HAL
open science

Information auxiliaire non paramétrique : géometrie de l'information, processus empirique et applications

Sofiane Arradi-Alaoui

To cite this version:

Sofiane Arradi-Alaoui. Information auxiliaire non paramétrique : géometrie de l'information, processus empirique et applications. Théorie de l'information [cs.IT]. Université Paul Sabatier - Toulouse III, 2022. Français. NNT : 2022TOU30096 . tel-03783626

HAL Id: tel-03783626

https://theses.hal.science/tel-03783626

Submitted on 22 Sep 2022

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.

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *13 juillet 2022* **par :**

ARRADI-ALAOUI Sofiane

Information auxiliaire non paramétrique : géométrie de l'information, processus empirique et applications

JURY

Philippe Berthet	Univ. Toulouse Paul Sabatier	Directeur de thèse
Elena Di Bernardino	Univ. Côte d'Azur, Nice	Rapporteur
Matthieu Lerasle	CNRS, CREST-Ensae, Paris	Rapporteur
Stéphane Puechmorel	ENAC, Toulouse	Membre du Jury
Jean-Claude Fort	Univ. Paris Descartes	Membre du Jury
Agnès Lagnoux	Univ. Toulouse Jean Jaurès	Membre du Jury

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées
Unité de Recherche :
Institut de Mathématiques de Toulouse (UMR 5219)
Directeur de Thèse :
Philippe Berthet
Rapporteurs :
Elena Di Bernardino et *Matthieu Lerasle*

Remerciements

En premier lieu, je souhaite remercier mon directeur de thèse Philippe Berthet qui m'a transmis sa passion de la statistique mathématique et sans qui cette thèse n'aurait jamais pu voir le jour. Sa gentillesse, sa bienveillance et son optimisme, caractéristique de mon directeur, m'ont permis d'entreprendre sereinement cette thèse et de rebondir lorsque je n'arrivais pas au bout de certaines pistes. De plus, ses idées profondes en mathématiques ont été pour moi une source d'inspiration bien que je n'ai pas eu le temps de poursuivre chacune d'entre elles. Enfin, je le remercie du temps précieux qu'il a su me consacrer tout au long de cette thèse.

Je remercie Elena Di Bernardino et Matthieu Lerasle d'avoir rapporté ma thèse et de m'avoir aidé à améliorer mon manuscrit. Je remercie également les membres du jury, Stéphane Puechmorel, Jean-Claude Fort et Agnès Lagnoux de leur présence à ce jury.

Je remercie l'ensemble des enseignants qui m'ont encouragé et aidé tout au long de mes études à commencer par mon professeur de mathématiques (Monsieur Anzelin) au lycée en terminale qui m'a donné l'envie d'étudier les mathématiques. Je remercie aussi particulièrement Franck Barthe et Philippe Berthet de m'avoir transmis le goût des probabilités et statistiques.

Je tiens également à remercier l'ensemble des doctorants de l'IMT à commencer par les doctorants avec qui j'ai partagé mon bureau. Enfin je remercie particulièrement Pierre, Laetitia, Mehdi, William et Clément avec qui j'ai pu échanger et rigoler sur des sujets divers et variés lors de nos longues conversations. Je leur souhaite la réussite dans l'ensemble des projets qu'ils entreprennent. Je remercie aussi l'ensemble des membres permanents de l'IMT que j'ai pu côtoyer tout au long de ce doctorat.

De plus, je tiens aussi à remercier Laurent Bakri et Guillaume Loizelet qui m'ont apporté de précieux conseils au niveau de l'enseignement et qui m'ont aidé à mûrir mon projet professionnel.

Enfin je finis ces remerciements par les personnes les plus chères dans ma vie à savoir ma famille

et mon épouse. Je remercie infiniment ma famille (mon père, ma mère, ma soeur et mon frère) qui ont su me soutenir lors de mes études et de mon doctorat. Leur aide et leur présence ont été essentiels à la réussite de mes études. Concernant mon épouse, que dire si ce n'est que je l'aime tellement et que je la remercie infiniment d'avoir été à mon coté durant ces trois années de thèse et de m'avoir soutenu à tout moment.

Table des matières

1 Approche géométrique de la notion d'information auxiliaire **10**

1.1 Motivations du chapitre 1 10

1.2 Préliminaires techniques : cadre géométrique 11

1.2.1 Métrique de Fisher, sous variété autoparallèle et projection de connexion . . . 11

1.2.2 α-connexion 12

1.2.3 Structure duale d'une variété 15

1.2.4 Construction d'une structure riemannienne duale à partir d'une divergence . . 16

1.2.5 Structure duale plate (DFS) et système dual de coordonnées 17

1.2.6 Divergence canonique construite à partir d'une DFS et relation triangulaire . . 20

1.2.7 Théorème de Pythagore généralisé et théorème de projection 21

1.2.8 α-variété et α-famille 23

1.2.9 Structure duale du modèle exponentiel et de mélange et divergence canonique associée à chaque modèle 25

1.3 Injection optimale d'une information auxiliaire 27

1.3.1 Projection de la mesure empirique sur $\mathscr{P}^I$ 27

1.3.2 Structure géométrique de $\mathscr{P}^I$ dans le cas d'une information auxiliaire forte apportée par des espérances 29

2 Mesure empirique informée par une information auxiliaire forte **35**

2.1 Motivations du chapitre 2 35

2.2 Étude des deux problèmes d'optimisation 36

2.2.1 Premier problème d'optimisation 36

2.2.2 Second problème d'optimisation 44

2.3 Approximation commune de la solution de ces deux problèmes d'optimisation et définition de la mesure empirique informée 51

2.4 Généralisation à une information auxiliaire forte donnée par une fonctionnelle de la mesure P et étude de la répartition des poids sur l'échantillon 59

2.4.1 Généralisation à une information auxiliaire forte donnée par une fonctionnelle de la mesure P . . . 59
2.4.2 Étude de la répartition des poids sur l'échantillon . . . 62

3 Résultats asymptotiques pour l'injection d'information auxiliaire forte 65
3.1 Résultats du type Glivenko-Cantelli et Donsker sous des hypothèses minimales dans le cas d'information auxiliaire donnée par des espérances . . . 65
3.2 Extension des résultats asymptotiques à des fonctionnelles . . . 70

4 Injection d'information auxiliaire faible et résultats asymptotiques 76
4.1 Information auxiliaire faible issue de sources indépendantes à l'échantillon initial . . . 76
4.1.1 Information auxiliaire faible donnée par des estimations issues de sources indépendantes . . . 76
4.1.2 Ajout de données auxiliaires à l'échantillon initial . . . 84
4.1.3 Un mélange des deux informations auxiliaires faibles précédentes . . . 86
4.2 Information auxiliaire donnée par une mesure des préférences d'un expert . . . 89
4.2.1 Choix entre un nombre fini de valeurs et généralisation au cas infini . . . 89
4.2.2 Information auxiliaire donnée par une densité de probabilité . . . 93

5 Concentration de la mesure empirique informée 95
5.1 Concentration du processus empirique informé et application . . . 95
5.2 Borne de concentration pour le processus empirique informé . . . 99

6 Information auxiliaire fausse et adaptativité de la mesure informée 109
6.1 Impact et détection d'une information auxiliaire forte fausse . . . 109
6.2 Adaptativité de la mesure informée par une information auxiliaire forte pour l'estimation de Pf . . . 112
6.3 Adaptativité de la mesure informée au temps n par une information auxiliaire générale en apprentissage . . . 115

7 Applications 118
7.1 Méthode de Monte-Carlo, variables auxiliaires et données manquantes massives . . . 119
7.1.1 Amélioration de la méthode de Monte-Carlo à l'aide d'une information auxiliaire 119
7.1.2 Utilisation de variables auxiliaires . . . 121
7.1.3 Données manquantes massives et information auxiliaire . . . 122
7.2 Applications en statistique paramétrique et non paramétrique . . . 127
7.2.1 Estimateur du maximum de vraisemblance informé et matrice d'information de Fisher informée . . . 127
7.2.2 Méthode des moments avec informations auxiliaires . . . 132
7.2.3 Estimateur à noyau informé . . . 133
7.2.4 Quantiles empiriques informés . . . 139
7.3 Estimateur des moindres carrés avec informations auxiliaires et modèle linéaire informé 144
7.3.1 Estimateur des moindres carrés avec informations auxiliaires . . . 144
7.3.2 Propriétés asymptotiques de l'estimateur des moindres carrés informé . . . 145

7.3.3 Interprétation géométrique et qualité d'ajustement dans le cadre d'un modèle linéaire avec informations auxiliaires . . . 148
7.3.4 Variables instrumentales et informations auxiliaires . . . 150
7.3.5 Tests statistiques asymptotiques sur le modèle informé (avec ou sans instrument) 152
7.4 Pistes méthodologiques : Analyse des données et apprentissage statistique avec informations auxiliaires . . . 154
7.4.1 Problème de bandits avec information auxiliaire . . . 154
7.4.2 Analyse en composantes principales avec informations auxiliaires . . . 155
7.4.3 Réseaux de neurones . . . 156
7.4.4 Bagging informé . . . 157
7.4.5 Algorithme des k-plus proches voisins avec informations auxiliaires . . . 157
7.4.6 Apprentissage non supervisé : k-moyennes avec informations auxiliaires . . . 158

Appendices **162**

A Annexe A : Rudiments de géométrie différentielle et riemannienne **163**
A.1 Variétés topologiques et différentielles . . . 163
A.2 Différentiabilité, sous variétés et plongements . . . 164
A.3 Espaces tangents : courbes et dérivations . . . 166
A.4 Différentielle, théorème d'inversion locale et théorème des fonctions implicites . . . 169
A.5 Fibré vectoriel et fibré tangent . . . 170
A.6 Champ de vecteurs et crochet de Lie . . . 173
A.7 Variété riemannienne, longueurs et fibré normal . . . 175
A.8 Connexions et symboles de Christoffel . . . 178
A.9 Géodésiques et champ parallèle le long d'une courbe . . . 181
A.10 Courbure et torsion . . . 184

B Annexe B : Optimal use of auxiliary information : information geometry and empirical process **189**
B.1 Introduction . . . 189
B.2 Framework . . . 191
B.3 A geometrical approach of auxiliary information . . . 191
B.3.1 Submanifold structure of $\mathscr{P}^I(\mathscr{Z})$ . . . 192
B.3.2 Existence of a projection . . . 196
B.4 Two measure projections . . . 197
B.4.1 First optimization problem . . . 197
B.4.2 Second optimization problem . . . 198
B.5 Towards a definition of the informed empirical measure . . . 204
B.5.1 Equivalence of projections and the informed empirical measure . . . 204
B.5.2 Weights of the informed empirical measure . . . 207
B.6 Asymptotic results and concentration . . . 210
B.6.1 P-Glivenko-Cantelli and P-Donsker properties under minimal assumptions . 210
B.6.2 Concentration of the informed empirical process . . . 214
B.7 Informed empirical quantiles . . . 216

Introduction

La notion d'information auxiliaire regroupe toute information extérieure à l'expérience statistique observée. Cette notion d'information auxiliaire tire son origine en théorie des sondages quelques siècles plus tôt. Plus précisément vers 1740, le magistrat Jean-Baptiste François de La Michodière voulut estimer la taille de la population française en supposant que le nombre des naissances, des mariages et des morts est proportionnel à la population entière. Cette hypothèse est donc une des premières traces d'une information auxiliaire dans l'histoire de la statistique. Elle permit de donner une approximation de la taille de la population française en multipliant le nombre de naissances par un facteur qui pouvait varier selon la personne et le lieu d'intérêt (par exemple la localité). Cette méthode, appelée estimateur par le ratio, a été validée par Laplace dans son mémoire *Sur les naissances, les mariages et les morts* [24] et fût très utilisée au sein de l'administration française de l'époque. On pourra consulter à ce sujet l'article de Bernard Bru [12].

Dans les siècles suivants, plusieurs auteurs en statistique se sont intéressés à développer des méthodes d'injection d'une information auxiliaire. Nous pouvons citer par exemple les références récentes [31], [37], [5], [4],[1]. Dans [5], la méthode du Raking-Ratio est utilisée afin d'incorporer une information auxiliaire donnée par les probabilités d'une ou plusieurs partitions selon un principe qui remonte à S.Kullback et C.T.Ireland [22]. Cette méthode est une procédure séquentielle permettant d'injecter une information auxiliaire l'une après l'autre. Dans [31], *Owen* a développé la méthode de la vraisemblance empirique qui peut être utilisée pour injecter une information auxiliaire donnée par des espérances. Cette méthode est devenue une méthode standard en statistique mathématique mais plus rarement utilisée sous l'angle d'une information auxiliaire. Dans [37], une méthode a été développée par *Tarima* et *Pavlov* afin d'injecter une information auxiliaire générale. Cette dernière consiste à minimiser la variance sur une classe d'estimateurs sans biais. La thèse de *Mickael Albertus* [4] utilise cette approche pour injecter une information auxiliaire apportée par des espérances, comme extension naturelle de probabilités d'ensembles comme [5].

Cependant il n'existe pas à proprement parler de théorie générale de l'information auxiliaire permettant de relier ces méthodes. Ainsi il est difficile de choisir parmi ces approches laquelle

serait optimale pour incorporer une information extérieure à une expérience statistique donnée. De ce fait, nous proposons de rechercher une mesure de probabilité discrète ayant pour support l'échantillon vérifiant l'information auxiliaire et qui soit la plus *proche* de la mesure empirique au sens de la géométrie des mesures de probabilité. Ainsi la motivation principale de cette thèse est de développer une méthode d'injection optimale d'une information auxiliaire au sens de la géométrie de l'information. Cela nous a amené, dans un premier temps, à clarifier la notion vague d'information auxiliaire en la classifiant en deux catégories :

1. **Une information auxiliaire forte** est une information auxiliaire donnée par la connaissance d'une ou plusieurs espérances et plus généralement par des fonctionnelles de la loi P qui génère les données. Entre dans ce cadre, par exemple, la connaissance d'un moment de la loi P, un quantile, la probabilité d'un évènement, la connaissance de la variance ou bien l'injection de variables auxiliaires dans le cadre d'un sondage, certaines formes de calibration.
2. **Une information auxiliaire faible** est une information auxiliaire amoindrie par rapport aux informations auxiliaires fortes de la première catégorie. Par exemple, l'information auxiliaire peut être donnée par une estimation issue de sources indépendantes ou bien par une mesure des préférences d'un ou plusieurs experts sur un ensemble de choix restreints. Cette seconde catégorie sera étudiée dans le chapitre 4.

