ; et que malgré tout cela, seules des mesures insuffisantes ont été prises aux niveaux politique et gouvernemental ([HCC-2021](#)). Pourtant, cette question est l'arbre qui cache la forêt : **que pouvons-nous réellement faire pour limiter la crise climatique, ou au moins nous y adapter ?**

Partant du plus haut niveau, une réponse partielle naturelle consiste à réduire les émissions anthropiques de gaz à effet de serre : cela est nécessaire pour respecter l'accord de Paris, qui exige que la température moyenne de la terre n'augmente pas de plus de 2°C avant 2100, par rapport à 1850. Évidemment, la réduction de notre production et de notre consommation aurait un impact rapide sur cet objectif. Cependant, la manière d'y parvenir et la question de savoir si nous voulons appliquer cette stratégie dépassent le cadre d'un débat académique et semblent très probablement appartenir à la sphère des citoyens. La manière dont nous produisons l'énergie et tout ce qui entoure ce sujet sont des champs d'action plus proches de notre rayon d'application concret, et cela reste néanmoins très pertinent pour répondre à la question qui nous intéresse.

Les dernières décennies ont été marquées par d'importants changements dans le panorama énergétique, avec une intégration croissante de la production d'énergie à partir de combustibles non fossiles. Par exemple, d'importants efforts de recherche et d'exploitation ont été déployés pour développer les énergies renouvelables ([RTE, 2022](#); [IEA, 2022a](#))⁶. En particulier, la France s'est engagée à atteindre la neutralité carbone d'ici 2050, et notamment à atteindre 1/3 d'énergies renouvelables dans la consommation finale brute d'énergie d'ici 2030. La France a également décidé de soutenir le développement des centrales nucléaires

⁶RTE est le Réseau français de Transport d'Électricité, tandis que l'IEA est l'Agence Internationale de l'Énergie.

afin d’atteindre un mix énergétique décarboné. En parallèle, de nombreux usages ont été électrifiés, ou sont en passe de l’être, comme les véhicules électriques et les stockages distribués. L’autoconsommation (aussi appelée consommateur-producteur, c’est-à-dire consommer l’énergie que l’on produit) ou encore les programmes de réponse à la demande (c’est-à-dire adapter la demande en fonction de la production, et non le traditionnel inverse) sont également fortement encouragés (Bakare et al., 2023).

La prolifération de ces nouveaux usages de l’électricité et l’importance croissante des énergies renouvelables intermittentes modifient profondément le paysage énergétique en Europe et sont à l’origine de transformations majeures des marchés européens de l’électricité. Ceux-ci deviennent notamment plus dépendants et plus volatils. **Par conséquent, une prévision précise des prix de l’électricité est nécessaire pour stabiliser la planification de la production d’énergie et ainsi réduire les émissions de carbone associées en augmentant les investissements dans les énergies renouvelables et les solutions de stockage. Dans cette thèse, nous nous concentrons sur les prix à court-terme.**

Marchés de l’électricité

Il y a 4 principaux marchés court-terme en France, et plus généralement en Europe.

- i) Le premier, sur lequel nous nous concentrerons, est le marché *spot*. Le marché spot de l’électricité est un marché d’enchères à l’aveugle dans lequel les producteurs et les fournisseurs font des offres pour chaque heure, ou pour un bloc d’heures, du jour suivant. Le marché ferme à 12 heures la veille de la livraison. Les 24 prix horaires sont définis par le principe du “pay-as-clear” : tous les acteurs échangeront des mégawattheures au même prix, qui, à première vue, peut être considéré comme le croisement entre l’offre et la demande globales. Cependant, la définition du prix est plus complexe, car elle prend en compte les interconnexions entre les différents pays, ainsi que les offres dites “en bloc”.
- ii) Le second est le marché *intraday*. Il s’agit d’un marché en continu, offrant des produits à l’heure, à la demi-heure et au quart d’heure. Contrairement au marché spot, les prix sont fixés à la volée afin de répondre aux ordres le plus rapidement possible, avec un moment de clôture de 5 à 15 minutes avant la livraison.
- iii) Enfin, les deux derniers marchés sont les marchés des *services système* et de *réserve*. Ces marchés sont gérés par le gestionnaire du système de transport et sont chargés d’assurer l’équilibre parfait entre l’offre et la demande à tout moment.

Ces marchés à court terme sont affectés par la transition décrite précédemment. D’une part, le besoin d’une plus grande sécurité d’approvisionnement en électricité à différentes échelles de temps conduit à une refonte des services système, avec la création de nouveaux marchés pour ces services au niveau européen, notamment dans le nouveau cadre réglementaire “Electricity balancing” adopté par la Commission européenne en 2017 (EU-2017/2195). D’autre part, la pénétration croissante des énergies renouvelables a accentué l’incertitude sur un horizon à court terme de la production d’électricité, affectant le fonctionnement

des marchés infra-journaliers, qui deviennent l'outil indispensable pour gérer les erreurs de prévision de la production renouvelable. Sur le marché allemand, on observe déjà de fortes corrélations entre les prix et la production éolienne, et ce n'est qu'une question de temps avant que ces phénomènes n'apparaissent en France. La présence d'actifs de stockage, dont le prix ne cesse de baisser – même s'il est actuellement assez élevé –, permet de mettre en place de nouvelles stratégies de marché pour stabiliser l'offre et réduire les coûts.

Prévision des prix de l'électricité

Dans ce contexte en pleine transformation, il est essentiel de disposer de méthodes performantes de prévision des prix sur l'ensemble des marchés à court terme.

En effet, de bonnes prévisions de prix sur les marchés successifs permettent de *mieux anticiper les flux financiers liés à la production renouvelable et d'optimiser le placement de la production sur les différents marchés*. C'est un des éléments essentiels pour une bonne valorisation de ces actifs de production, qui permettra *d'encourager les investissements dans ces actifs bas carbone*.

De plus, une prévision précise des prix, à la fois sur les marchés successifs et sur les différents prix horaires d'un même marché, permet *d'optimiser la gestion des flexibilités* (batterie physique ou contrat d'effacement de consommation à court terme, flexibilité d'ajustement à la hausse et à la baisse des centrales thermiques, etc.). En particulier, l'augmentation de la valeur de ces flexibilités encouragera les acteurs à *investir dans ces actifs, ce qui conduira à un système électrique plus sûr*.

Or, la prévision des prix de l'électricité est un véritable défi en raison de toutes les spécificités susmentionnées de l'électricité : adéquation entre la demande et la production à tout moment, caractère non stockable de l'électricité, échanges entre différents pays via les interconnexions, caractère variable des moyens de production, etc. Plus précisément, ces caractéristiques conduisent à des prix négatifs ou extrêmement élevés dont l'occurrence n'est pas négligeable (voir Figure F.1). Sans parler des récents événements malheureux et fortuits qui ont eu un impact considérable sur les marchés, les rendant hautement non stationnaires, tels que la pandémie de Covid-19 en 2020-2021 (IEA, 2021), le problème de corrosion sous contrainte qui a affecté les centrales nucléaires françaises en 2022 ou la crise des marchés du gaz déclenchée par l'invasion de l'Ukraine par la Russie (IEA, 2022b). Malgré le nombre croissant de données historiques disponibles, les modèles de pointe (Weron, 2014; Lago et al., 2021) (de la prévision classique des séries temporelles aux méthodes d'apprentissage profond), ainsi que les études internes d'EDF R&D⁷, ne permettent pas d'obtenir des erreurs de prévision inférieures à 10% du prix réalisé⁸. À titre de référence, les prévisions de la consommation nationale n'atteignent des erreurs que de l'ordre de 1% de la consommation réalisée.