Pour une information auxiliaire donnée, nous souhaitons donc modifier de manière optimale les poids uniforme de la mesure empirique $\mathbb{P}_n$ afin d'incorporer cette information exacte. Dans un premier temps, nous commençons par préciser dans quel sens l'optimalité doit être comprise. Puisque nous cherchons la mesure informée la plus proche de la mesure empirique, il a semblé naturel de considérer que la voie géométrique est le cadre privilégié pour répondre à cette question. La proximité des mesures de probabilités est comprise au sens de la géométrie de l'information. Pour cela, nous considérons la mesure empirique $\mathbb{P}_n$ comme un point d'une certaine variété et nous souhaitons définir une notion de projection orthogonale de $\mathbb{P}_n$ sur la sous-variété dite informée. Ainsi nous avons appliqué les théorèmes principaux de la géométrie de l'information afin d'en déduire une méthode d'injection de l'information auxiliaire forte (théorème 1.3.1). Cette approche fait l'objet du chapitre 1.

À l'issue du chapitre 1, nous disposons d'une méthode d'injection de l'information auxiliaire forte. Cette dernière établit qu'en fonction de l'autoparallélisme de la sous variété informée la projection orthogonale de $\mathbb{P}_n$ est solution d'un des deux problèmes de minimisation mentionnés dans le théorème 1.3.1. Dans le chapitre 2, nous étudions ces deux problèmes d'optimisation en établissant l'existence et l'unicité des solutions sous certaines hypothèses. Ainsi nous obtenons deux mesures empiriques distinctes données par les égalités (2.10) et (2.11) dont les poids ne sont malheureusement pas explicites. C'est pourquoi nous prouvons qu'il existe une approximation commune explicite de ces deux familles de poids vérifiant l'information auxiliaire forte. Par la suite, nous définissons la mesure empirique informée notée $\mathbb{P}_n^I$ à l'aide de ces poids approximés. Nous montrons que cette dernière admet les mêmes propriétés asymptotiques que les deux mesures empiriques précédentes et que la répartition des poids est quasi similaire à ces deux dernières. À la fin du chapitre 2, nous généralisons la définition de la mesure empirique informée à des informations auxiliaires fortes données par des fonctionnelles de la loi P et nous étudions la répartition des poids

de la mesure empirique informée $\mathbb{P}_n^I$.

Dans le chapitre 3, nous étudions les propriétés asymptotiques de la mesure empirique informée dans le cadre d'une information auxiliaire forte. Plus précisément, nous établissons des résultats du type P-Glivenko-Cantelli et P-Donsker pour la mesure $\mathbb{P}_n^I$. De plus ces résultats sont énoncés sous des hypothèses minimales. Dans un premier temps, nous prouvons ces résultats dans le cas d'une information auxiliaire forte apportée par la connaissance d'une ou plusieurs espérances puis dans un second temps nous généralisons cela pour une information auxiliaire forte donnée par une fonctionnelle de la loi P. Comme nous remarquons une diminution uniforme de la variance asymptotique du processus limite, la méthode d'injection d'une information auxiliaire via $\mathbb{P}_n^I$ est ainsi bien justifiée.

Dans le chapitre 4, nous souhaitons injecter une information auxiliaire faible. Dans un premier temps, nous étudions les informations auxiliaires faibles issues de sources indépendantes. Par exemple l'information auxiliaire peut être donnée par des données auxiliaires indépendantes ou par une estimation d'une espérance en utilisant une base de données indépendantes. Dans un second temps, nous étudions les informations auxiliaires données par une mesure de préférence d'un expert. Dans les deux sections, nous généralisons la définition de la mesure empirique informée à une information auxiliaire faible puis nous établissons des résultats asymptotiques du type P-Glivenko-Cantelli et P-Donsker. Ceci met de nouveau en évidence une diminution uniforme de la variance asymptotique du processus limite.

Dans le chapitre 5, nous étudions la concentration à n fixé. Nous montrons que le processus empirique informé $\sqrt{n}\left(\mathbb{P}_n^I - P\right)$ est plus concentré que le processus empirique standard $\sqrt{n}\left(\mathbb{P}_n - P\right)$ pour une information auxiliaire générale. Ainsi la mesure empirique informée est plus proche de P que ne l'est la mesure empirique $\mathbb{P}_n$ à partir d'un certain n. Cela montre que l'injection d'une information auxiliaire est bénéfique dans un cadre non asymptotique. De plus, dans le cas d'une information auxiliaire forte apportée par des espérances, nous établissons une borne de concentration plus fine pour la mesure empirique informée que la borne usuelle donnée par l'inégalité de Bernstein.

Dans le chapitre 6, nous nous intéressons à l'impact de l'utilisation d'une information auxiliaire fausse et à des problématiques d'adaptativité. Plus précisément, nous souhaitons mesurer l'impact négatif d'une information auxiliaire fausse, détecter les informations auxiliaires fausses et sélectionner les informations auxiliaires réellement informatives pour un objectif donné comme par exemple l'estimation d'un paramètre d'intérêt. Notre démarche est structurée de la manière suivante. Dans un premier temps, on s'intéresse à l'impact de l'utilisation d'une information auxiliaire forte fausse et nous présentons un test statistique afin de détecter les informations fausses parmi un ensemble d'informations auxiliaires fortes. Dans un second temps, nous proposons une méthode pour sélectionner les informations auxiliaires réellement informatives au temps n afin d'estimer au mieux un paramètre d'intérêt. Dans un troisième temps, nous nous plaçons dans un contexte d'apprentissage statistique et classons les informations auxiliaires en utilisant une approche par validation croisée.

Dans le chapitre 7, nous montrons comment utiliser la mesure empirique informée dans

quelques problèmes classiques en statistique, et en particulier comment reconnaître l'information auxiliaire en pivot de nouvelles méthodes. L'objectif de ce chapitre est de prouver que l'injection d'une information auxiliaire peut être pertinente dans des situations pratiques variées. Nous proposons des applications à des méthodes génériques en vigueur en simulation stochastique et en statistique. Dans la première section, nous proposons dans un premier temps une amélioration significative de la méthode de Monte-Carlo. Puis par la suite nous nous intéressons à l'utilisation de variables auxiliaires et au problème de données manquantes massives dans un contexte d'apprentissage statistique. En particulier nous montrons que l'utilisation des données partielles vues comme une information auxiliaire peut être bénéfique pour l'estimation statistique. Dans la seconde section, nous revisitons les principales méthodes d'estimation en statistique paramétrique et non paramétrique en injectant de l'information auxiliaire, pour tirer partie de la moindre variance asymptotique de $\mathbb{P}_n^I$. Dans la troisième section, nous étudions l'estimateur des moindres carrés avec informations auxiliaires dans le cadre du modèle linéaire. Nous montrons que l'information impactante doit porter sur la variable expliquée elle-même. Enfin dans la quatrième section, nous donnons des pistes méthodologiques pour incorporer une information auxiliaire à des problèmes d'apprentissage statistique couvrant un large spectre.

De plus, nous avons joint deux annexes. La première annexe porte sur des rudiments de géométrie différentielle et riemannienne nécessaires à la compréhension du chapitre 1. La seconde annexe porte sur un article que nous avons prépublié [8].

À travers ces chapitres, nous avons pu définir une mesure empirique informée $\mathbb{P}_n^I$ ayant de meilleures propriétés asymptotiques que $\mathbb{P}_n$ et dont les poids ont une forme close. Elle est ainsi très facilement calculable d'un point de vue algorithmique. Cette mesure empirique informée possède les mêmes propriétés asymptotiques que les deux mesures empiriques optimales au sens de la géométrie de l'information données par les égalités (2.10) et (2.11). De plus, nous avons montré que l'utilisation d'une information auxiliaire est bénéfique pour améliorer les méthodes d'estimation en statistique et en simulation stochastique. Des méthodes nouvelles peuvent émerger en reconnaissant des éléments d'un problème complexe pouvant jouer le rôle pivot d'une information auxiliaire. Cela est d'autant plus intéressant du fait que notre définition de l'information auxiliaire permet de couvrir un panel assez large de ce qu'on pourrait encore imaginer comme information auxiliaire. Cela ouvre la voie à l'utilisation d'une information auxiliaire par la mesure empirique informée à l'ensemble des méthodes d'apprentissage statistique (réseaux de neurones, bagging etc). Des pistes méthodologiques ont été proposées en fin de chapitre 7 mais qui n'ont pas encore été développées.

1

Approche géométrique de la notion d'information auxiliaire

Ce chapitre a pour objectif de poser le cadre géométrique permettant de projeter la mesure empirique sur l'ensemble des mesures de probabilités discrètes à support dans l'échantillon noté $\mathscr{P}^I$ afin d'en déduire une méthode d'injection de l'information auxiliaire. De plus, nous étudions la structure géométrique $\mathscr{P}^I$ dans le cas d'une information auxiliaire forte apportée par des espérances.

1.1 Motivations du chapitre 1

Dans ce chapitre, on s'intéresse à la question suivante :

Comment injecter une information auxiliaire de manière optimale ?

La notion d'optimalité peut faire débat mais l'approche géométrique semble être une réponse cohérente et intéressante pour l'injection d'information auxiliaire. Le cadre de ce chapitre est le suivant. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et identiquement distribuées (*i.i.d.*) à valeurs dans $\mathscr{Y}$ de loi P inconnue. Soit $n \in \mathbb{N}^*$. On note $\mathbb{P}_n$ la mesure empirique définie par

$$\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

où δ_{X_i} désigne la mesure (aléatoire) de Dirac en X_i pour tout $i \in [\![1, n]\!]$. Posons $\mathscr{X} = \{X_1, \cdots, X_n\}$ l'ensemble aléatoire des observations et $\mathscr{P}(\mathscr{X})$ l'ensemble des mesures de probabilité sur $\mathscr{X}$ à poids strictement positifs. On suppose qu'on dispose d'une information auxiliaire notée I et on note $Q \sim I$ si Q satisfait l'information auxiliaire. Par exemple, si l'information auxiliaire I est donnée par Pg alors

$$Q \sim I \Longleftrightarrow Qg = Pg.$$

L'ensemble des mesures de probabilité $Q \in \mathscr{P}(\mathscr{X})$ vérifiant l'information auxiliaire I est noté

$$\mathscr{P}^I = \left\{Q \in \mathscr{P}(\mathscr{X}),\ \ Q \sim I\right\}.$$

L'objectif de ce chapitre est de poser un cadre géométrique permettant de projeter la mesure empirique $\mathbb{P}_n \in \mathscr{P}(\mathscr{X})$ sur $\mathscr{P}^I$ et d'en déduire une méthode optimale d'injection de l'information auxiliaire.

Dans la première section, on souhaite définir **au sens de la géométrie de l'information** la notion de **projection** sur une sous variété. On supposera connues les bases de la géométrie différentielle et riemannienne. On pourra consulter à ce sujet l'annexe A.

L'objectif de la seconde section est d'en déduire une **méthode d'injection de l'information auxiliaire**.

Remarque 1.1.1. *Afin d'obtenir rapidement une réponse à cette introduction, on pourra directement consulter la section 1.3 pour une première lecture.*

1.2 Préliminaires techniques : cadre géométrique

Cette section a pour seul objectif de poser le cadre géométrique nécessaire à la définition de la projection. Les lecteurs non familiers avec les outils de la géométrie riemannienne pourront directement consulter l'annexe A pour des rappels ou bien admettre le cadre géométrique et passer à la section 1.3 permettant d'obtenir la méthode optimale d'injection de l'information auxiliaire. Les résultats de cette section sont tirés du livre Methods of Information Geometry de Amari et Nagaoka [6]. On pourra aussi consulter les démonstrations des énoncés de cette section dans cet ouvrage.

Concernant les notations, on omettra à certains moments de cette section de préciser la carte dans laquelle on se place (sous entendu $\theta \in \Theta$) et les indices (sous-entendus $i, j, k \in [\![1, m]\!]$ par exemple). De plus, on utilisera souvent la convention de sommation d'Einstein à savoir $y = \sum_{i=1}^n c_i x^i$ sera noté $y = c_i x^i$.

1.2.1 Métrique de Fisher, sous variété autoparallèle et projection de connexion

Soit $\mathscr{P}$ une famille de mesures de probabilités sur un espace topologique $\mathscr{X}$. Alors $\mathscr{P}$ hérite de la structure topologique de $\mathscr{X}$ par la topologie faible des mesures [25]. On suppose que $\mathscr{P}$ peut être paramétrisée au moins localement par des homéomorphismes

$$\psi : U \to \Theta \subset \mathbb{R}^m$$

où U est un ouvert de $\mathscr{P}$ et Θ un ouvert de $\mathbb{R}^m$. On peut alors munir $\mathscr{P}$ d'une structure de variété différentielle pourvu que les paramétrisations locales soient **compatibles**. Plaçons nous pour un temps dans une carte munie d'une certaine paramétrisation. En se plaçant dans une carte, on peut alors décrire $\mathscr{P}$ au moins localement comme $\{P_\theta, \theta \in \Theta\}$. On supposera que le modèle statistique associé est dominé par une mesure μ et on notera p_θ la densité de Radon-Nikodym de P_θ par rapport à la mesure μ pour tout $\theta \in \Theta$. De plus, on supposera que μ-pp $x, \theta \mapsto p_\theta(x)$ est C^∞. Sous

ces conditions, on peut équiper $\mathscr{P}$ d'une structure de variété riemannienne. Notons pour tout $\theta \in \Theta$, $l_\theta = \log(p_\theta)$ la fonction score et définissons la **métrique de Fisher** pour tout $i, j \in [\![1, m]\!]$

$$g_{i,j}(\theta) = g_{P_\theta}(\partial_i, \partial_j) = E_\theta((\partial_i l_\theta)(\partial_j l_\theta))$$

où $\partial_i = \frac{\partial}{\partial \theta_i}$. En notant $I(\theta) = \left(g_{i,j}(\theta)\right)_{i,j \in [\![1,m]\!]}$ l'information de Fisher, la métrique de Fisher peut s'écrire

$$g_\theta(u, v) = u^T I(\theta) v$$

où $u, v \in \mathbb{R}^m$.

Soit ∇ une connexion sur la variété différentielle $\mathscr{P}$ et soit M une sous variété de $\mathscr{P}$. On rappelle qu'on note TM le fibré tangent de la sous variété M, T_pM l'espace tangent de M au point p et $C^\infty(TM)$ l'ensemble des sections du fibré tangent c'est à dire des champs de vecteurs. On dit que M est **autoparallèle** si pour tout $X, Y \in C^\infty(TM)$

$$\nabla_X Y \in C^\infty(TM).$$

Notons $[\cdot, \cdot]$ le crochet de Lie. Dans le cas où la sous variété n'est pas autoparallèle, il est possible de définir une projection de ∇ sur M. Supposons qu'on dispose pour tout $p \in M$ d'une application linéaire $\pi_p : T_p\mathscr{P} \to T_pM$ telle que pour tout $D \in T_pM$, $\pi_d(D) = D$. De plus, supposons aussi que $p \mapsto \pi_p$ est C^∞. On définit, pour tout $X, Y \in C^\infty(TM)$, $\nabla_X^{(\pi)} Y \in C^\infty(TM)$ par

$$(\nabla_X^{(\pi)} Y)_p = \pi_p((\nabla_X Y)_p).$$

Si g est une métrique riemannienne, on peut prendre la projection orthogonale vérifiant

$$g_p(\pi_p(D), D') = g_p(D, D')$$

pour tout $D \in T_p\mathscr{P}$ et $D' \in T_pM$. Dans ce cas, on appelle $\nabla^{(\pi)}$ la projection de ∇ sur M par rapport à la métrique riemannienne g. De plus, un fait remarquable est que si ∇ est sans torsion alors $\nabla^{(\pi)}$ est aussi sans torsion.

Un théorème utile reliant autoparallélisme et planéité est le suivant :

Théorème 1.2.1. *Soient S une variété et M une sous variété. Si S est plate, alors une condition nécessaire et suffisante pour que M soit autoparallèle est que M soit exprimée en tant que sous espace affine (ou ouvert d'un sous espace affine) de S par rapport à un système de coordonnées affines. De plus si M est autoparallèle alors M est aussi plate.*

1.2.2 α-connexion

Dans cette sous-section, nous souhaitons définir la notion de α-connexion avec $\alpha \in \mathbb{R}$. Pour cela, définissons tout d'abord pour tout $i, j, k \in [\![1, m]\!]$ les fonctions $\Gamma_{ij,k}^{(\alpha)}$ de la manière suivante dans une carte

$$\Gamma_{ij,k}^{(\alpha)}(\theta) = E_\theta\left[\left(\partial_i \partial_j l_\theta + \frac{1-\alpha}{2} \partial_i l_\theta \partial_j l_\theta\right) \partial_k l_\theta\right]$$

où $\alpha \in \mathbb{R}$. On définit alors la connexion $\nabla^{(\alpha)}$ par

$$g\left(\nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k\right) = \Gamma^{(\alpha)}_{ij,k}$$

où g est la métrique de Fisher. On appelle $\nabla^{(\alpha)}$ la α-**connexion**. De plus la connexion est **symétrique** c'est à dire sans torsion car $\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(\alpha)}_{ji,k}$ et donc puisque le crochet de Lie $[\partial_i, \partial_j] = 0$, on a que $\nabla^{(\alpha)}_{\partial_i}\partial_j = \nabla^{(\alpha)}_{\partial_j}\partial_i$. La β-connexion (avec $\beta \in \mathbb{R}$) et la α-connexion sont reliées de la manière suivante

$$\Gamma^{(\beta)}_{ij,k} = \Gamma^{(\alpha)}_{ij,k} + \frac{\alpha - \beta}{2} T_{ijk} \tag{1.1}$$

où T_{ijk} est un tenseur symétrique de type $(3,0)$ défini par

$$T_{ijk}(\theta) = E_\theta(\partial_i l_\theta \partial_j l_\theta \partial_k l_\theta).$$

En utilisant (1.1), on montre que

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}.$$

En utilisant que $g_{ij}(\theta) = E_\theta(\partial_i l_\theta \partial_j l_\theta)$, on a

$$\partial_k g_{ij} = \Gamma^{(0)}_{ki,j} + \Gamma^{(0)}_{kj,i}.$$

C'est la traduction dans une carte de la définition d'une connexion métrique. D'où la 0-connexion est métrique, sans torsion qui est donc la connexion de Levi-Civita par rapport à la métrique de Fisher. En général pour $\alpha \neq 0$, $\nabla^{(\alpha)}$ n'est pas métrique.

Exemple 1.2.1. Nous présentons deux modèles statistiques et nous étudions leur planéité par rapport à la 1-connexion et la (-1)-connexion.

- Concernant le **modèle exponentiel**, les densités associées à ce modèle sont données pour tout $x \in \mathcal{X}$ par

$$p_\theta(x) = \exp\left(C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)\right)$$

où θ^i est la coordonnée en F_i (fonction de θ). Puisque p_θ est une densité, on en déduit que

$$\psi(\theta) = \log\left(\int \exp\left(C(x) + \sum_{i=1}^{n} \theta^i F_i(x)\right) d\mu(x)\right).$$

L'application $\theta \mapsto P_\theta$ est injective si et seulement si $\{1, F_1, \cdots, F_n\}$ est une famille libre. Dans le cas d'une loi $\mathcal{N}(\mu, \sigma^2)$, on a

$$C(x) = 0,\ F_1(x) = x,\ F_2(x) = x^2,\ \theta^1 = \frac{\mu}{\sigma^2},\ \theta^2 = -\frac{1}{2\sigma^2},\ \psi(\theta) = \frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi}\sigma).$$

En se plaçant dans la carte $\phi(\theta) = (\theta^1, \cdots, \theta^n)$ où $\theta^i = \phi_i(\theta)$ et en notant $\partial_i = \frac{\partial}{\partial \theta^i}$, on a

$$\begin{aligned} \partial_i l_\theta &= F_i(x) - \partial_i \psi(\theta), \\ \partial_i \partial_j l_\theta &= -\partial_i \partial_j \psi(\theta). \end{aligned}$$

D'où $\Gamma^{(1)}_{ij,k}(\theta) = -\partial_i \partial_j \psi(\theta) E_\theta(\partial_k l_\theta)$. En supposant certaines conditions de régularité (interversion intégrale et dérivée), on a $\Gamma^{(1)}_{ij,k}(\theta) = 0$. On obtient que **le modèle exponentiel est 1-plat** c'est à dire plat pour la 1-connexion. On dit aussi que $\phi = [\theta^i]$ est un 1-**système de coordonnées affines** c'est à dire que pour tout $i = 1, ..., n$, $\partial_i = \frac{\partial}{\partial \theta^i}$ est parallèle à $\mathscr{P}$. On appelle $\nabla^{(1)}$ la **connexion exponentielle** notée $\nabla^{(e)}$. Ainsi $\mathscr{P}$ est dit e-plat.

- Maintenant supposons que le modèle statistique soit un **famille de mélange** c'est à dire pour tout $x \in \mathscr{Z}$

$$\begin{aligned} p_\theta(x) &= \sum_{i=1}^m \theta^i p_i(x) + \left(1 - \sum_{i=1}^m \theta^i\right) p_0(x) \\ &= p_0(x) + \sum_{i=1}^m \theta^i (p_i(x) - p_0(x)). \end{aligned}$$

où $\{p_0, \cdots, p_m\}$ sont des densités sur $\mathscr{Z}$ et pour tout $i \in [\![1, m]\!]$, $\theta^i > 0$, $\sum_{i=1}^m \theta^i < 1$. On peut alors écrire p_θ de la manière suivante

$$p_\theta(x) = C(x) + \sum_{i=1}^m \theta^i F_i(x).$$

Ainsi dans la carte $[\theta^i]$ où on note $\partial_i = \frac{\partial}{\partial \theta^i}$, on a

$$\begin{aligned} \partial_i l_\theta(x) &= \frac{F_i(x)}{p_\theta(x)}, \\ \partial_i \partial_j l_\theta(x) &= -\frac{F_i(x) F_j(x)}{p_\theta(x)^2}. \end{aligned}$$

D'où

$$\partial_i \partial_j l_\theta + \partial_i l_\theta \partial_j l_\theta = 0.$$

Ainsi $\Gamma^{(-1)}_{ij,k} = 0$. Ainsi $\mathscr{P}$ est (-1)-plat et $[\theta^i]$ est un (-1)-système de coordonnées affines. On appelle $\nabla^{(-1)}$ la **connexion mélange** notée $\nabla^{(m)}$. Ainsi $\mathscr{P}$ est dit m-plat.

Remarque 1.2.2. *Faisons une remarque essentielle, pour $\mathscr{X} = \{x_1, \cdots, x_n\}$, l'ensemble des mesures de probabilités sur $\mathscr{X}$ à poids strictement positifs noté $\mathscr{P}(\mathscr{X})$ est un modèle de mélange. En effet, soit*

$Q = \sum_{i=1}^{n} p_i \delta_{x_i} \in \mathscr{P}(\mathscr{X})$ et prenons μ la mesure de comptage sur $\mathscr{X}$. Alors il suffit de poser

$$\begin{aligned} C(x) &= 1_{x_n}(x) \\ m &= n-1 \\ F_i(x) &= 1_{\{x_i\}}(x) - 1_{\{x_n\}}(x), \ \ \forall\ i \in [\![1, n-1]\!] \\ \theta^i &= p_i, \ \ \forall\ i \in [\![1, n-1]\!]. \end{aligned}$$

1.2.3 Structure duale d'une variété

Afin de mieux comprendre la structure géométrique du modèle statistique, nous avons besoin de la notion de **connexions duales**. Soit (M, g) une variété riemannienne. Deux connexions ∇ et ∇^* sont dites **duales** si pour tout $X, Y, Z \in C^\infty(TM)$,

$$Z(g(X,Y)) = g(\nabla_Z X, Y) + g(X, \nabla^*_Z Y). \tag{1.2}$$

Remarquons la similitude entre la notion de connexion métrique et celle de connexions duales. Ainsi une connexion est dite métrique si elle est auto-duale c'est à dire $\nabla^* = \nabla$. On appelle (g, ∇, ∇^*) une **structure duale** sur M. En se plaçant dans une carte (système de coordonnées et repère local) et en notant $\Gamma_{ij,k}$ et $\Gamma^*_{ij,k}$ les symboles pour les connexions ∇ et ∇^* respectivement, on a

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}. \tag{1.3}$$

Etant données une métrique g et une connexion ∇, il existe une unique connexion duale ∇^* de ∇. De plus, on voit rapidement que $(\nabla^*)^* = \nabla$. On peut aussi remarquer que $\frac{\nabla+\nabla^*}{2}$ est une connexion métrique. Réciproquement, si ∇' a la même torsion que ∇^* et si $\frac{\nabla+\nabla^*}{2}$ est métrique alors $\nabla' = \nabla^*$. Pour revenir aux modèles statistiques, on a le résultat suivant :

Théorème 1.2.3. *Soit $\alpha \in \mathbb{R}$. Pour tout modèle statistique, la α-connexion et la $(-\alpha)$-connexion sont duales par rapport à la métrique de Fisher.*

Démonstration.

Fixons $\alpha \in \mathbb{R}$ et calculons

$$\begin{aligned} \Gamma^{\alpha}_{ki,j}(\theta) + \Gamma^{(-\alpha)}_{kj,i}(\theta) &= E_\theta\left((\partial_k \partial_i l_\theta + \frac{1-\alpha}{2}\partial_k l_\theta \partial_i l_\theta)\partial_j l_\theta\right) + \\ &\quad E_\theta\left((\partial_k \partial_j l_\theta + \frac{1+\alpha}{2}\partial_k l_\theta \partial_j l_\theta)\partial_i l_\theta\right) \\ &= E_\theta(\partial_k l_\theta \partial_i l_\theta \partial_j l_\theta) + E_\theta((\partial_k \partial_i l_\theta)\partial_j l_\theta) + E_\theta((\partial_k \partial_j l_\theta)\partial_i l_\theta) \\ &= \partial_k g_{ij}(\theta). \end{aligned}$$

□

Remarque 1.2.4. *En particulier, on a la dualité entre la e-connexion et la m-connexion.*

De plus la structure duale peut s'étendre à une sous variété de la manière suivante. Supposons dorénavant qu'on a une structure duale sur la variété riemannienne (M, g, ∇, ∇^*). Soit N une sous variété de M. On note ∇_N et ∇_N^* les projections orthogonales par rapport à g sur N de ∇ et ∇^* respectivement. Ainsi cela induit une structure duale sur la sous variété N qui est $(N, g_N, \nabla_n, \nabla_N^*)$.

Enfin mentionnons un résultat important pour la courbure. Il est possible de relier la courbure pour ∇ et la courbure pour ∇^* de la manière suivante. Pour tout $X, Y, Z, W \in C^\infty(TM)$, on a

$$g(R(X,Y)Z, W) = -g(R^*(X,Y)W, Z) \tag{1.4}$$

où R (resp. R^*) est la courbure pour la connexion ∇ (resp. ∇^*). D'où

$$R = 0 \Leftrightarrow R^* = 0. \tag{1.5}$$

On en déduit alors que la planéité de ∇ est équivalente à celle de ∇^*.