⁷Notons ici que les outils de prévision opérationnelle disponibles à EDF-Trading peuvent être plus efficaces, mais ils utilisent des informations en temps réel qui ne sont pas disponibles en tant que données historiques.

⁸De manière surprenante, cela vaut pour les prévisions avant 2020 comme après 2020 : les erreurs sont plus importantes après 2020, mais comme les prix sont également plus élevés, l'erreur relative est plus importante après 2020.

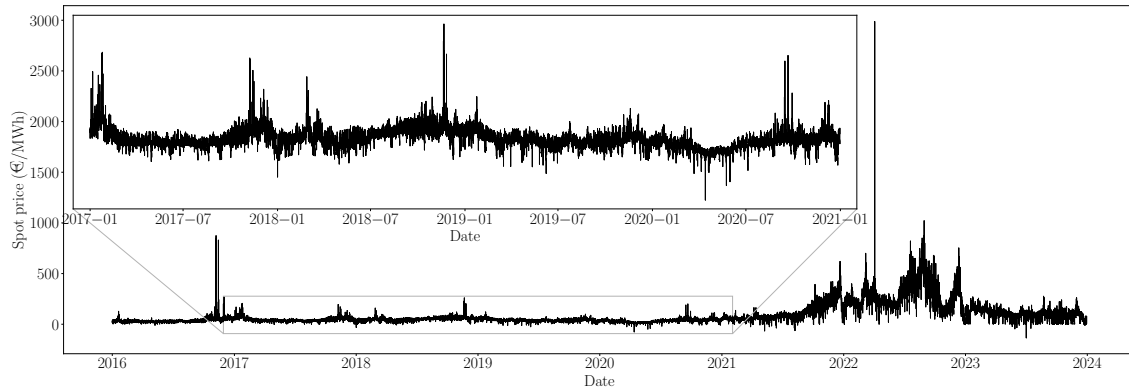


Figure F.1: Temporal evolution of the French electricity spot prices between 2016 and 2021.

Tirant parti de l'émergence de plateformes de données ouvertes telles que la plateforme de transparence ENTSO-E⁹ ou la plateforme Eco2Mix alimentée par RTE permettrait probablement d'améliorer les prévisions de prix de l'électricité. Cependant, l'agrégation de différentes sources de données introduit un nouveau cadre complexe : l'occurrence de *valeurs manquantes* qui s'accompagne de défis computationnels et statistiques. Par exemple, elle peut être causée par des fréquences temporelles ou des horizons de marché différents entre des variables explicatives fondamentalement différentes. En outre, la qualité des données évolue avec le temps (au fur et à mesure que les processus se consolident) et des anomalies peuvent être observées.

Prévision probabiliste des prix de l'électricité

De manière cruciale, ces méthodes de prévision fournissent des prédictions ad hoc, sans indication du degré de confiance que l'on peut leur accorder. Pour garantir la confiance des acteurs clés des marchés de l'énergie à l'égard de ces outils d'aide à la décision, il est essentiel de **quantifier leur incertitude prédictive**.

En outre, les décisions en matière de commerce et de gestion de l'énergie (telles que celles mentionnées précédemment) nécessitent des outils de gestion des risques qui sont basés sur des prévisions probabilistes des prix de l'électricité, ce qui a conduit à une expansion rapide de la littérature dans ce domaine (voir la revue de [Nowotarski and Weron, 2018](#)). Toutefois, les prévisions probabilistes traditionnelles ne sont valables que de manière asymptotique ou sur la base d'hypothèses fortes sur les données qui ne sont généralement pas respectées par les prix de l'électricité (gaussianité, stationnarité).

Cela favorise le développement d'approches probabilistes adaptatives pour la prévision des prix, capables d'apprendre continuellement et de s'adapter aux comportements évolutifs des prix de l'électricité, ce qui permet d'obtenir des prévisions probabilistes précises et fiables, même sur des *séries temporelles non stationnaires*.

⁹ENTSO-E est le réseau européen des gestionnaires de réseaux de transport d'électricité..

Dans cette thèse de doctorat, nous proposons de fournir des **outils théoriques** capables de quantifier l'incertitude prédictive sous de **faibles hypothèses sur la distribution des données sous-jacentes** et dont les garanties ne dépendent pas de l'algorithme de prédiction. Nous envisageons des méthodes **post-hoc**, afin de permettre leur utilisation de manière plug-in : tout acteur des marchés de l'énergie pourrait conserver son pipeline opérationnel préféré et transformer les prédictions résultantes en prévisions probabilistes garanties.

Aperçu rapide des contributions

La prédiction conforme par partition (SCP, [Vovk et al., 2005](#); [Papadopoulos et al., 2002](#); [Lei et al., 2018](#)) est une procédure polyvalente associant des intervalles prédictifs à tout modèle de prédiction. Contrairement aux méthodes de prédiction probabilistes existantes, CP est hautement prometteuse car elle offre des garanties théoriques à taille d'échantillon finie, sous la seule hypothèse distributionnelle que les données sont échangeables (c'est-à-dire que la distribution des données est invariante par permutation, ce qui est plus faible que des données indépendantes et identiquement distribuées).

Formellement, supposons que nous disposons de n données $(X_i, Y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ où Y est la variable à prédire (e.g., le prix de l'électricité) et $X \in \mathbb{R}^d$ les d covariables (e.g., les productions). L'utilisateur fixe un *taux de non-couverture* $\alpha \in [0, 1]$ (typiquement 0.1 ou 0.05). SCP construit un intervalle prédictif $\mathcal{C}_{n,\alpha}$ tel que $\mathbb{P}\{Y_{n+1} \in \mathcal{C}_{n,\alpha}(X_{n+1})\} \geq 1 - \alpha$: on dit que $\mathcal{C}_{n,\alpha}$ est *valide* marginalement. Sa longueur doit être la plus petite possible pour qu'il soit informatif (*efficace*). Un exemple de tel intervalle est donné en Figure F.2.

Cependant, SCP n'est pas applicable sur une séries temporelles (telles que les prix de l'électricité) car elles ne sont pas échangeables en raison de leur dépendance temporelle. Pour remédier à cette limitation, une première approche ([Gibbs and Candès, 2021](#)) repose sur l'utilisation d'un taux de non-couverture adaptatif α_t , qui est mis à jour en fonction des performances passées et d'un hyperparamètre $\gamma > 0$, jouant le rôle d'un taux d'apprentissage. En utilisant la théorie des chaînes de Markov, la première contribution de cette thèse analyse