1.2.4 Construction d'une structure riemannienne duale à partir d'une divergence

Dans cette sous-section, nous montrons qu'il est possible de définir une structure riemannienne duale à partir d'une divergence D. Soit $\mathscr{P}$ une variété différentielle. Soit $D : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ une fonction lisse satisfaisant pour tout $p, q \in \mathscr{P}$,

$$D(p||q) \geq 0, \quad \text{et} \quad D(p||q) = 0 \Leftrightarrow p = q. \tag{1.6}$$

Soit $(p, q) \in \mathscr{P}^2$. Notons $[\xi^i]$ (resp $[\xi'^i]$) un système de coordonnées pour p (resp q). On définit pour $X_1, ..., X_r$ et $Y_1, ..., Y_l$ des champs de vecteurs

$$D[X_1 ... X_r || Y_1 ... Y_l](p) = (X_1)_p ... (X_r)_p (Y_1)_q ... (Y_l)_q D(p||q)_{|p=q}, \tag{1.7}$$

$$D[X_1 ... X_r || \cdot](p) = (X_1)_p ... (X_r)_p D(p||q)_{|p=q}. \tag{1.8}$$

Notation 1.2.1. Pour $i \in [\![1, r]\!]$, on note $(X_i)_p$ l'élément $X_i(p) \in T_p\mathscr{P}$ car $X_i : \mathscr{P} \to T\mathscr{P}$. On pourra consulter la définition de champ de vecteurs dans la section A.6.

Puisque $q \mapsto D(p||q)$ et $p \mapsto D(p||q)$ admettent un minimum global en $p = q$, en notant $\partial_i = \frac{\partial}{\partial \xi^i}$ (on peut aussi varier les systèmes de coordonnées et on obtient le même résultat) on a

$$D[\partial_i || \cdot] = D[\cdot || \partial_i] = 0. \tag{1.9}$$

De plus

$$D[\partial_i \partial_j || \cdot] = D[\cdot || \partial_i \partial_j] = -D[\partial_i || \partial_j]. \tag{1.10}$$

On définit $g^D_{ij} = -D[\partial_i||\partial_j]$. On obtient une matrice $G^{(D)} = (g^D_{ij})$ symétrique semi définie positive car l'application $\varphi : p \mapsto D(p||q)$ admet un minimum global en $p = q$. Lorsque la matrice est définie positive alors on dit que D est une **divergence** ou un **contraste**. On définit alors

$$g^{(D)}(X, Y) = -D[X||Y] \tag{1.11}$$

où X, Y sont des champs de vecteurs. De plus on a $D(p||q) = \varphi(p) = \varphi \circ \xi^{-1}(\xi(p)) = \bar{\varphi}(\xi(p))$ où $\bar{\varphi} = \varphi \circ \xi^{-1}$ et ξ est un système de coordonnées. Ainsi on peut faire un développement limité dans la carte ξ

$$\begin{aligned}
D(p||q) &= \bar{\varphi}(\xi(p)) \\
&= \bar{\varphi}(\xi(q) + (\xi(p) - \xi(q))) \\
&= \frac{1}{2}(\xi(p) - \xi(q))^T G^{(D)}(q)(\xi(p) - \xi(q)) + o(\|\xi(p) - \xi(q)\|^2) \\
&= \frac{1}{2} g^{(D)}_{ij}(q)\Delta\xi^i \Delta\xi^j + o(\|\Delta\xi\|^2)
\end{aligned}$$

où $\Delta\xi^i = \xi^i(p) - \xi^i(q)$. De la même manière, on peut définir la connexion $\nabla^{(D)}$ par les coefficients $\Gamma^{(D)}_{ij,k} = -D[\partial_i\partial_j||\partial_k]$ et plus généralement pour tout X, Y, Z (des champs de vecteurs)

$$g^{(D)}(\nabla^{(D)}_X Y, Z) = -D[XY||Z]. \tag{1.12}$$

On peut remarquer que la connexion $\nabla^{(D)}$ est sans torsion puisque $\Gamma_{ij,k} = \Gamma_{ji,k}$. Définissons maintenant la divergence duale D^* d'une divergence D par $D^*(p||q) = D(q||p)$. Puisque $D[\partial_i||\partial_j] = D[\partial_j||\partial_i]$, on a

$$g^{(D^*)} = g^{(D)}, \tag{1.13}$$

$$\Gamma^{(D^*)}_{ij,k} = -D^*[\partial_i\partial_j||\partial_k] = -D[\partial_k||\partial_i\partial_j]. \tag{1.14}$$

Par cela, on peut définir la connexion $\nabla^{(D^*)}$ et on peut montrer que les connexions $\nabla^{(D)}$ et $\nabla^{(D^*)}$ sont duales :

Théorème 1.2.5. *$\nabla^{(D)}$ et $\nabla^{(D^*)}$ sont duales par rapport à $g^{(D)}$.*

Ainsi une divergence engendre une structure riemannienne duale sans torsion $(M, g^{(D)}, \nabla^{(D)}, \nabla^{(D^*)})$. Et réciproquement à partir d'une structure duale sans torsion (M, g, ∇, ∇^*) (sans torsion pour chaque connexion) alors on peut définir globalement une divergence D (*K.Matsumoto* [28]).

1.2.5 Structure duale plate (DFS) et système dual de coordonnées

Définissons à présent la notion de **structure duale plate** (Dually flat spaces) qu'on notera sous l'acronyme DFS. Soit (g, ∇, ∇^*) une structure duale sur une variété différentielle S. Si ∇ et ∇^* sont

sans torsion, alors la planéité de ∇ équivaut à celle ∇^* (par la relation entre R et R^*). Ainsi pour les α-divergence avec $\alpha \in \mathbb{R}$, **on a l'équivalence S est α-plate si et seulement si S est $(-\alpha)$-plate** [43]. Ainsi le modèle exponentiel et la famille mélangée sont (± 1)-plats.

On dit que (S, g, ∇, ∇^*) est une **DFS** si S est à la fois ∇-plate et ∇^*-plate. La structure DFS se transmet aux sous variétés autoparallèles à S

Théorème 1.2.6. *Soient (S, g, ∇, ∇^*) une DFS et M une sous variété autoparallèle par rapport à ∇* ***ou*** *∇^*. Alors $(M, g_M, \nabla_M, \nabla^*_M)$ est une DFS.*

Ainsi par exemple, les sous variétés m-autoparallèles (resp. e-autoparallèles) d'une famille exponentielle (resp. d'une famille mélangée) sont ± 1-plates.

Nous souhaitons définir deux fonctions potentielles associées à une structure duale plate. Considérons (S, g, ∇, ∇^*) une DFS et $[\theta^i]$ un système de coordonnées affines par rapport à ∇. On note $\partial_i = \frac{\partial}{\partial \theta^i}$. On définit les champs de vecteurs ∂^j par

$$g(\partial_i, \partial^j) = \delta_{ij} = 1_{i=j}$$

pour tout $i = 1, .., m$. Au passage remarquons que cela détermine bien les champs de vecteurs ∂^j pour tout $j \in [\![1, m]\!]$. Fixons un $j \in [\![1, m]\!]$. En écrivant $\partial^j = \sum_k \alpha^j_k \partial_k$ et en notant $g = (g(\partial_i, \partial_k)_{i,k})$,$\alpha^{(j)} = (\alpha^j_k)_k$, e_j le j-ième vecteur de la base canonique de $\mathbb{R}^m$, on obtient

$$g\alpha^{(j)} = e_j.$$

Ainsi cela définit bien le champ de vecteur ∂^j. De plus pour tout champ de vecteurs X vu comme une dérivation on a par la structure duale

$$0 = Xg(\partial_i, \partial^j) = g(\nabla_X \partial_i, \partial^j) + g(\partial_i, \nabla^*_X \partial^j).$$

Puisque $[\theta^i]$ est un système de coordonnées affines alors les champs de vecteurs ∂_i sont parallèles. Par conséquent, on a

$$g(\partial_i, \nabla^*_X \partial^j) = 0.$$

Montrons que les ∂^j sont parallèles par rapport à ∇^* i.e pour tout k, j, on a $\nabla^*_{\partial^k} \partial^j = 0$. Ecrivons $\nabla^*_{\partial^k} \partial^j = \sum_l \Gamma^{*l}_{kj} \partial^l$. En utilisant la métrique g, on a pour tout i,

$$\begin{aligned} 0 = g(\partial_i, \nabla^*_{\partial^k} \partial^j) &= \sum_l \Gamma^{*l}_{kj} g(\partial_i, \partial^l) \\ &= \Gamma^{*i}_{jk} \\ &= 0. \end{aligned}$$

Ainsi ∂^j sont parallèles par rapport à ∇^*. On peut trouver un système de coordonnées affines $[\eta^j]$ tel que $\frac{\partial}{\partial \eta^j} = \partial^j$. En effet en écrivant $\partial^j = \sum_k \alpha_k \partial_k$, il suffit de poser $\eta^j = \sum_k \alpha_k \theta^k$. On appelle $[\theta^i]$ et $[\eta^j]$ un **système dual de coordonnées**. Notons maintenant

$$\begin{aligned} g_{ij} &= g(\partial_i, \partial_j), \\ g^{ij} &= g(\partial^i, \partial^j). \end{aligned}$$

En utilisant les formules de changement de variables $\partial^j = \partial^j \theta^i \partial_i$ et $\partial_i = \partial_i \eta^j \partial^j$, on a

$$\begin{aligned} g_{ij} &= g(\partial_i, \partial_j) \\ &= \partial_i \eta^k g(\partial^k, \partial_j) \\ &= \partial_i \eta^j \\ &= \frac{\partial \eta^j}{\partial \theta^i}. \end{aligned}$$

De même $g^{ij} = \frac{\partial \theta^i}{\partial \eta^j}$. De plus, on a $g_{ij} g^{jk} = \delta_{ik}$ en utilisant la convention d'Einstein. En effet

$$\begin{aligned} \delta_{ij} = g(\partial_i, \partial^j) &= (\partial_i \eta^k)(\partial^j \theta^l) g(\partial^k, \partial_l) \\ &= (\partial_i \eta^k)(\partial^j \theta^k) \\ &= g_{ik} g^{kj}. \end{aligned}$$

Intéressons nous maintenant à cette équation pour tout i,

$$\partial_i \psi = \eta^i.$$

Une solution à cette équation existe si et seulement si $\partial_j \eta^i = \partial_i \eta^j$. Dans notre cas, $\partial_j \eta^i = g_{ij} = \partial_i \eta^j$ donc une solution existe. On a alors

$$\partial_i \partial_j \psi = g_{ij}.$$

Donc ψ est strictement convexe. De même, on résout pour tout i,

$$\partial^i \varphi = \theta^i.$$

En utilisant une solution ψ de l'équation $\partial_i \psi = \eta^i$, on pose $\varphi = \theta^i \eta^i - \psi$ où $\psi : \theta \mapsto \psi(\theta)$ (avec la convention d'Enstein). On remarque que

$$\partial^i \varphi = \theta^i + \partial^i \theta^j \eta^i - \partial^i \eta^j \partial_j \psi = \theta^i,$$

De plus, on a aussi

$$\partial^i \partial^j \varphi = g^{ij}.$$

D'où φ est une fonction strictement convexe aussi. De plus, on a que :

$$\varphi(\eta) + \psi(\theta) = \theta^i \eta^i.$$

Remarquons qu'on écrira souvent $\psi(p) = \psi(\theta(p))$. Ainsi $\varphi(p) + \psi(p) = \theta^i(p)\eta^i(p)$. Elle vérifie l'inégalité de Young, ainsi

$$\varphi(q) = \max_{p\in S}\{\theta^i(p)\eta^i(q) - \psi(p)\}, \tag{1.15}$$

$$\psi(p) = \max_{q\in S}\{\theta^i(q)\eta^i(p) - \varphi(q)\}. \tag{1.16}$$

Qu'on peut réécrire différemment

$$\varphi(\eta) = \max_{\theta}\{\theta^i\eta^i - \psi(\theta)\},$$
$$\psi(\theta) = \max_{\eta}\{\theta^i\eta^i - \varphi(\eta)\}.$$

Ces deux fonctions φ et ψ sont appelés des **potentiels** et permettent de relier les deux systèmes de coordonnées. De plus puisque $g_{ij} = \partial_i\partial_j\psi$ et $\nabla_{\partial_i}\partial_j = 0$ (donc $\Gamma_{ij,k} = 0$), on a

$$\Gamma^*_{ij,k} = g(\nabla^*_{\partial_i}\partial_j, \partial_k) = \partial_i\partial_j\partial_k\psi.$$

De même pour les mêmes raisons (∂^i parallèles par rapport à ∇^*), on a que $\Gamma^{*ij,k} = 0$ et

$$\Gamma^{ij,k} = g(\nabla_{\partial^i}\partial^j, \partial^k) = \partial^i\partial^j\partial^k\varphi.$$

1.2.6 Divergence canonique construite à partir d'une DFS et relation triangulaire

Soit (S, g, ∇, ∇^*) une DFS de coordonnées $\{[\theta^i], [\eta^i]\}$ et de fonctions potentielles $\{\psi, \varphi\}$. Pour $p, q \in S$, on définit

$$D(p||q) = \psi(p) + \varphi(q) - \theta^i(p)\eta^i(q).$$

En utilisant (1.15) et (1.16), on a $D(p||q) \geq 0$ et $D(p||q) = 0 \Leftrightarrow p = q$. Ainsi D est une divergence. De plus,

$$\begin{aligned} D((\partial_i\partial_j)_p||q) &= (\partial_i\partial_j)_p D(p||q) \\ &= (\partial_i\partial_j)_p\psi(p) \\ &= g_{ij}(p). \end{aligned}$$

De même,

$$D(p||(\partial_i\partial_j)_q) = g^{ij}(q).$$

Ainsi la métrique induite par D est g. Puisque $\Gamma_{ij,k} = \Gamma^{*ij,k} = 0$, on a $\nabla^{(D)} = \nabla$ et $\nabla^{(D^*)} = \nabla^*$. On appelle D la **divergence canonique** de (S, g, ∇, ∇^*) ou la (g, ∇)-**divergence** sur S. De même en échangeant le rôle de ∇ et ∇^*, on peut trouver une (g, ∇^*)- divergence D^* qui vérifie $D^*(p||q) = D(q||p)$. De plus, pour M une sous variété autoparallèle par rapport à ∇ ou ∇^*, on peut considérer la DFS $(M, g_M, \nabla_M, \nabla^*_M)$ et il est possible de montrer que la (g_M, ∇_M)-divergence est $D_M = D_{|M\times M}$. Le même argument s'applique pour D^*_M.

Exemple 1.2.2. Soit ∇ une connexion métrique (donc autoduale $\nabla = \nabla^*$). Puisque ∇ est plate, alors $g(\partial_i, \partial_j)$ est constante pour un système de coordonnées affine donné. De plus, l'ensemble des systèmes de coordonnées affines sont reliés par une transformation affine. Donc il existe un système de coordonnées euclidien (noté $[\theta^i]$) c'est à dire qui vérifie que $g(\partial_i, \partial_j) = \delta_{ij}$ (où ∂_i est un autre système de coordonnées que le précédent). Puisque la connexion est métrique, alors $\theta = \eta$. Etant donné que $g_{ij} = g(\partial_i, \partial_j) = \partial_i \partial_j \psi$ et $\partial_i \psi = \theta^i$, on a que $\psi = \varphi = \frac{1}{2}\sum_i (\theta^i)^2$. Ainsi pour $p, q \in U$ ((U, θ) étant la carte),

$$\begin{aligned} D(p||q) &= \psi(p) + \varphi(q) - \theta^i(p)\theta^i(q) \\ &= \frac{1}{2}\sum_i \Big((\theta^i(p))^2 + (\theta^i(q))^2 - 2\theta^i(p)\theta^i(q) \Big) \\ &= \frac{1}{2}(d(p,q))^2 \end{aligned}$$

où $d(p,q) = \sqrt{\sum_i (\theta^i(p) - \theta^i(q))^2}$. Ainsi dans le cas métrique (de la connexion), la divergence canonique est égale à facteur $1/2$ de la distance euclidienne au carré.

Donnons maintenant une condition nécessaire et suffisante pour qu'une divergence D sur une DFS soit une divergence canonique.

Théorème 1.2.7. *Soit $\{[\theta^i, \eta^i\}$ un système dual de coordonnées affines d'une DFS (S, g, ∇, ∇^*). Soit D une divergence sur S. Alors une condition nécessaire et suffisante pour que D soit une (g, ∇)-divergence est que pour tout $p, q, r \in S$, la relation triangulaire soit vérifiée*

$$D(p||q) + D(q||r) - D(p||r) = (\theta^i(p) - \theta^i(q))(\eta^i(r) - \eta^i(q)). \tag{1.17}$$

De même pour que D^* soit une (g, ∇^*)-divergence est que pour tout $p, q, r \in S$, la relation triangulaire soit vérifiée

$$D^*(p||q) + D^*(q||r) - D^*(p||r) = (\theta^i(r) - \theta^i(q))(\eta^i(p) - \eta^i(q)). \tag{1.18}$$

1.2.7 Théorème de Pythagore généralisé et théorème de projection

Nous pouvons dorénavant énoncer le **théorème de Pythagore** pour une DFS (S, g, ∇, ∇^*).

Théorème 1.2.8. *Soient $p, q, r \in S$, γ_1 une ∇-géodésique reliant p et q et γ_2 une ∇^*-géodésique reliant q et r. Si à l'intersection q, les courbes γ_1 et γ_2 sont orthogonales (pour la métrique g), alors nous*

obtenons la ***relation de Pythagore***

$$D(p||r) = D(p||q) + D(q||r) \tag{1.19}$$

On a aussi la relation de Pythagore pour la divergence duale D^* si γ_1 une ∇^*-géodésique reliant p et q et γ_2 une ∇-géodésique reliant q et r et à l'intersection q, les courbes γ_1 et γ_2 sont orthogonales (pour la métrique g) alors on a

$$D^*(p||r) = D^*(p||q) + D^*(q||r).$$

Un théorème de projection découle directement comme corollaire du théorème précédent.

Théorème 1.2.9. *Soient $p \in S$ et M une sous variété de S qui est ∇^*-autoparallèle. Alors une condition nécessaire et suffisante pour qu'un point $q \in M$ vérifie :*

$$D(p||q) = \min_{r \in M} D(p||r). \tag{1.20}$$

*est que la ∇-géodésique reliant p et q soit orthogonale à M en q. Le point q est appelé la ∇-**projection de p sur** M.*

De même si M est ∇-autoparallèle. Alors une condition nécessaire et suffisante pour qu'un point $q \in M$ vérifie

$$D^*(p||q) = \min_{r \in M} D^*(p||r) \tag{1.21}$$

*est que la ∇^*-géodésique reliant p et q soit orthogonale à M en q. Le point q est appelé la ∇^*-**projection de p sur** M.*

Il est possible de relâcher l'hypothèse de sous variété autoparallèle :

Théorème 1.2.10. *Soient $p \in S$ et M une sous variété de S. Une condition nécessaire et suffisante pour qu'un point q soit un point stationnaire de la fonction $D(p||\cdot) : r \mapsto D(p||r)$ restreinte à M (i.e les dérivées partielles par rapport à un système de coordonnées de la fonction sont toutes nulles) est que la ∇-géodésique reliant p et q soit orthogonale à M en q.*

De même une condition nécessaire et suffisante pour qu'un point q soit un point stationnaire de la fonction $D^*(p||\cdot) : r \mapsto D^*(p||r)$ restreinte à M (i.e les dérivées partielles par rapport à un système de coordonnées de la fonction sont toutes nulles) est que la ∇^*-géodésique reliant p et q soit orthogonale à M en q.

Enonçons maintenant un corollaire intéressant à ce théorème :

Corollaire 1.2.11. *Soient $p \in S$ et $c \in \mathbb{R}_+^*$. On suppose que la D-sphère $M = \{q \in S, D(p||q) = c\}$ soit une hypersurface de S. Alors toute ∇-géodésique passant par le centre p intersecte orthogonalement la D-sphère M.*

Démonstration.
On remarque que la fonction $q \mapsto D(p||q)$ est constante sur la D-sphère M. Ainsi tout point $q \in M$ est un point stationnaire de la fonction $q \mapsto D(p||q)$ restreinte à M. Par le théorème 1.2.10, la ∇-géodésique reliant p et q est orthogonale à la D-sphère M. Puisque q est quelconque, on en déduit que toute géodésique passant par le centre p intersecte orthogonalement la D-sphère M. □

1.2.8 α-variété et α-famille

Dans cette sous-section, nous souhaitons définir la notion de α-**variété** et de α-**famille**. On considère S un modèle statistique régulier dominé par une mesure μ (sur $\mathscr{X}$). On note $\mathscr{P}(\mathscr{X})$ l'ensemble des mesures de probabilité sur $\mathscr{X}$ dominées par μ. On note aussi

$$\tilde{\mathscr{P}}(\mathscr{X}) = \left\{ p : \mathscr{X} \to \mathbb{R},\ p > 0\ (\mu - pp),\ \int_{\mathscr{X}} p d\mu < \infty \right\}.$$

C'est une extension de $\mathscr{P}(\mathscr{X})$ où on relâche l'hypothèse de normalisation. De par la régularité du modèle, $S = \{P_\theta\}_\theta$ (dans une carte) est une variété riemannienne de métrique de fisher g et on note $\nabla^{(\alpha)}$ la α-connexion. Posons pour $\alpha \in \mathbb{R}^{\setminus\{1\}}$, $L^{(\alpha)}(u) = \frac{2}{1-\alpha} u^{\frac{1-\alpha}{2}}$ et $L^{(1)}(u) = \log u$ pour tout $u \in \mathbb{R}_+^*$. On définit

$$l_\theta^{(\alpha)}(x) = L^{(\alpha)}(p_\theta(x)).$$

On notera $l_\theta^\alpha := l_\theta^{(\alpha)}$ et $l := l^{(1)}$. On définit la α-représentation de $X \in T_\theta S$ par

$$X^\alpha(x) = X l_\theta^\alpha(x).$$

En calculant, on a

$$\begin{aligned} \partial_i l^\alpha &= p^{\frac{1-\alpha}{2}} \partial_i l, \\ \partial_i \partial_j l^\alpha &= p^{\frac{1-\alpha}{2}} \left(\partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l \right). \end{aligned}$$

En utilisant ces égalités, on obtient

$$\begin{aligned} g_{ij}(\theta) &= \int \partial_i l_\theta^\alpha \partial_j l_\theta^{-\alpha} d\mu, \\ \Gamma_{ij,k}^\alpha(\theta) &= \int \partial_i \partial_j l_\theta^\alpha \partial_k l_\theta^{-\alpha} d\mu. \end{aligned}$$

On dit que S est une **α-variété** s'il existe un système de coordonnées $[\theta^i]$ tel que

$$\partial_i \partial_j l_\theta^\alpha = 0.$$

Dans ce cas $[\theta^i]$ est un α-système de coordonnées affines et donc S est α-plat. Une condition équivalente à la condition précédente est qu'il existe $n+1$ fonctions $\{C, F_1, ..., F_n\}$ où $n = \dim(S)$ telles que

$$l_\theta^\alpha(x) = C(x) + \sum_{i=1}^{n} \theta^i F_i(x). \tag{1.22}$$

Ainsi il est clair qu'une famille de mélange est une (-1)-variété. Cependant, une famille exponentielle n'est pas une 1-variété. En effet $l^1 = l$ et par 1.2 $\partial_i \partial_j l_\theta = -\partial_i \partial_j \psi(\theta)$. Donc une famille exponentielle n'est pas en général une 1-variété.

Pour $\mathscr{X} = \{x_1, \cdots, x_r\}$ fini, on peut montrer que $\tilde{\mathscr{P}}(\mathscr{X})$ est une α-variété pour tout $\alpha \in \mathbb{R}$. En effet, en posant $\theta^i = L^\alpha(p(x_i))$ et en définissant $F_i : \mathscr{X} \to \mathbb{R}$ par $F_i(x_j) = \delta_{ij}$ pour tout $i, j \in [\![1, r]\!]$, on obtient le résultat voulu.

Maintenant, nous allons **dénormaliser** la variété S de telle manière à la voir comme une sous variété et la version dénormalisée sera une α-variété. Définissons la version dénormalisée

$$\tilde{S} = \{\tau p_\theta, \theta \in \Theta, \tau > 0\}.$$

Dans ce cas, on voit S comme une sous variété de $\tilde{S}$ avec $\dim(\tilde{S}) = \dim S + 1$. Ainsi, on a un nouveau système de coordonnées $[\theta, \tau] = [\theta^1, ..., \theta^n, \tau]$ et la base naturelle devient $\tilde{\partial}_i = \partial_i$ et $\tilde{\partial}_\tau = \frac{\partial}{\partial \tau}$. Par la dénormalisation de la variété différentielle S, il est possible de montrer que :

Théorème 1.2.12. *S est (-1)-autoparallèle dans $\tilde{S}$.*

De même, on a le résultat suivant :

Théorème 1.2.13. *Soit M une sous variété de S et notons $\tilde{M}$ sa dénormalisation. Alors pour tout $\alpha \in \mathbb{R}$, on a équivalence entre :*

1. *M est α-autoparallèle dans S.*
2. *$\tilde{M}$ est α-autoparallèle dans $\tilde{S}$.*

On dit qu'un modèle statistique régulier S est une **α-famille** si $\tilde{S}$ est une α-variété.

Ainsi pour $\mathscr{X}$ fini, $\mathscr{P}(\mathscr{X})$ est une α-famille pour tout $\alpha \in \mathbb{R}$ car $\tilde{\mathscr{P}}(\mathscr{X})$ est une α-variété. On a vu que la famille exponentielle n'est pas forcément une 1-variété. Intéressons nous à sa version dénormalisée. Pour $\alpha = 1$, on a $L^1 = \log$ et notons pour tout $x \in \mathbb{R}^p$ pour un certain $p \in \mathbb{N}^*$

$$p_\theta(x) = \exp\left(C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)\right).$$

Ainsi

$$\log(\tau p_\theta) = \log\tau + \log p_\theta = C(x) + \log\tau + \sum_{i=1}^{\dim S} \theta^i F_i(x) - \psi(\theta).$$

En posant $F_0 = 1$ et $\theta^0 = \log\tau - \psi(\theta)$, on a

$$\log(\tau p_\theta) = C(x) + \sum_{i=0}^{\dim S} \theta^i F_i.$$

En conclusion, le modèle exponentiel est une 1-famille car $\tilde{S}$ est une 1-variété et une famille de mélange est une (-1)-variété.

1.2.9 Structure duale du modèle exponentiel et de mélange et divergence canonique associée à chaque modèle

Etudions maintenant la structure duale du modèle exponentiel. Tout d'abord par ce qui précède, cette structure est une DFS. Trouvons le système dual de coordonnées. On note les densités de ce modèle exponentiel S,

$$p_\theta(x) = \exp\Big(C(x) + \theta^i F_i(x) - \psi(\theta)\Big).$$

On note $[\theta^i]$ les paramètres naturels du modèle. Nous avons vu que ce système de coordonnées forme un 1-système de coordonnées affines. Définissons

$$\eta^i(\theta) = E_\theta(F_i) = \int F_i p_\theta \, d\mu. \tag{1.23}$$

Un simple calcul nous montre que $\eta^i = \partial_i \psi$. De plus, on remarque aussi que $\partial_i \partial_j \psi = g_{ij}$. Ces équations caractérisent le système de coordonnées dual à $[\theta^i]$. Puisque $[\theta^i]$ est un 1-système de coordonnées affines alors $[\eta^i]$ est (-1)-système de coordonnées affines. Au passage précisons que ψ est la fonction potentielle pour le système de coordonnnées $[\theta^i]$. Cherchons la fonction potentielle φ du système de coordonnées $[\eta^i]$. On avait défini φ de la manière suivante

$$\begin{aligned}\varphi(\theta) := \varphi(\eta(\theta)) &= \theta^i \eta^i(\theta) - \psi(\theta)\\ &= E_\theta(\log p_\theta - C)\\ &= -H(p_\theta) - E_\theta C.\end{aligned}$$

où H est l'entropie définie par

$$H(p) = -\int p \log p d\mu.$$

Calculons la divergence canonique associée à cette paire de fonctions potentielles. En notant $p = p_\theta$ et $q = p_{\theta'}$, on a

$$\begin{aligned}D(P_\theta || P_{\theta'}) &= \psi(p) + \varphi(q) - \theta^i(p)\eta^i(q)\\ &= \varphi(q) - (\theta^i(p)\eta^i(q) - \psi(p))\\ &= E_{\theta'}(\log p_{\theta'} - C) - \big(E_{\theta'}(\log p_\theta - C)\big)\\ &= E_{\theta'}(\log p_{\theta'} - \log p_\theta)\\ &= KL(P_{\theta'} || P_\theta)\\ &= KL^*(P_\theta || P_{\theta'}).\end{aligned}$$

Ainsi la divergence canonique (la $(g, \nabla^{(1)})$-divergence) est la divergence duale de Kullback-Leibler. Par conséquent, la $(g, \nabla^{(-1)})$-divergence est la divergence duale de la divergence canonique, autrement dit c'est la divergence de Kullback-Leibler.