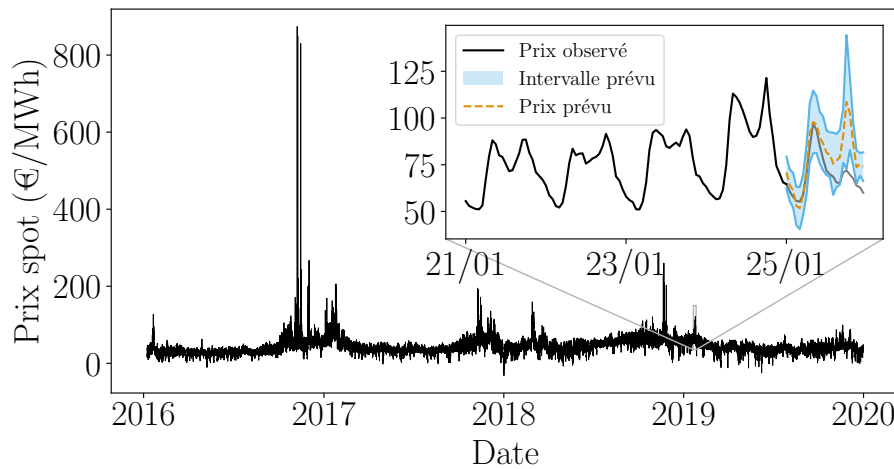


Figure F.2: intervalles prédictifs pour les prix de l'électricité.

l'influence de γ sur l'efficacité des intervalles prédictifs associés. Cela a permis de proposer une nouvelle méthode ne nécessitant pas le choix de γ —et donc utilisable en pratique—basée sur l'agrégation d'experts en ligne. Suite à l'explosion des prix de l'électricité en 2021, la deuxième contribution de cette thèse étudie l'impact de cette non-stationnarité accrue sur les prévisions probabilistes, et les améliorations apportées par différentes surcouches adaptatives telles que SCP et l'agrégation en ligne.

Néanmoins, pour améliorer les prévisions des prix de l'électricité, nous pourrions tirer parti de l'émergence de plateformes de données ouvertes pour intégrer davantage de variables explicatives telles que les prix des matières premières ou les prix d'autres marchés corrélés. Cependant, l'agrégation de différentes sources de données s'accompagne de défis méthodologiques, tels que le traitement des valeurs manquantes, comme les fréquences temporelles et les horizons de marché peuvent différer. Les données manquantes sont courantes dans la pratique statistique et, paradoxalement, leur nombre augmente avec la quantité de données.

Une approche traditionnelle pour obtenir des prédictions ponctuelles consiste à remplacer (imputer) les valeurs manquantes (NAs) par des valeurs plausibles, puis à entraîner n'importe quel algorithme d'apprentissage sur les données complétées. Cependant, il n'existe aucune méthode permettant de quantifier l'incertitude prédictive avec les NAs. Les troisième et quatrième contributions de cette thèse montrent que SCP appliquée à un jeu de données imputé bénéficie exactement des mêmes garanties de *validité* marginales que sur des données complètes. La force de ce résultat réside dans sa généralité : il implique que l'utilisateur peut choisir n'importe quelle imputation, même naïve, sans affecter la validité des intervalles, même pour des NAs informatives (un scénario complexe et rarement étudié). Cependant, Les troisième et quatrième contributions de cette thèse constatent que les NA génèrent de l'hétéroscédasticité : la validité des intervalles dépend de quelles variables explicatives sont observées. Ils proposent les premiers algorithmes pour résoudre ce problème, qui sont extrêmement simples à mettre en pratique. Théoriquement valides, les hypothèses sur lesquelles ils reposent sont presque minimales d'après de nouveaux résultats d'impossibilité.

Plan de la thèse et principales contributions

Ce manuscrit est divisé en trois parties principales. La Part I est organisée comme suit. Ce chapitre 2 donne un aperçu rapide du plan et des principales contributions. Chapter 3 est une introduction pédagogique aux méthodes de prédiction conforme (voir Table F.3 pour un guide de lecture), basée sur un tutoriel conçu pendant la réalisation de ce doctorat. Enfin, dans Chapter 4 nous donnons un résumé plus technique et détaillé de nos contributions.

Part II étudie la quantification de l'incertitude prédictive post hoc pour **séries temporelles**. Le premier obstacle à l'application de méthodes conformes permettant d'obtenir des prévisions probabilistes garanties des prix de l'électricité de manière post hoc est l'aspect temporel hautement non stationnaire des prix de l'électricité, qui rompt l'hypothèse d'échangeabilité. Dans Chapter 5 (basé sur un travail conjoint avec Olivier Féron, Yannig Goude, Julie Josse et Aymeric Dieuleveut), nous proposons un algorithme sans paramètre adapté aux séries temporelles, qui est basé sur l'analyse théorique de l'efficacité de l'Inférence Conformale Adaptative (Gibbs and Candès, 2021). Afin d'étudier plus en profondeur comment des prévisions probabilistes post-hoc adaptatives des prix de l'électricité peuvent être obtenues, dans Chapter 6 (basé sur le stage de Grégoire Dutot, co-supervisé avec Olivier Féron et Yannig Goude), nous menons une étude applicative approfondie sur un nouvel ensemble de données de prix spot français récents et turbulents en 2020 et 2021.

Un autre défi auquel la quantification de l'incertitude prédictive pour la prévision des prix de l'électricité est confrontée est l'occurrence de **données manquantes**. Par conséquent, dans Part III (basé sur des travaux conjoints avec Aymeric Dieuleveut, Julie Josse et Yaniv Romano), nous analysons l'interaction entre les valeurs manquantes et la quantification de l'incertitude prédictive. Dans Chapter 7, nous soulignons que les données manquantes induisent une hétéroscédasticité, conduisant à une couverture inégale en fonction des variables explicatives observées. Nous concevons deux algorithmes qui obtiennent une couverture égalisée pour tout schéma de données manquantes sous des hypothèses distributionnelles sur le mécanisme de données manquantes. Dans Chapter 8, nous approfondissons l'analyse théorique pour comprendre précisément quelles hypothèses de distribution sont inévitables pour l'informativité théorique. Nous unifions également les algorithmes proposés précédemment dans un cadre général qui démontre la robustesse empirique aux violations de la distribution supposée des données manquantes.

Toutes ces contributions sont mises en œuvre à l'aide d'un code source ouvert disponible sur [cette GitHub](#). Le tutoriel sur lequel Chapter 3 est basé a également été mis en libre accès sur [ce site web](#).

Chaque chapitre étant autonome, les notations peuvent varier légèrement d'un chapitre à l'autre.

Contribution associée		Section pertinentes de Chapter 3				
Ch. 3	Tutoriel à: ► <i>MASPIN days 2023</i> (national), avec C. Boyer ► <i>ENBIS 2023</i> (européen) ► <i>UAI 2024</i> (international), avec A. Dieuleveut ► <i>ICML 2024</i> (international), avec A. Dieuleveut	Split CP, Section 3.2	Validité conditionnelle, Section 3.3	Full et K -fold CP, Section 3.4	Non échangeable, Section 3.5	
Ch. 5	M. Zaffran, O. Féron, Y. Goude, J. Josse and A. Dieuleveut <i>ICML 2022</i> ¹	✓			✓	
Ch. 6	G. Dutot*, M. Zaffran*, O. Féron and Y. Goude soumis à <i>Applied Energy</i> ²	✓			✓	
Ch. 7	M. Zaffran, A. Dieuleveut, J. Josse and Y. Romano <i>ICML 2023</i> ³	✓	(✓)	(✓)		
Ch. 8	M. Zaffran, J. Josse, Y. Romano and A. Dieuleveut soumis à <i>JMLR</i> ⁴	✓	✓	✓		

Table F.3: Résumé de la production scientifique (* signifie contribution équivalente), avec des indications pour une lecture parcimonieuse du Chapter 3.

Résumé des chapitres

Ci-dessous se trouvent des résumés plus détaillés de chacun des chapitres contenant les contributions principales de cette thèse.

Chapitre 5

La quantification de l'incertitude des modèles prédictifs est cruciale dans les problèmes de prise de décision. La prédiction conforme est une solution générale et théoriquement solide. Cependant, elle nécessite des données échangeables, ce qui exclut les séries temporelles. Bien que des travaux récents aient abordé cette question, nous soutenons que l'inférence adaptative conforme (ACI, [Gibbs and Candès, 2021](#)), développée pour les séries temporelles avec changement de distribution, est une bonne procédure pour les séries temporelles avec une dépendance générale. Nous analysons théoriquement l'impact du taux d'apprentissage sur son efficacité dans le cas échangeable et auto-régressif. Nous proposons une méthode sans paramètre, AgACI, qui s'appuie de manière adaptative sur l'ACI en se basant sur l'agrégation d'experts en ligne. Nous menons des simulations complètes et équitables contre

¹“Adaptive Conformal Predictions for Time Series”.

²“Adaptive Probabilistic Forecasting of French Electricity Spot Prices”.

³“Conformal Prediction with Missing Values”.

⁴“Predictive Uncertainty Quantification with Missing Covariates”.

des méthodes concurrentes qui plaident en faveur de l'utilisation de l'ACI pour les séries temporelles. Nous menons une étude de cas réelle : la prévision des prix de l'électricité. L'algorithme d'agrégation proposé fournit des intervalles de prédiction efficaces pour les prévisions à horizon un jour. L'ensemble du code et des données permettant de reproduire les expériences sont disponibles sur [GitHub](#).

Chapitre 6

La prévision des prix de l'électricité (EPF) joue un rôle majeur pour les compagnies d'électricité en tant qu'élément fondamental pour les décisions commerciales ou les opérations de gestion de l'énergie. L'électricité ne pouvant être stockée, les prix de l'électricité sont très volatils, ce qui rend la prévision des prix de l'électricité particulièrement difficile. Cela est d'autant plus vrai lorsque des événements dramatiques et fortuits perturbent les marchés. Les décisions en matière de commerce et, plus généralement, de gestion de l'énergie nécessitent des outils de gestion des risques basés sur l'EPF probabiliste (PEPF). Dans ce contexte difficile, nous plaçons en faveur du déploiement de stratégies de boîtes noires hautement adaptatives permettant de transformer toute prévision en un intervalle prédictif adaptatif robuste, comme la prédiction conforme et l'agrégation en ligne, en tant que dernière couche fondamentale de tout pipeline opérationnel.

Nous proposons d'étudier un nouvel ensemble de données contenant les prix spot de l'électricité en France pendant les années 2020-2021, et de construire une nouvelle variable explicative révélant un pouvoir prédictif élevé, à savoir la disponibilité du nucléaire. L'analyse comparative de l'état de l'art du PEPF sur cet ensemble de données met en évidence la difficulté de choisir un modèle donné, car ils se comportent tous très différemment dans la pratique, et aucun d'entre eux n'est fiable. Cependant, nous proposons une conformalisation adéquate, `OSSCP-horizon`, qui améliore les performances des méthodes PEPF, même dans la période la plus hasardeuse de la fin de l'année 2021. Enfin, nous soulignons que la combinaison avec l'agrégation en ligne surpasse de manière significative toutes les autres approches, et devrait être la solution préférée, car elle fournit des prévisions probabilistes fiables.

Chapitre 7

La prédiction conforme est un cadre théorique pour la construction d'intervalles prédictifs. Nous étudions la prédiction conforme avec des valeurs manquantes dans les covariables, un cadre qui pose de nouveaux défis à la quantification de l'incertitude. Nous montrons tout d'abord que la garantie de couverture marginale de la prédiction conforme est valable pour les données imputées, quelle que soit la distribution des valeurs manquantes et pour la quasi-totalité des fonctions d'imputation. Cependant, nous soulignons que la couverture moyenne varie en fonction de la structure des valeurs manquantes : les méthodes conformes ont tendance à construire des intervalles de prédiction qui ne couvrent pas suffisamment la réponse sous certaines structures de données manquantes. Cela motive notre nouveau cadre de régression quantile conformalisée généralisé, l'augmentation des données manquantes, qui produit des intervalles de prédiction qui sont valides conditionnellement

aux modèles de valeurs manquantes, malgré leur nombre exponentiel. Nous montrons ensuite qu'un algorithme de régression quantile universellement cohérent, entraîné sur les données imputées, est Bayes-optimal en ce qui concerne le risque pinball, ce qui permet d'obtenir une couverture valide conditionnellement à tout point donné. En outre, nous examinons le cas d'un modèle linéaire, ce qui démontre l'importance de notre proposition pour surmonter l'hétéroscédasticité induite par les valeurs manquantes. En utilisant des données synthétiques et des données de soins intensifs, nous corroborons notre théorie et rapportons une amélioration de la performance de nos méthodes.

Chapitre 8

La quantification de l'incertitude prédictive est cruciale dans les problèmes de prise de décision. Nous étudions comment quantifier de manière adéquate l'incertitude prédictive avec des covariables manquantes. Le fait que les valeurs manquantes induisent une hétéroscédasticité sur la distribution prédictive de la réponse compte tenu des covariables observées constitue une limitation. Nous nous concentrons donc sur la construction d'ensembles prédictifs pour la réponse qui sont valides *conditionnellement* au modèle des valeurs manquantes. Nous montrons qu'il est impossible d'atteindre cet objectif de manière informative sans hypothèse de distribution, et nous proposons des restrictions utiles sur la classe de distribution. Motivés par ces résultats d'impossibilité, nous caractérisons la façon dont les valeurs manquantes et l'incertitude prédictive s'entremêlent. En particulier, nous formalisons rigoureusement l'idée selon laquelle plus il y a de valeurs manquantes, plus l'incertitude prédictive est élevée. Ensuite, nous introduisons un cadre généralisé, appelé **CP-MDA-Nested***, qui produit des ensembles prédictifs à la fois dans la régression et la classification. Sous réserve d'indépendance entre le modèle de valeurs manquantes et les caractéristiques et la réponse (une hypothèse justifiée par nos résultats de dureté), ces ensembles prédictifs sont valides conditionnellement à tout modèle de valeurs manquantes. En outre, ils offrent une grande souplesse dans le compromis entre *la variabilité statistique* et *l'efficacité*. Enfin, nous évaluons expérimentalement les performances de **CP-MDA-Nested*** au-delà de son champ de validité théorique, en démontrant des résultats prometteurs dans des configurations plus difficiles que l'indépendance.