Intéressons nous maintenant à la structure duale du modèle de mélange dont les densités sont données par

$$\begin{aligned} p_\eta(x) &= \sum_{i=1}^m \eta^i q_i(x) + \left(1 - \sum_{i=1}^m \eta^i\right) q_0(x) \\ &= \sum_{i=1}^m \eta^i (q_i(x) - q_0(x)) + q_0(x) \\ &= \eta^i (q_i(x) - q_0(x)) + q_0(x) \end{aligned}$$

en utilisant la convention d'Einstein. On rappelle que les q_i pour $i \in [\![1, m]\!]$ sont des densités de probabilité définies sur $\mathscr{X}$ par rapport à μ. On rappelle aussi que le système de coordonnées $[\eta^i]$ est un (-1)-système de coordonnées affines. On pose $\partial^i = \frac{\partial}{\partial \eta^i}$ et

$$\varphi(\eta) = \int_{\mathscr{X}} p_\eta \log p_\eta \, d\mu$$

où p_η est une densité de probabilité. Puisque le modèle est régulier, on peut montrer que

$$\partial^i \partial^j \varphi = g^{ij} = g(\partial^i, \partial^j)$$

où g est la métrique de Fisher. En effet

$$\begin{aligned} g^{ij}(\eta) &= \mathbb{E}_\eta \left(\partial^i \log p_\eta \partial^j \log p_\eta\right) \\ &= \int_{\mathscr{X}} \frac{(q_j - q_0)(q_i - q_0)}{p_\eta} d\mu. \end{aligned}$$

et

$$\begin{aligned} \partial^j \varphi(\eta) &= \int_{\mathscr{X}} (q_j - q_0) \log p_\eta \, d\mu + \int_{\mathscr{X}} p_\eta \frac{\partial^j p_\eta}{p_\eta} d\mu \\ &= \int_{\mathscr{X}} (q_j - q_0) \log p_\eta \, d\mu + \int_{\mathscr{X}} (q_j - q_0) d\mu. \end{aligned}$$

Puisque q_j et q_0 sont des densités de probabilité, on a

$$\partial^j \varphi(\eta) = \int_{\mathscr{X}} (q_j - q_0) \log p_\eta \, d\mu.$$

Ainsi

$$\begin{aligned} \partial^i \partial^j \varphi(\eta) &= \int_{\mathscr{X}} (q_j - q_0) \frac{\partial^i p_\eta}{p_\eta} d\mu \\ &= \int_{\mathscr{X}} (q_j - q_0) \frac{(q_i - q_0)}{p_\eta} d\mu \\ &= \mathbb{E}_\eta \left(\partial^i \log p_\eta \partial^j \log p_\eta\right) \\ &= g^{ij}(\eta). \end{aligned}$$

On définit le $[\theta^i]$ système de coordonnées de la manière suivante

$$\begin{aligned}\theta^i &= \partial^i \varphi \\ &= \int_{\mathscr{X}} (q_i - q_0) \log p_\eta d\mu.\end{aligned}$$

Ces équations caractérisent la structure duale, ainsi $[\theta^i]$ est donc un 1-système de coordonnées affines. Déterminons la fonction potentielle ψ. Puisque $\psi(\eta) + \varphi(\eta) = \theta^i(\eta)\eta^i$, on a

$$\begin{aligned}\psi(\eta) &= \theta^i(\eta)\eta^i - \varphi(\eta) \\ &= \int_{\mathscr{X}} \eta^i (q_i - q_0) \log p_\eta d\mu - \int_{\mathscr{X}} p_\eta \log p_\eta d\mu.\end{aligned}$$

Or $p_\eta = \eta^i(q_i - q_0) + q_0$, on en déduit

$$\psi(\eta) = -\int_{\mathscr{X}} q_0 \log p_\eta d\mu.$$

Nous pouvons maintenant déterminer la divergence canonique associée au modèle de mélange. Rappelons que $\eta^i(\tilde{\eta}) := \eta^i(p_{\tilde{\eta}}) = \tilde{\eta}^i$ et $\theta^i(\eta) := \theta^i(p_\eta)$. La divergence canonique est alors

$$\begin{aligned}D(p_\eta || p_{\eta'}) &= \psi(\eta) + \varphi(\eta') - (\eta')^i \theta^i(\eta) \\ &= \int_{\mathscr{X}} p_{\eta'} \log p_{\eta'} d\mu - \int_{\mathscr{X}} q_0 \log p_\eta d\mu - \int_{\mathscr{X}} (\eta')^i (q_i - q_0) \log p_\eta d\mu \\ &= \int_{\mathscr{X}} p_{\eta'} \log p_{\eta'} d\mu - \int_{\mathscr{X}} \Big((\eta')^i (q_i - q_0) + q_0\Big) \log p_\eta d\mu \\ &= \int_{\mathscr{X}} p_{\eta'} \log p_{\eta'} d\mu - \int_{\mathscr{X}} p_{\eta'} \log p_\eta d\mu \\ &= \int_{\mathscr{X}} p_{\eta'} \log\left(\frac{p_{\eta'}}{p_\eta}\right) d\mu \\ &= KL(p_{\eta'} || p_\eta).\end{aligned}$$

Ainsi on retrouve la **même divergence canonique** que pour le modèle exponentiel.

1.3 Injection optimale d'une information auxiliaire

1.3.1 Projection de la mesure empirique sur $\mathscr{P}^I$

On rappelle les notations énoncées dans la section 1.1. Fixons $n \in \mathbb{N}^*$. Soit $\mathscr{X} = \{X_1, \cdots, X_n\}$ l'ensemble des observations et notons $\mathscr{P}(\mathscr{X})$ l'ensemble des mesures de probabilités à poids strictement positifs sur $\mathscr{X}$ c'est à dire

$$\mathscr{P}(\mathscr{X}) = \left\{ \sum_{i=1}^n q_i \delta_{X_i},\ \ \forall i \in [\![1, n]\!]\ q_i > 0\ ,\ \ \sum_{i=1}^n q_i = 1 \right\}.$$

Notons I l'information auxiliaire qu'on dispose et on note $Q \sim I$ si Q vérifie l'information auxiliaire. Par exemple, si l'information auxiliaire I est donnée par Pg alors

$$Q \sim I \Longleftrightarrow Qg = Pg.$$

On note

$$\mathscr{P}^I = \left\{Q \in \mathscr{P}(\mathscr{X}),\ \ Q \sim I\right\}$$

l'ensemble des mesures de probabilité sur $\mathscr{X}$ vérifiant l'information auxiliaire I. Enfin notons

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

la mesure empirique des observations.

Par ce qui précède dans la section 1.2, $\mathscr{P}(\mathscr{X})$ est un modèle de mélange et admet une structure duale plate $(\mathscr{P}(\mathscr{X}), g, \nabla^{(1)}, \nabla^{(-1)})$. La divergence canonique associée à $\mathscr{P}(\mathscr{X})$ est la **divergence duale de Kullback-Leibler**.

Par application du théorème de projection 1.2.9, il est possible d'injecter de l'information auxiliaire en prenant le projeté $\mathbb{Q}_n^I$ de la mesure empirique :

Théorème 1.3.1. *On suppose que $\mathscr{P}^I$ est une **sous variété** de $\mathscr{P}(\mathscr{X})$. Alors*

- *Si $\mathscr{P}^I$ est une sous variété $\nabla^{(-1)}$-autoparallèle alors $\mathbb{Q}_n^I$ est la $\nabla^{(1)}$-projection de $\mathbb{P}_n$ sur $\mathscr{P}^I$ c'est à dire*

$$\mathbb{Q}_n^I \in arg \min_{Q\in\mathscr{P}^I} KL^*(\mathbb{P}_n||Q) = arg \min_{Q\in\mathscr{P}^I} KL(Q||\mathbb{P}_n).$$

- *Si $\mathscr{P}^I$ est une sous variété $\nabla^{(1)}$-autoparallèle alors $\mathbb{Q}_n^I$ est la $\nabla^{(-1)}$-projection de $\mathbb{P}_n$ sur $\mathscr{P}^I$*

$$\mathbb{Q}_n^I \in arg \min_{Q\in\mathscr{P}^I} KL(\mathbb{P}_n||Q).$$

- *Dans le cas où $\mathscr{P}^I$ n'est autoparallèle à aucune des ces deux connexions alors $\mathbb{Q}_n^I$ est un point stationnaire d'une de ces deux applications ci-dessous*

$$Q \longmapsto KL(\mathbb{P}_n||Q),$$
$$Q \longmapsto KL^*(\mathbb{P}_n||Q).$$

Remarque 1.3.2. *Notons que ce théorème ne permet pas de définir la mesure empirique informée puisque nous ne savons pas si la sous variété $\mathscr{P}^I$ est autoparallèle.*

1.3.2 Structure géométrique de $\mathscr{P}^I$ dans le cas d'une information auxiliaire forte apportée par des espérances

Afin de pouvoir appliquer le théorème précédent, on souhaite dorénavant montrer que $\mathscr{P}^I$ est une sous variété dans le cas où l'information auxiliaire forte est donnée par Pg où g est une fonction intégrable par rapport à P.

Proposition 1.3.3. *Soit $\mathscr{X} = \{X_1, \cdots, X_n\}$ l'ensemble des observations. On suppose que :*
- *Il existe $i \neq j$ tel que $g(X_i) \neq g\left(X_j\right)$.*
- *Pg appartient à l'enveloppe convexe de $\left\{g(X_1), \cdots, g(X_n)\right\}$.*

*Alors $\mathscr{P}^I$ est une **sous variété** de $\mathscr{P}(\mathscr{X})$ de dimension $n-2$.*

Démonstration.
Tout d'abord rappelons que $]0,1[^n$ est une sous variété de $\mathbb{R}^n$ puisque c'est un ouvert de $\mathbb{R}^n$. Posons

$$\mathscr{S} = \left\{ q \in]0,1[^n,\ \sum_{i=1}^{n} q_i = 1 \right\}.$$

le simplexe. Montrons dans un premier temps que $\mathscr{S}$ est une sous variété de dimension $n-1$ de $]0,1[^n$. Définissons la fonction $\theta :]0,1[^n \to \mathbb{R}$

$$\theta(q) = \sum_{i=1}^{n} q_i.$$

Remarquons que $\mathscr{S} = \theta^{-1}(\{1\})$. Il nous suffit donc de montrer que θ est une submersion. En effet, θ est différentiable et pour tout $q \in]0,1[^n$

$$\forall h \in \mathbb{R}^n,\ D\theta(q)(h) = \theta(h).$$

Ainsi $D\theta(q)$ est surjective. D'où θ est une submersion et on en déduit que $\mathscr{S}$ est une sous variété de dimension $n-1$ dans $]0,1[^n$. On munit $\mathscr{S}$ de la carte globale $(\mathscr{S}, \pi)$ avec

$$\begin{aligned} \pi : \mathscr{S} &\to U \subset \mathbb{R}^{n-1} \\ q &\mapsto (q_1, \cdots, q_{n-1}) \end{aligned}$$

et $U = \left\{(q_1, \cdots, q_{n-1}) \in \mathbb{R}^{n-1},\ q_i > 0\ \forall i \in [\![1, n-1]\!],\ \sum_{i=1}^{n-1} q_i < 1\right\}$. De même on munit $\mathscr{P}(\mathscr{X})$ de la carte globale $\left(\mathscr{P}(\mathscr{X}), \varphi\right)$ avec

$$\begin{aligned} \varphi : \mathscr{P}(\mathscr{X}) &\to U \subset \mathbb{R}^{n-1} \\ Q &\mapsto (q_1, \cdots, q_{n-1}). \end{aligned}$$

On s'intéresse dorénavant à la fonction bijective

$$\begin{aligned} \psi : \mathscr{P}(\mathscr{X}) &\to \mathscr{S} \\ Q &\mapsto q. \end{aligned}$$

Remarquons que $\psi = \pi^{-1} \circ \varphi$. *Puisque* π *et* φ *sont des difféomorphismes et* $dim\mathscr{P}(\mathscr{X}) = dim\mathscr{S} = n-1$, *on en déduit que* ψ *est un difféomorphisme. Ainsi* ψ^{-1} *est aussi un difféomorphisme et donc un plongement. Notons que*

$$\mathscr{P}^I = \psi^{-1}(\mathscr{E})$$

où $\mathscr{E} = \left\{q \in \mathscr{S},\ \sum_{i=1}^n q_i g(X_i) = Pg\right\}$. *Remarquons que cet ensemble est non vide puisque* Pg *appartient à l'ensemble convexe de* $\left\{g(X_1), \cdots, g(X_n)\right\}$. *Ainsi il nous suffit de montrer que* $\mathscr{E}$ *est une sous variété de* $\mathscr{S}$ *de dimension* $n-2$.