Bibliography

- L. Amabile, D. Bresch-Pietri, G. El Hajje, S. Labbé, and N. Petit. Optimizing the self-consumption of residential photovoltaic energy and quantification of the impact of production forecast uncertainties. *Advances in Applied Energy*, 2:100020, 2021. (p. [106](#).)
- Y. Amara-Ouali, M. Fasiolo, Y. Goude, and H. Yan. Daily peak electrical load forecasting with a multi-resolution approach. *International Journal of Forecasting*, 39(3):1272–1286, 2023. (p. [112](#).)
- A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4), 2023. (pp. [32](#), [62](#), [132](#), [148](#), [152](#), and [185](#).)
- A. N. Angelopoulos, E. J. Candes, and R. J. Tibshirani. Conformal PID Control for Time Series Prediction. In *Advances in Neural Information Processing Systems*, 2023. (pp. [53](#) and [118](#).)
- A. N. Angelopoulos, R. F. Barber, and S. Bates. Online conformal prediction with decaying step sizes, 2024. (pp. [53](#) and [118](#).)
- A. Ayme, C. Boyer, A. Dieuleveut, and E. Scornet. Near-optimal rate of consistency for linear models with missing values. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, 162, 1211–1243. PMLR, 2022. (pp. [129](#), [133](#), [144](#), [150](#), [168](#), [180](#), [204](#), and [218](#).)
- A. Ayme, C. Boyer, A. Dieuleveut, and E. Scornet. Naive imputation implicitly regularizes high-dimensional linear models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, 202 of *Proceedings of Machine Learning Research*, 1320–1340. PMLR, 23–29 Jul 2023. (pp. [168](#) and [183](#).)
- A. Ayme, C. Boyer, A. Dieuleveut, and E. Scornet. Random features models: a way to study the success of naive imputation, 2024. (pp. [168](#) and [183](#).)
- M. S. Bakare, A. Abdulkarim, M. Zeeshan, and A. N. Shuaibu. A comprehensive overview on demand side energy management towards smart grids: challenges, solutions, and future direction. *Energy Informatics*, 6(1), Mar. 2023. (pp. [4](#) and [221](#).)

- R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2), 2021a. (pp. [23](#), [39](#), [138](#), [145](#), [167](#), and [173](#).)
- R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1), 2021b. (pp. [44](#), [46](#), [47](#), [133](#), [137](#), [148](#), [172](#), [184](#), [189](#), [190](#), [208](#), [209](#), [211](#), [212](#), and [219](#).)
- R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), Apr. 2023. (pp. [51](#) and [148](#).)
- R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions: Theory and application of isotonic regression*. John Wiley & Sons, 1972. (p. [218](#).)
- O. Bastani, V. Gupta, C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022. (pp. [53](#) and [118](#).)
- J. Berrisch and F. Ziel. Crps learning. *Journal of Econometrics*, 237(2):105221, 2023. (pp. [69](#) and [115](#).)
- J. Berrisch and F. Ziel. Multivariate probabilistic crps learning with an application to day-ahead electricity prices. *International Journal of Forecasting*, 2024a. (p. [107](#).)
- J. Berrisch and F. Ziel. *The profoc Package: An R package for probabilistic forecast combination using CRPS Learning*, 2024b. R package version 1.3.1. (p. [115](#).)
- A. Bhatnagar, H. Wang, C. Xiong, and Y. Bai. Improved online conformal prediction via strongly adaptive online learning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, 202 of *Proceedings of Machine Learning Research*, 2337–2363. PMLR, 23–29 Jul 2023. (pp. [53](#) and [118](#).)
- M. Bian and R. F. Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023. (p. [49](#).)
- R. Bjorgan, C.-C. Liu, and J. Lawarree. Financial risk management in a competitive electricity market. *IEEE Transactions on power systems*, 14(4):1285–1291, 1999. (p. [107](#).)
- D. Bunn, A. Andresen, D. Chen, and S. Westgaard. Analysis and forecasting of electricity price risks with quantile factor models. *The Energy Journal*, 37(1), 2016. (p. [106](#).)
- E. Burnaev and V. Vovk. Efficiency of conformalized ridge regression. In M. F. Balcan, V. Feldman, and C. Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, 35 of *Proceedings of Machine Learning Research*, 605–622, Barcelona, Spain, 13–15 Jun 2014. PMLR. (p. [43](#).)

- Y. Cai and N. Davies. A simple bootstrap method for time series. *Communications in Statistics-Simulation and Computation*, 41(5):621–631, 2012. (p. [95](#).)
- E. Candès, L. Lei, and Z. Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, Jan. 2023. (p. [50](#).)
- M. Cauchois, S. Gupta, and J. C. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021. (pp. [24](#) and [217](#).)
- M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, page 1–66, Feb. 2024. (pp. [51](#) and [63](#).)
- N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006. (pp. [55](#), [68](#), [69](#), [107](#), [115](#), and [118](#).)
- W. Chen, Z. Wang, W. Ha, and R. F. Barber. Trimmed conformal prediction for high-dimensional models, 2016. (p. [43](#).)
- W. Chen, K. Chun, and R. F. Barber. Discretized conformal prediction for efficient distribution-free inference. *Stat*, 7(1), Jan. 2018. (p. [43](#).)
- V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017. (p. [217](#).)
- V. Chernozhukov, K. Wüthrich, and Z. Yinchu. Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data. In *Conference On Learning Theory*. PMLR, 2018. (pp. [49](#), [63](#), and [116](#).)
- V. Chernozhukov, K. Wüthrich, and Y. Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48), 2021. (p. [24](#).)
- G. Cherubin, K. Chatzikokolakis, and M. Jaggi. Exact optimization of conformal predictors via incremental and decremental learning. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021. (p. [43](#).)
- C. Cornell, N. T. Dinh, and S. A. Pourmousavi. A probabilistic forecast methodology for volatile electricity prices in the australian national electricity market. *International Journal of Forecasting*, 2024. (p. [107](#).)
- M. Dashevskiy and Z. Luo. Network traffic demand prediction with confidence. In *IEEE Global Telecommunications Conference*. IEEE, 2008. (p. [63](#).)
- M. Dashevskiy and Z. Luo. Time series prediction with performance guarantee. *IET communications*, 5(8):1044–1051, 2011. (p. [63](#).)
- T. Deschatre, O. Féron, and P. Gruet. A survey of electricity spot and futures price models for risk management applications. *Energy Economics*, 102:105504, 2021. (p. [107](#).)

- T. Ding, A. Angelopoulos, S. Bates, M. Jordan, and R. J. Tibshirani. Class-conditional conformal prediction with many classes. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, 36, 64555–64576. Curran Associates, Inc., 2023. (p. 40.)
- D. Dua and C. Graff. UCI machine learning repository, 2017. (p. 141.)
- M. L. Eaton. *Multivariate statistics*. John Wiley & Sons, Nashville, TN, 1983. (p. 149.)
- EU-2017/2195. Commission Regulation (EU) 2017/2195 of 23 November 2017 establishing a guideline on electricity balancing (text with EEA relevance.). <https://eur-lex.europa.eu/eli/reg/2017/2195/oj>. Published: 2017-11-23. (pp. 4 and 221.)
- EUPHEMIA. Euphemia public description, single price coupling algorithm, April 2019. (p. 74.)
- C. Fannjiang, S. Bates, A. N. Angelopoulos, J. Listgarten, and M. I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022. (p. 50.)
- M. Fasiolo, S. N. Wood, M. Zaffran, R. Nedellec, and Y. Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 116(535):1402–1412, mar 2020. (p. 112.)
- M. Fasiolo, S. N. Wood, M. Zaffran, R. Nedellec, and Y. Goude. qgam: Bayesian nonparametric quantile regression modeling in R. *Journal of Statistical Software*, 100(9):1–31, 2021. (p. 112.)
- S. Feldman, S. Bates, and Y. Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023. (p. 217.)
- France-2023-491. Loi n° 2023-491 du 22 juin 2023 relative à l’accélération des procédures liées à la construction de nouvelles installations nucléaires à proximité de sites nucléaires existants et au fonctionnement des installations existantes. <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000047715784/>. Published: 2023-06-22. (p. 3.)
- J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232, 2001. (p. 112.)
- J. H. Friedman, E. Grosse, and W. Stuetzle. Multidimensional additive spline approximation. *SIAM J. Sci. Stat. Comput.*, 1983. (p. 70.)
- P. Gaillard and Y. Goude. *OPERA*, 2021. R package version 1.2.0. (pp. 69 and 115.)
- P. Gaillard, G. Stoltz, and T. Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, 176–196. PMLR, 2014. (p. 69.)