Définissons la fonction

$$\begin{aligned} f : \mathscr{S} &\to \mathbb{R} \\ q &\mapsto \sum_{i=1}^n q_i g(X_i). \end{aligned}$$

Montrons que f *est une submersion. Posons* $\gamma := f \circ \pi^{-1} : U \to \mathbb{R}$. *Ainsi*

$$\gamma(q) = \sum_{i=1}^{n-1} q_i g(X_i) + \left(1 - \sum_{i=1}^{n-1} q_i\right) g(X_n).$$

Ainsi γ *est différentiable et pour tout* $q \in U$

$$D\gamma(q) = \left(g(X_1) - g(X_n), \cdots, g(X_{n-1}) - g(X_n)\right).$$

Puisqu'il existe $i \neq j$ *tel que* $X_i \neq X_j$, *on en déduit que* $D\gamma(q)$ *est surjective. Ainsi* f *est une submersion et* $\mathscr{E}$ *est une sous variété de dimension* $n-2$ *par le lemme A.2.3. On en déduit que* $\mathscr{P}^I = \psi^{-1}(\mathscr{E})$ *est une sous variété de dimension* $n-2$ *car* ψ^{-1} *est un plongement en utilisant le lemme A.2.2.* □

Ce résultat peut se généraliser pour $g = (g_1, \cdots, g_m)^T$ avec $m \in [\![1, n-1]\!]$. Par simplicité d'écriture, **on supposera que** $Pg = 0$ **quitte à poser** $h = g - Pg$**.**

Posons

$$\begin{aligned} &\forall j \in [\![1, m]\!],\ N_j(X) = (g_j(X_1) - g_j(X_n), \cdots, g_j(X_{n-1}) - g_j(X_n))^T, \\ &\forall j \in [\![1, m]\!],\ g_j(X) = (g_j(X_1), \cdots, g_j(X_n))^T. \end{aligned}$$

Remarquons que

$$\dim\left(Vect(g(X_1) - g(X_n), \cdots, g(X_{n-1}) - g(X_n))\right) = \dim\left(Vect((N_j(X))_{1 \le j \le m})\right). \tag{1.24}$$

En effet définissons la matrice A et donnons sa transposée

$$A = \begin{pmatrix} g_1(X_1) - g_1(X_n) & \cdots & g_1(X_{n-1}) - g_1(X_n) \\ \vdots & \cdots & \vdots \\ g_m(X_1) - g_m(X_n) & \cdots & g_m(X_{n-1}) - g_k(X_n) \end{pmatrix},$$

$$A^T = \begin{pmatrix} g_1(X_1) - g_1(X_n) & \cdots & g_m(X_1) - g_m(X_n) \\ \vdots & \cdots & \vdots \\ g_1(X_{n-1}) - g_1(X_n) & \cdots & g_m(X_{n-1}) - g_m(X_n) \end{pmatrix}.$$

Puisque $rg(A) = rg(A^T)$, on obtient le résultat voulu. Nous pouvons à présent énoncer le théorème suivant.

Proposition 1.3.4. *Soient $\mathscr{X} = \{X_1, ..., X_n\}$ l'ensemble des observations et $g = \left(g_1, \cdots, g_m\right)$ avec $m \in [\![1, n-1]\!]$. On suppose que*

- *0 appartient à l'enveloppe convexe de $\left\{g(X_1), ..., g(X_n)\right\}$.*
- *L'égalité des dimensions est vérifiée,*

$$l := \dim\left(Vect(g_1(X), \cdots, g_m(X))\right) = \dim\left(Vect(g(X_1) - g(X_n), \cdots, g(X_{n-1}) - g(X_n))\right) \leq m.$$

*Alors $\mathscr{P}^I$ est une **sous variété** de $\mathscr{P}(\mathscr{X})$ de dimension $n-1-l$.*

Démonstration (de la proposition).
On reprend la démonstration du théorème précédent avec les mêmes notations. Il nous reste à montrer que $\mathscr{E} = \left\{q \in \mathscr{S},\ \sum_{i=1}^n q_i g(X_i) = 0\right\}$ est une sous variété de dimension $n-1-l$. Pour cela, nous avons besoin d'un lemme technique d'algèbre linéaire.

Lemme 1.3.5. *Soient $x^1, ..., x^k \in \mathbb{R}^n$ avec $n \in \mathbb{N}^*$ et $k \in [\![1, n-1]\!]$. Définissons*

$$\forall j \in [\![1, k]\!],\ \tilde{x}^j = \left(x_1^j - x_n^j, \cdots, x_{n-1}^j - x_n^j\right)^T \in \mathbb{R}^{n-1}.$$

Alors

$$\dim\left(Vect(\tilde{x}^1, \cdots, \tilde{x}^k)\right) \leq \dim\left(Vect(x^1, \cdots, x^k)\right).$$

De plus si

$$l := \dim\left(Vect(\tilde{x}^1, \cdots, \tilde{x}^k)\right) = \dim\left(Vect(x^1, \cdots, x^k)\right)$$

alors il existe une partie $J \subset [\![1, k]\!]$ de cardinal l telle que

$$\dim\left(Vect\left((\tilde{x}^j)_{j\in J}\right)\right) = \dim\left(Vect\left((x^j)_{j\in J}\right)\right).$$

Démonstration (du lemme).
Si $k = \dim\left(Vect(x^1, \cdots, x^k)\right)$ alors il n'y a rien à démontrer. On suppose que $\dim\left(Vect(x^1, \cdots, x^k)\right) < k$. Ainsi il existe $j \in [\![1, k]\!]$ tel que

$$x^j = \sum_{i \neq j} \lambda_i x^i$$

où $\lambda_1, \cdots, \lambda_k \in \mathbb{R}$. *Ainsi pour tout* $r \in [\![1, n]\!]$

$$x_r^j = \sum_{i \neq j} \lambda_i x_r^i.$$

D'où

$$x_r^j - x_n^j = \sum_{i \neq j} \lambda_i x_r^i - \sum_{i \neq j} \lambda_i x_n^i = \sum_{i \neq j} \lambda_i (x_r^i - x_n^i).$$

Autrement dit

$$\tilde{x}^j = \sum_{i \neq j} \lambda_i \tilde{x}^i.$$

On peut donc conclure que

$$\dim\Big(Vect(\tilde{x}^1, \cdots, \tilde{x}^k)\Big) \leq \dim\Big(Vect(x^1, \cdots, x^k)\Big).$$

On suppose maintenant que

$$l := \dim\Big(Vect(\tilde{x}^1, \cdots, \tilde{x}^k)\Big) = \dim\Big(Vect(x^1, \cdots, x^k)\Big)$$

Ainsi il existe une partie $J \subset [\![1, k]\!]$ *de taille* l *telle que*

$$l = \dim\Big(Vect\Big((\tilde{x}^j)_{j \in J}\Big)\Big).$$

Supposons par l'absurde que $\dim\big(Vect\big((x^j)_{j \in J}\big)\big) < l$. *Ainsi il existe* $r \in J$ *tel que*

$$x^r = \sum_{i \in J,\ i \neq r} \lambda_i x^i$$

où $\lambda_1, \cdots, \lambda_k \in \mathbb{R}$. *Par ce qui précède, on en déduit*

$$\tilde{x}^r = \sum_{i \in J,\ i \neq r} \lambda_i \tilde{x}^i.$$

Cela contredit le fait que

$$l = \dim\Big(Vect\Big((\tilde{x}^j)_{j \in J}\Big)\Big).$$

□

On peut donc appliquer le lemme technique à $g_1(X), \cdots, g_m(X)$. *Rappelons que (voir (1.24))*

$$\dim\big(Vect(g(X_1) - g(X_n), \cdots, g(X_{n-1}) - g(X_n))\big) = \dim\big(Vect((N_j(X))_{1 \leq j \leq m})\big)$$

où

$$\forall j \in [\![1, m]\!],\ N_j(X) = \big(g_j(X_1) - g_j(X_n), \cdots, g_j(X_{n-1}) - g_j(X_n)\big)^T.$$

Puisque l'égalité des dimensions est vérifiée (hypothèse de la proposition), on a

$$l := \dim\big(Vect(g_1(X),\cdots,g_m(X))\big) = \dim\big(Vect(g(X_1)-g(X_n),\cdots,g(X_{n-1})-g(X_n))\big).$$

Ainsi par le lemme technique, il existe une partie $J \subset [\![1,m]\!]$ de taille l telle que

$$\begin{aligned} l = \dim\big(Vect\big(g_1(X),\cdots,g_m(X)\big)\big) &= \dim\big(Vect\big((g_j(X))_{j\in J}\big)\big) \\ &= \dim\big(Vect\big((N_j(X))_{j\in J}\big)\big) \\ &= \dim\big(Vect(g(X_1)-g(X_n),\cdots,g(X_{n-1})-g(X_n))\big). \end{aligned}$$

Notons $\widetilde{g} = (g_j)_{j\in J}$ et remarquons que

$$\big\{Q \in \mathscr{P}(\mathscr{X}),\ \ Qg = 0\big\} = \big\{Q \in \mathscr{P}(\mathscr{X}),\ \ Q\widetilde{g} = 0\big\}.$$

Ainsi on peut retenir seulement les contraintes $\widetilde{g}$. Par simplicité d'écriture, on notera toujours $g = (g_j)_{j\in J}$. Définissons la fonction

$$\begin{aligned} f : \mathscr{S} &\to \mathbb{R}^l \\ q &\mapsto \sum_{i=1}^{n} q_i g(X_i). \end{aligned}$$

Montrons que f est une submersion. Posons $\gamma := f \circ \pi^{-1} : U \to \mathbb{R}^l$. Ainsi

$$\gamma(q) = \sum_{i=1}^{n-1} q_i g(X_i) + \left(1 - \sum_{i=1}^{n-1} q_i\right) g(X_n).$$

Ainsi γ est différentiable et pour tout $q \in U$

$$D\gamma(q) = (g(X_1)-g(X_n),\cdots,g(X_{n-1})-g(X_n)).$$

Puisque

$$rg\big(D\gamma(q)\big) = \dim\big(Vect(g(X_1)-g(X_n),\cdots,g(X_{n-1})-g(X_n))\big) = l$$

on en déduit que $D\gamma(q)$ est surjective. Ainsi f est une submersion et $\mathscr{E}$ est une sous variété de dimension $n-1-l$ par le lemme A.2.3. On peut donc conclure que $\mathscr{P}^I = \psi^{-1}(\mathscr{E})$ est une sous variété de dimension $n-1-l$ du fait que ψ^{-1} est un plongement en utilisant le lemme A.2.2. □

Enfin énonçons une proposition donnant une condition suffisante pour vérifier l'égalité des dimensions de la proposition 1.3.4.

Proposition 1.3.6. *Notons pour tout $k \in [\![1,m]\!]$, $M_k(X) = g_k(X) = \big(g_k(X_1),\cdots,g_k(X_n)\big)^T$. Supposons que*

$$\dim Vect\big((1,\cdots,1)^T, M_1(X),\cdots,M_m(X)\big) = m+1.$$

Alors on vérifie l'égalité des dimensions de la proposition 1.3.4.

Démonstration.
Puisque

$$\dim Vect\left((1,\cdots,1)^T, M_1(X),\cdots,M_m(X)\right) = m+1$$

on en déduit que

$$\dim Vect\,(M_1(X),\cdots,M_m(X)) = m.$$

Ainsi il suffit de montrer que

$$\dim\left(Vect(g(X_1)-g(X_n),\cdots,g(X_{n-1})-g(X_n))\right) = m.$$

Par la remarque B.1, cela est équivalent à montrer que

$$\dim\,(Vect(N_1(X),\cdots,N_m(X))) = m$$

où pour tout $j \in [\![1,m]\!]$ *:*

$$N_j(X) = (g_j(X_1)-g_j(X_n),\cdots,g_j(X_{n-1})-g_j(X_n))^T$$

Remarquons que pour tout $j \in [\![1,m]\!]$

$$M_j(X) - g_j(X_n)(1,\cdots,1)^T = (g_j(X_1)-g_j(X_n),\cdots,g_j(X_{n-1})-g_j(X_n),0)^T.$$

Ainsi

$$\dim\,(Vect(N_1(X),\cdots,N_m(X))) = \dim\left(Vect\left((M_j(X)-g_j(X_n)(1,\cdots,1)^T)_{1\le j\le m}\right)\right).$$

Puisque $\dim Vect\left((1,\cdots,1)^T, M_1(X),\cdots,M_m(X)\right) = m+1$*, on en déduit que*

$$\dim\left(Vect\left((M_j(X)-g_j(X_n)(1,\cdots,1)^T)_{1\le j\le m}\right)\right) = m.$$

□

2

Mesure empirique informée par une information auxiliaire forte

Ce chapitre a pour objectif de définir la mesure empirique informée en exploitant le théorème 1.3.1. Ce dernier nous conduit à nous intéresser à deux problèmes d'optimisation distincts. La mesure empirique informée s'obtient comme une approximation commune de la solution de chacun de ces deux problèmes de minimisation donnés. Par la suite, nous nous sommes intéressés à la répartition des poids de la mesure empirique informée. Enfin nous proposons une généralisation de la mesure empirique informée pour le cas d'une information auxiliaire forte donnée par une fonctionnelle de la mesure P.

2.1 Motivations du chapitre 2

Au chapitre 1, nous avons injecté de l'information auxiliaire en projetant la mesure empirique sur la sous variété informée. Cela nous a conduit à nous intéresser à la minimisation de la divergence de Kullback-Leibler et de sa version duale. Puisque la divergence de Kullback-Leibler n'est pas symétrique, nous avons affaire à deux problèmes d'optimisation distincts. Le **premier problème d'optimisation** est le suivant

$$arg \min_{Q \in \mathscr{P}^I} KL(\mathbb{P}_n || Q)$$

où $\mathscr{P}^I$ est l'ensemble des mesures de probabilités à coefficients strictement positifs vérifiant l'information auxiliaire I. Calculons

$$\begin{aligned} KL(\mathbb{P}_n || Q) &= \sum_{i=1}^{n} \frac{1}{n} \log\left(\frac{1}{nq_i}\right) \\ &= -\log(n) - \frac{1}{n} \sum_{i=1}^{n} \log q_i. \end{aligned}$$

Ainsi le problème de minimisation se transforme en un problème de maximisation,

$$\begin{aligned} arg \min_{Q\in\mathscr{P}^I} KL(\mathbb{P}_n||Q) &= arg \max_{Q\in\mathscr{P}^I} \sum_{i=1}^{n} \log q_i \\ &= arg \max_{Q\in\mathscr{P}^I} \prod_{i=1}^{n} q_i. \end{aligned}$$

Dans le cas où l'information auxiliaire I est donnée par la connaissance d'une espérance d'un vecteur de fonctions Pg, on retrouve la notion de **vraisemblance empirique** développée par *Owen* [31] [32].