- P. Gaillard, Y. Goude, and R. Nedellec. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, 32(3):1038–1050, 2016. (p. 62.)
- I. Gibbs and E. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021. (pp. iii, v, 7, 52, 60, 63, 64, 65, 107, 117, 216, 224, 226, and 227.)
- I. Gibbs and E. Candès. Conformal inference for online prediction with arbitrary distribution shifts, 2023. (pp. 53, 118, and 173.)
- I. Gibbs, J. J. Cherian, and E. J. Candès. Conformal prediction with conditional guarantees, 2023. arXiv: 2305.12616. (p. 24.)
- B. Goehry. Random forests for time-dependent processes. *ESAIM: Probability and Statistics*, 24:801–826, 2020. (p. 95.)
- B. Goehry, H. Yan, Y. Goude, P. Massart, and J.-M. Poggi. Random forests for time series. *HAL hal-03129751*, 2021. (p. 95.)
- L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35:507–520, 2022. (p. 112.)
- L. Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1), 2022. (pp. 24 and 173.)
- Y. Gui, R. F. Barber, and C. Ma. Conformalized matrix completion, 2023a. (p. 170.)
- Y. Gui, R. Hore, Z. Ren, and R. F. Barber. Conformalized survival analysis with adaptive cut-offs. *Biometrika*, Dec. 2023b. (p. 50.)
- C. Gupta, A. K. Kuchibhotla, and A. Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022. (pp. 34, 45, 47, 137, 172, 184, 189, and 209.)
- M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d : A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165, 2021. (p. 217.)
- W. Härdle, J. Horowitz, and J.-P. Kreiss. Bootstrap methods for time series. *International Statistical Review*, 71(2):435–459, 2003. (p. 95.)
- T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986. (p. 111.)
- E. Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019. (p. 69.)

- HCC-2021. Rapport annuel 2021 “renforcer l’atténuation, engager l’adaptation”, haut conseil pour le climat. <https://www.hautconseilclimat.fr/wp-content/uploads/2021/06/HCC-rapport-annuel-2021.pdf>. Published: 2021-06-30. (pp. 3 and 220.)
- O. Himych, A. Durand, and Y. Goude. Adaptive Forecasting of Extreme Electricity Load. working paper or preprint, Feb. 2024. (p. 123.)
- K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. (p. 218.)
- T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016. (p. 107.)
- P. IEA. Covid-19 impact on electricity. Technical report, Technical report, 2021. (pp. 5, 106, and 222.)
- P. IEA. Renewable electricity. Technical report, Technical report, 2022a. (pp. 3, 106, and 220.)
- P. IEA. World energy outlook 2022. Technical report, Technical report, 2022b. (pp. 5, 106, and 222.)
- R. Izbicki, G. Shimizu, and R. Stern. Flexible distribution-free conditional predictive bands using density estimators. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 108, 3068–3077. PMLR, 2020. (pp. 143 and 145.)
- R. Izbicki, G. Shimizu, and R. B. Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022. (pp. 24, 143, and 145.)
- L. P. Z. J. M. Robins, A. Rotnitzky. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89 (427):846–866, 1994. (p. 218.)
- A. Jędrzejewski, J. Lago, G. Marcjasz, and R. Weron. Electricity price forecasting: The dawn of machine learning. *IEEE Power and Energy Magazine*, 20(3):24–31, 2022. (p. 106.)
- W. Jiang, M. Bogdan, J. Josse, S. Majewski, B. Miasojedow, V. Ročková, and TraumaBase® Group. Adaptive bayesian slope: Model selection with incomplete data. *Journal of Computational and Graphical Statistics*, 31(1):113–137, 2022. (p. 142.)
- Y. Jin and E. J. Candès. Model-free selective inference under covariate shift via weighted conformal p-values, 2023. (p. 50.)
- J. Josse and J. P. Reiter. Introduction to the Special Section on Missing Data. *Statistical Science*, 33(2):139 – 141, 2018. (pp. 129 and 168.)
- J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values, 2019. (pp. 129, 168, and 183.)

- C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023. (pp. [24](#) and [173](#).)
- C. Kath and F. Ziel. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2):777–799, 2021. (pp. [63](#) and [64](#).)
- B. Kim and R. F. Barber. Black-box tests for algorithmic stability. *Information and Inference: A Journal of the IMA*, 12(4):2690–2719, Sept. 2023. (pp. [42](#) and [209](#).)
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. (p. [159](#).)
- D. Kivaranovic, K. D. Johnson, and H. Leeb. Adaptive, Distribution-Free Prediction Intervals for Deep Networks. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020. (p. [24](#).)
- R. Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. (p. [111](#).)
- J.-P. Kreiss and E. Paparoditis. The hybrid wild bootstrap for time series. *Journal of the American Statistical Association*, 107(499):1073–1084, 2012. (p. [95](#).)
- H. R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, 1217–1241, 1989. (p. [120](#).)
- J. Lago, F. De Ridder, and B. De Schutter. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221:386–405, 2018. (p. [62](#).)
- J. Lago, G. Marcjasz, B. De Schutter, and R. Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021. (pp. [5](#), [62](#), [106](#), and [222](#).)
- M. Le Morvan, J. Josse, T. Moreau, E. Scornet, and G. Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, 33, 5980–5990. Curran Associates, Inc., 2020a. (pp. [129](#), [150](#), and [168](#).)
- M. Le Morvan, N. Prost, J. Josse, E. Scornet, and G. Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 108, 3165–3174. PMLR, 2020b. (pp. [129](#), [148](#), [150](#), [168](#), [180](#), and [204](#).)
- M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux. What’s a good imputation to predict with missing values? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 34,

- 11530–11540. Curran Associates, Inc., 2021. (pp. 56, 129, 131, 138, 139, 156, 157, 168, and 183.)
- Y. Lee, E. Dobriban, and E. T. Tchetgen. Simultaneous conformal prediction of missing outcomes with propensity score ε -discretization, 2024. (p. 170.)
- J. Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4), 2019. (p. 43.)
- J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 2014. (pp. 23, 35, 36, 138, 144, 167, 175, and 176.)
- J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 2018. (pp. iii, v, 17, 21, 34, 62, 79, 114, 129, 132, 148, 152, 153, 185, and 224.)
- L. Lei and E. J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5): 911–938, Oct. 2021. (p. 50.)
- R. Liang and R. F. Barber. Algorithmic stability implies training-conditional coverage for distribution-free prediction methods. *arXiv preprint arXiv:2311.04295*, 2023. (p. 49.)
- Z. Lin, S. Trivedi, and J. Sun. Locally valid and discriminative prediction intervals for deep learning models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 34, 8378–8391. Curran Associates, Inc., 2021. (pp. 143 and 145.)
- R. J. A. Little. *Statistical analysis with missing data, third edition*. John Wiley & Sons, Nashville, TN, 3 edition, 2019. (pp. 129 and 168.)
- S. Loizidis, A. Kyprianou, and G. E. Georghiou. Electricity market price forecasting using elm and bootstrap analysis: A case study of the german and finnish day-ahead markets. *Applied Energy*, 363:123058, 2024. (p. 107.)
- K. Maciejowska, J. Nowotarski, and R. Weron. Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. *International Journal of Forecasting*, 32(3):957–965, 2016. (p. 62.)
- K. Maciejowska, B. Uniejewski, and R. Weron. Forecasting electricity prices, 2022. (p. 109.)
- S. Makridakis, E. Spiliotis, and V. Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022. (p. 112.)
- V. Manokhin. Awesome conformal prediction, Apr. 2022. <https://github.com/valeman/awesome-conformal-prediction>. (pp. 132 and 185.)

- G. Marcjasz, T. Serafin, and R. Weron. Selection of calibration windows for day-ahead electricity price forecasting. *Energies*, 11(9), 2018. (pp. 106 and 121.)
- G. Marcjasz, M. Narajewski, R. Weron, and F. Ziel. Distributional neural networks for electricity price forecasting. *Energy Economics*, 125:106843, 2023. (p. 107.)
- I. Mayer, A. Sportisse, J. Josse, N. Tierney, and N. Vialaneix. R-miss-tastic: a unified platform for missing values methods and workflows, 2019. (pp. 129 and 168.)
- N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35): 983–999, 2006. (p. 112.)
- MEPS. Medical expenditure panel survey. https://meps.ahrq.gov/mepsweb/data_stats/data_overview.jsp. (p. 141.)
- S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012. (pp. 83, 92, and 93.)
- B. M. Moreno, M. Brégère, P. Gaillard, and N. Oudjane. A mirror descent approach for mean field control applied to demande-side management. 2023. (p. 106.)
- B. Muzellec, J. Josse, C. Boyer, and M. Cuturi. Missing data imputation using optimal transport. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, 119, 7130–7140. PMLR, 2020. (p. 194.)
- N. Nassar, D. Silva, and H. Morais. Hierarchical energy management solution for smart charging. In *CIREP Porto Workshop 2022: E-mobility and power distribution systems*, 2022, 721–725, 2022. (p. 106.)
- E. Ndiaye. Stable conformal prediction sets. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022. (p. 43.)
- E. Ndiaye and I. Takeuchi. Computing full conformal prediction set with approximate homotopy. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. (p. 43.)
- E. Ndiaye and I. Takeuchi. Root-finding approaches for computing conformal prediction set. *Machine Learning*, 112(1), 2022. (p. 43.)
- D. Nickelsen and G. Müller. Bayesian hierarchical probabilistic forecasting of intraday electricity prices. *arXiv preprint arXiv:2403.05441*, 2024. (p. 107.)
- W. Nitka, T. Serafin, and D. Sotiros. Forecasting electricity prices: Autoregressive hybrid nearest neighbors (arhnn) method. In M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. Sloot, editors, *Computational Science – ICCS 2021*, 312–325, Cham, 2021. Springer International Publishing. (p. 106.)
- I. Nouretdinov, T. Melluish, and V. Vovk. Ridge regression confidence machine. In *Proceedings of the 18th International Conference on Machine Learning*, 2001. (p. 43.)

- J. Nowotarski and R. Weron. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30(3):791–803, Aug. 2014. (p. [115](#).)
- J. Nowotarski and R. Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018. (pp. [6](#), [62](#), and [223](#).)
- H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, 345–356. Springer, 2002. (pp. [iii](#), [v](#), [17](#), [20](#), [62](#), [63](#), [107](#), [113](#), [114](#), [129](#), [132](#), [148](#), [152](#), [185](#), and [224](#).)
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. (pp. [112](#), [139](#), and [191](#).)
- A. Podkopaev and A. Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. PMLR, 2021. (pp. [50](#) and [51](#).)
- D. N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994. (p. [120](#).)
- C. Remlinger, C. Alasseur, M. Brière, and J. Mikael. Expert aggregation for financial forecasting. *The Journal of Finance and Data Science*, 9:100108, 2023. (pp. [69](#) and [107](#).)
- Y. Romano, E. Patterson, and E. Candès. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. (pp. [vii](#), [24](#), [25](#), [27](#), [34](#), [56](#), [62](#), [114](#), [128](#), [129](#), [132](#), [134](#), [145](#), [146](#), [153](#), and [188](#).)
- Y. Romano, R. F. Barber, C. Sabatti, and E. Candès. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2), 2020a. (pp. [24](#), [150](#), [152](#), and [173](#).)
- Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020b. (pp. [31](#) and [32](#).)
- J. RTE. Bilan électrique 2022, 2022. (pp. [3](#), [106](#), and [220](#).)
- D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. (pp. [131](#) and [168](#).)
- M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, June 2018. (p. [30](#).)

- A. Saha, S. Basu, and A. Datta. Random forests for spatially dependent data. *Journal of the American Statistical Association*, 0(0):1–19, 2021. (p. [95](#).)
- N. Seedat, A. Jeffares, F. Imrie, and M. van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 206 of *Proceedings of Machine Learning Research*, 10160–10177. PMLR, 25–27 Apr 2023. (p. [170](#).)
- M. Sesia and E. J. Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020. (pp. [27](#), [41](#), [119](#), [138](#), [143](#), and [145](#).)
- M. Sesia and Y. Romano. Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021. (pp. [24](#), [139](#), [159](#), and [191](#).)
- G. Shafer and V. Vovk. A Tutorial on Conformal Prediction. *JMLR*, 9:51, 2008. (p. [62](#).)
- S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2-3):115–142, 2007. (p. [69](#).)
- M. Shao and Y. Zhang. Distribution-free matrix prediction under arbitrary missing pattern, 2023. (p. [170](#).)
- A. Sportisse, C. Boyer, A. Dieuleveut, and J. Josse. Debiasing averaged stochastic gradient descent to handle missing values. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, 33, 12957–12967. Curran Associates, Inc., 2020. (pp. [143](#) and [162](#).)
- H. Susmann, A. Chambaz, J. Josse, M. Wargon, P. Aegerter, and E. Bacry. Probabilistic Prediction of Arrivals and Hospitalizations in Emergency Departments in Île-de-France. working paper or preprint, Apr. 2024. (p. [123](#).)
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. (p. [111](#).)
- R. J. Tibshirani, R. F. Barber, E. Candès, and A. Ramdas. Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. (pp. [21](#), [50](#), [51](#), and [63](#).)
- L. Tschora, E. Pierre, M. Plantevit, and C. Robardet. Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy*, 313:118752, 2022. (p. [106](#).)
- A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, 2009. (pp. [201](#) and [202](#).)
- J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, 1975. (p. [217](#).)

- B. Uniejewski and R. Weron. Regularized quantile regression averaging for probabilistic electricity price forecasting. *Energy Economics*, 95:105121, 2021. (pp. [62](#) and [107](#).)
- B. Uniejewski, J. Nowotarski, and R. Weron. Automated variable selection and shrinkage for day-ahead electricity price forecasting. *Energies*, 9(8), 2016. (p. [106](#).)
- B. Uniejewski, R. Weron, and F. Ziel. Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, 33(2):2219–2229, 2018. (p. [106](#).)
- S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. (p. [184](#).)
- M. Van Ness, T. M. Bosschieter, R. Halpin-Gregorio, and M. Udell. The missing indicator method: From low to high dimensions, 2022. (pp. [129](#) and [168](#).)
- V. Vovk. Conditional Validity of Inductive Conformal Predictors. In *Asian Conference on Machine Learning*. PMLR, 2012. (pp. [23](#), [35](#), [36](#), [40](#), [49](#), [138](#), [145](#), [167](#), [171](#), [175](#), and [176](#).)
- V. Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74: 9–28, 2015. (pp. [47](#), [172](#), [189](#), and [208](#).)
- V. Vovk, A. Gammerman, and C. Saunders. Machine-Learning Applications of Algorithmic Randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 444–453. Morgan Kaufmann Publishers Inc., 1999. (pp. [62](#), [107](#), and [113](#).)
- V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer US, 2005. (pp. [iii](#), [v](#), [17](#), [20](#), [24](#), [34](#), [42](#), [62](#), [107](#), [113](#), [116](#), [129](#), [132](#), [133](#), [148](#), [171](#), [173](#), [184](#), [185](#), and [224](#).)
- V. G. Vovk. Aggregating strategies. *Proc. of Computational Learning Theory*, 1990. (pp. [69](#) and [94](#).)
- R. Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081, 2014. (pp. [5](#), [62](#), [106](#), and [222](#).)
- O. Wintenberger. Optimal learning with bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017. (pp. [69](#), [94](#), and [115](#).)
- W. Wisniewski, D. Lindsay, and S. Lindsay. Application of conformal prediction interval estimations to market makers’ net positions. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, 128 of *Proceedings of Machine Learning Research*, 285–301. PMLR, 2020. (pp. [63](#), [64](#), [72](#), [114](#), and [116](#).)
- S. N. Wood, Y. Goude, and M. Fasiolo. *Interpretability in Generalized Additive Models*, page 85–123. Springer International Publishing, 2022. (p. [123](#).)

- C. Xu and Y. Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, 139 of *Proceedings of Machine Learning Research*, 11559–11569. PMLR, 2021. (pp. [63](#), [72](#), [73](#), [75](#), and [95](#).)
- M. Yang. *Features Handling by Conformal Predictors*. PhD thesis, Royal Holloway, University of London, 2015. (p. [133](#).)
- Y. Yang, J. Guo, Y. Li, and J. Zhou. Forecasting day-ahead electricity prices with spatial dependence. *International Journal of Forecasting*, 2023. (p. [106](#).)
- Z. Yang, E. Candès, and L. Lei. Bellman conformal inference: Calibrating prediction intervals for time series, 2024. (p. [53](#).)
- M. Zaffran, O. Féron, Y. Goude, J. Josse, and A. Dieuleveut. Adaptive conformal predictions for time series. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, 162, 25834–25866. PMLR, 2022. (pp. [53](#), [107](#), [114](#), [116](#), [117](#), [118](#), and [141](#).)
- M. Zaffran, A. Dieuleveut, J. Josse, and Y. Romano. Conformal prediction with missing values. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, 202 of *Proceedings of Machine Learning Research*, 40578–40604. PMLR, 23–29 Jul 2023. (pp. [168](#), [169](#), [170](#), [171](#), [172](#), [177](#), [180](#), [185](#), [186](#), [189](#), [190](#), [191](#), [193](#), [194](#), and [204](#).)
- F. Ziel and R. Weron. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, 70: 396–420, feb 2018. (p. [109](#).)
- Çağatay Berke Bozlak and C. F. Yaşar. An optimized deep learning approach for forecasting day-ahead electricity prices. *Electric Power Systems Research*, 229:110129, 2024. (p. [106](#).)

Titre : Quantification post-hoc de l'incertitude prédictive : méthodes avec applications à la prévision des prix de l'électricité

Mots clés : Quantification d'incertitude prédictive, apprentissage statistique, prévision de séries temporelles, données manquantes, marchés de l'énergie

Résumé : L'essor d'algorithmes d'apprentissage statistique offre des perspectives prometteuses pour prévoir les prix de l'électricité. Cependant, ces méthodes fournissent des prévisions ponctuelles, sans indication du degré de confiance à leur accorder. Pour garantir un déploiement sûr de ces modèles prédictifs, il est crucial de quantifier leur incertitude prédictive. Cette thèse porte sur le développement d'intervalles prédictifs pour tout algorithme de prédiction. Bien que motivées par le secteur électrique, les méthodes développées, basées sur la prédiction conforme par partition (SCP), sont génériques : elles peuvent être appliquées dans de nombreux autres domaines sensibles.

Dans un premier temps, cette thèse étudie la quantification post-hoc de l'incertitude prédictive pour les séries temporelles. Le premier obstacle à l'application de SCP pour obtenir des prévisions probabilistes théoriquement valides des prix de l'électricité de manière post-hoc est l'aspect temporel hautement non-stationnaire des prix de l'électricité, brisant l'hypothèse d'échangeabilité. La première contribution propose un algorithme qui ne dépend pas d'un paramètre et adapté aux séries temporelles, reposant

sur l'analyse théorique de l'efficacité d'une méthode pré-existante, l'Inférence Conforme Adaptative. La deuxième contribution mène une étude d'application détaillée sur un nouveau jeu de données de prix spot français récents et turbulents en 2020 et 2021.

Un autre défi sont les valeurs manquantes (NAs). Dans un deuxième temps, cette thèse analyse l'interaction entre les NAs et la quantification de l'incertitude prédictive. La troisième contribution montre que les NAs induisent de l'hétéroscédasticité, ce qui conduit à une couverture inégale en fonction de quelles valeurs sont manquantes. Deux algorithmes sont conçus afin d'assurer une couverture constante quelque soit le schéma de NAs, ceci étant assuré sous des hypothèses distributionnelles sur les NAs. La quatrième contribution approfondit l'analyse théorique afin de comprendre précisément quelles hypothèses de distribution sont inévitables pour construire des régions prédictives informatives. Elle unifie également les algorithmes proposés précédemment dans un cadre général qui démontre empiriquement être robuste aux violations des hypothèses distributionnelles sur les NAs.

Title : Post-hoc predictive uncertainty quantification: methods with applications to electricity price forecasting

Keywords : Predictive uncertainty quantification, statistical learning, time series forecasting, missing values, energy markets

Abstract : The surge of more and more powerful statistical learning algorithms offers promising prospects for electricity prices forecasting. However, these methods provide ad hoc forecasts, with no indication of the degree of confidence to be placed in them. To ensure the safe deployment of these predictive models, it is crucial to quantify their predictive uncertainty. This PhD thesis focuses on developing predictive intervals for any underlying algorithm. While motivated by the electrical sector, the methods developed, based on Split Conformal Prediction (SCP), are generic : they can be applied in many sensitive fields.

First, this thesis studies post-hoc predictive uncertainty quantification for time series. The first bottleneck to apply SCP in order to obtain guaranteed probabilistic electricity price forecasting in a post-hoc fashion is the highly non-stationary temporal aspect of electricity prices, breaking the exchangeability assumption. The first contribution proposes a parameter-free algorithm tailored for time series,

which is based on theoretically analysing the efficiency of the existing Adaptive Conformal Inference method. The second contribution conducts an extensive application study on novel data set of recent turbulent French spot prices in 2020 and 2021.

Another challenge are missing values (NAs). In a second part, this thesis analyzes the interplay between NAs and predictive uncertainty quantification. The third contribution highlights that NAs induce heteroskedasticity, leading to uneven coverage depending on which features are observed. Two algorithms recovering equalized coverage for any NAs under distributional assumptions on the missingness mechanism are designed. The fourth contribution pushes forwards the theoretical analysis to understand precisely which distributional assumptions are unavoidable for theoretical informativeness. It also unifies the previously proposed algorithms into a general framework that demonstrates empirical robustness to violations of the supposed missingness distribution.