Le **second problème d'optimisation** est le suivant

$$arg \min_{Q\in\mathscr{P}^I} KL(Q||\mathbb{P}_n).$$

Calculons

$$\begin{aligned} KL(Q||\mathbb{P}_n) &= \sum_{i=1}^{n} q_i \log(nq_i) \\ &= \log n + \sum_{i=1}^{n} q_i \log q_i. \end{aligned}$$

D'où

$$arg \min_{Q\in\mathscr{P}^I} KL(Q||\mathbb{P}_n) = arg \min_{Q\in\mathscr{P}^I} \sum_{i=1}^{n} q_i \log q_i.$$

Dans ce chapitre, on s'intéressera à l'étude de ces deux problèmes d'optimisation dans le cas où l'information auxiliaire forte est donnée par un vecteur d'espérance. Une généralisation sera donnée en fin de chapitre.

Notation 2.1.1. Soient $(X_n)_{n\in\mathbb{N}^*} \subset \mathbb{R}^k$ avec $k \geq 1$ et $(Y_n)_{n\in\mathbb{N}^*} \subset \mathbb{R}$ deux suites de variables aléatoires. On note :

- $X_n = o_p(Y_n)$ si la suite $\left(\frac{\|X_n\|}{|Y_n|}\right)_{n\in\mathbb{N}^*}$ tend en probabilité vers 0 lorsque n tend vers $+\infty$.
- $X_n = O_p(Y_n)$ si $\left(\frac{\|X_n\|}{|Y_n|}\right)_{n\in\mathbb{N}^*}$ est une suite tendue.

2.2 Étude des deux problèmes d'optimisation

2.2.1 Premier problème d'optimisation

Soit $X_1, \cdots, X_n$ un échantillon de taille $n \in \mathbb{N}^*$ *i.i.d.* de variables aléatoires à valeurs dans $\mathscr{X}$ et de loi inconnue P. On suppose qu'on dispose d'une information auxiliaire forte I donnée par Pg où $g = (g_1, \cdots, g_m)^T : \mathscr{X} \to \mathbb{R}^m$ est une fonction intégrable et $m \in \mathbb{N}^*$. Ainsi

$$\mathscr{P}^I = \left\{Q \in \mathscr{P}(\mathscr{X}),\ Qg = Pg\right\}.$$

Le **premier problème d'optimisation** peut s'écrire de la manière suivante

$$\begin{cases} \max_{\mathbf{q}} \sum_{i=1}^{n} \log(q_i) \\ \sum_{i=1}^{n} q_i = 1, \\ \forall i \in [\![1,n]\!]\ ,\ q_i > 0, \\ n\sum_{i=1}^{n} q_i(g(X_i) - Pg) = 0. \end{cases}$$

Remarque 2.2.1. *La condition $n\sum_{i=1}^{n} q_i(g(X_i) - Pg) = 0$ est donnée sous cette forme afin de faciliter l'expression de la solution donnée dans le théorème 2.2.2.*

Ce problème d'optimisation est connu sous le nom de **vraisemblance empirique** étudiée par *Owen* [31] [32]. On notera $\mathscr{C}_n$ l'ensemble des contraintes

$$\mathscr{C}_n = \left\{ q \in [0,1]^n,\ \sum_{i=1}^{n} q_i = 1,\ n\sum_{i=1}^{n} q_i(g(X_i) - Pg) = 0 \right\}.$$

Notons

$$\forall k \in [\![1,m]\!],\ M_k(X) = (g_k(X_1), \cdots, g_k(X_n))^T,$$
$$\forall q \in]0,1[^n,\ f(q) = \sum_{i=1}^{n} \log q_i.$$

Énonçons le théorème suivant :

Théorème 2.2.2. *Supposons que Pg appartienne à l'enveloppe convexe de $\{g(X_1), \cdots, g(X_n)\}$ et*

$$\dim Vect\left((1,\cdots,1)^T, M_1(X), \cdots, M_m(X)\right) = m+1. \tag{2.1}$$

Alors le premier problème d'optimisation admet une unique solution $q^* = (q_1^*, \cdots, q_n^*) \in \mathscr{C}_n$. De plus pour tout $i \in [\![1,n]\!]$

$$q_i^* > 0,$$
$$q_i^* = \frac{1}{n}\frac{1}{1 + \langle \lambda^*, g(X_i) - Pg\rangle}$$

pour un unique $\lambda^* \in \bigcap_{i=1}^{n} \left\{\lambda \in \mathbb{R}^m,\ 1 + \langle \lambda, g(X_i) - Pg\rangle \geq \frac{1}{n}\right\}$.

Remarque 2.2.3. *L'hypothèse (2.1) n'est pas restrictive puisque si elle n'est pas vérifiée cela signifie que certaines contraintes sont inutiles. De plus, ces deux hypothèses sont des conditions suffisantes pour s'assurer que $\mathscr{P}^I$ est une sous-variété (proposition 1.3.4).*

Démonstration.

Existence : *Puisque Pg appartient à l'enveloppe convexe de $\{g(X_1), \cdots, g(X_n)\}$, l'ensemble des contraintes $\mathscr{C}_n$ est non vide. Remarquons que $\mathscr{C}_n$ est fermé et borné. Ainsi $\mathscr{C}_n$ est compact. On en déduit par continuité de f qu'il existe $q^* \in \mathscr{C}_n$ tel que*

$$\sum_{i=1}^{n} \log q_i^* = \max_{\boldsymbol{q}} \sum_{i=1}^{n} \log q_i.$$

Notons que pour tout $i \in [\![1, n]\!]$, $q_i^* > 0$ (dans le cas contraire, on aurait que $\sum_{i=1}^{n} \log q_i^* = -\infty$). Ainsi $q^* \in \widetilde{\mathscr{C}}_n = \left\{ q \in]0,1[^n,\ \sum_{i=1}^{n} q_i = 1,\ n\sum_{i=1}^{n} q_i (g(X_i) - Pg) = 0 \right\}$.

Unicité : *Par l'hypothèse (2.1) et que q^* est un maximum global de f sur $\widetilde{\mathscr{C}}_n$, on peut donc appliquer le théorème des extrema liés. Ainsi il existe $(\lambda^*, \mu) \in \mathbb{R}^m \times \mathbb{R}$ tel que*

$$Df(q^*) = \mu(1, \cdots, 1)^T + n \sum_{j=1}^{m} \lambda_j^* \left(g_j(X_1) - Pg_j, \cdots, g_j(X_n) - Pg_j \right)^T.$$

On en déduit que pour tout $i \in [\![1, n]\!]$,

$$\frac{1}{q_i^*} = \mu + n\langle \lambda^*, g(X_i) - Pg \rangle.$$

Puisque $\sum_{i=1}^{n} q_i^* = 1$, on a que

$$\mu = n,$$
$$q_i^* = \frac{1}{n} \frac{1}{1 + \langle \lambda^*, g(X_i) - Pg \rangle}.$$

Remarquons que

$$\forall i \in [\![1, n]\!],\ \left(q_i^* \in]0,1] \Longleftrightarrow 1 + \langle \lambda^*, g(X_i) - Pg \rangle \geq \frac{1}{n} \right),$$
$$f(q^*) = -\sum_{i=1}^{n} \log\left(1 + \langle \lambda^*, g(X_i) - Pg \rangle\right) - n \log n.$$

Posons

$$C = \bigcap_{i=1}^{n} \left\{ \lambda \in \mathbb{R}^m,\ 1 + \langle \lambda^*, g(X_i) - Pg \rangle \geq \frac{1}{n} \right\},$$
$$\forall \lambda \in C,\ L(\lambda) := -\sum_{i=1}^{n} \log\left(1 + \langle \lambda, g(X_i) - Pg \rangle\right).$$

Notons que C est un ensemble convexe, fermé et non vide. Par dualité convexe, cela revient à minimiser L sur C. Il nous suffit de démontrer que λ^* est l'unique minimum global de L. Puisque L est une fonction différentiable et convexe, on a

$$\nabla L(\lambda) = 0 \Longleftrightarrow \lambda \in arg\min_C L.$$

En remarquant que

$$\nabla L(\lambda) = 0 \Longleftrightarrow \sum_{i=1}^{n} \frac{g(X_i) - Pg}{1 + \langle \lambda, g(X_i) - Pg \rangle} = 0$$

on en déduit que $\nabla L(\lambda^*) = 0$ car $\sum_{i=1}^{n} q_i^*(g(X_i) - Pg) = 0$. Ainsi λ^* est un minimum global de L sur C. Pour montrer l'unicité, il nous suffit de prouver que L est une fonction strictement convexe. La matrice hessienne de L est

$$Hess_L = \sum_{i=1}^{n} \frac{(g(X_i) - Pg)(g(X_i) - Pg)^T}{\left(1 + \langle \lambda, g(X_i) - Pg \rangle\right)^2}.$$

Par l'hypothèse 2.1, on a

$$\dim Vect\left(M_1(X) - Pg_1(1, \cdots, 1)^T, \cdots, M_m(X) - Pg_m(1, \cdots, 1)^T\right) = m.$$

En posant la matrice $A \in \mathcal{M}_{n,m}(\mathbb{R})$

$$A = \begin{pmatrix} g_1(X_1) - Pg_1 & \cdots & g_m(X_1) - Pg_m \\ \vdots & \cdots & \vdots \\ g_1(X_n) - Pg_1 & \cdots & g_m(X_n) - Pg_m \end{pmatrix}$$

et puisque $m = rg(A) = rg(A^T)$, on en déduit que

$$\dim Vect\left(g(X_1) - Pg, \cdots, g(X_n) - Pg\right) = m.$$

Ainsi la matrice $Hess_L$ est définie positive. On peut donc conclure que L est strictement convexe et

$$\lambda^* = arg\min_{\lambda \in C} L(\lambda).$$

□

Afin de déterminer λ^* numériquement, introduisons la fonction suivante définie pour tout $z \in \mathbb{R}$,

$$\log_*(z) = \begin{cases} \log(z) \ \text{ si } z > \frac{1}{n} \\ \log(1/n) - 3/2 + 2nz - (nz)^2/2 \ \text{ si } z \leq \frac{1}{n} \end{cases}.$$

Remarquons que le polynôme atteint son maximum en $z = \frac{2}{n} > \frac{1}{n}$. On pose alors pour tout $\lambda \in \mathbb{R}^m$

$$L_*(\lambda) = -\sum_{i=1}^{n} \log_*(1 + \langle \lambda, (g(X_i) - Pg) \rangle). \tag{2.2}$$

Énonçons maintenant la proposition suivante montrant qu'il suffit de minimiser L_* sur $\mathbb{R}^m$:

Proposition 2.2.4. *Supposons que Pg appartienne à l'enveloppe convexe de $\{g(X_1), \cdots, g(X_n)\}$ et*

$$\dim Vect\left((1, \cdots, 1)^T, M_1(X), \cdots, M_m(X)\right) = m+1. \tag{2.3}$$

Alors

$$\lambda^* = arg \min_{\lambda \in \mathbb{R}^m} L_*(\lambda)$$

avec L_* donnée par (2.2).

Démonstration.
Par ce qui précède, il existe une unique solution q^* telle que

$$\forall i \in [\![1, n]\!],\ q_i^* = \frac{1}{n} \frac{1}{1 + \langle \lambda^*, g(X_i) - Pg \rangle}$$

pour un unique $\lambda^* \in \bigcap_{i=1}^n \left\{\lambda \in \mathbb{R}^m,\ 1 + \langle \lambda, g(X_i) - Pg \rangle \geq \frac{1}{n}\right\}$. Puisque L_* est une fonction différentiable et convexe, il suffit de montrer que L_* est strictement convexe et $\nabla L_*(\lambda^*) = 0$. Tout d'abord calculons le gradient de L_* pour tout $\lambda \in \mathbb{R}^m$

$$\nabla L_*(\lambda) = \sum_{i=1}^n l_i(\lambda)$$

où pour tout $i \in [\![1, n]\!]$

$$l_i(\lambda) = \begin{cases} -\frac{g(X_i) - Pg}{1 + \langle \lambda, (g(X_i) - Pg) \rangle)} & si\ 1 + \langle \lambda, g(X_i) - Pg \rangle > \frac{1}{n} \\ 2n + n^2 (1 + \langle \lambda, g(X_i) - Pg \rangle)(g(X_i) - Pg) & si\ 1 + \langle \lambda, g(X_i) - Pg \rangle \leq \frac{1}{n} \end{cases}.$$

Puisque $\lambda^* \in \bigcap_{i=1}^n \left\{\lambda \in \mathbb{R}^m,\ 1 + \langle \lambda, g(X_i) - Pg \rangle \geq \frac{1}{n}\right\}$, on en déduit que

$$\nabla L_*(\lambda^*) = 0.$$

Montrons que la fonction L_* est strictement convexe sur $\mathbb{R}^m$. La matrice hessienne de L_* est

$$\forall \lambda \in \mathbb{R}^m,\ Hess_{L_*}(\lambda) = \sum_{i=1}^n h_i(\lambda)$$

où pour tout $i \in [\![1, n]\!]$

$$h_i(\lambda) = \begin{cases} \frac{(g(X_i) - Pg)(g(X_i) - Pg)^T}{(1 + \langle \lambda, (g(X_i) - Pg) \rangle)^2} & si\ 1 + \langle \lambda, (g(X_i) - Pg) > \frac{1}{n} \\ n^2 (g(X_i) - Pg)(g(X_i) - Pg)^T & si\ 1 + \langle \lambda, (g(X_i) - Pg) \leq \frac{1}{n} \end{cases}.$$

Par l'hypothèse 2.3, on en déduit que la matrice hessienne de L_* est définie positive. Ainsi L_* est strictement convexe. □

Dorénavant on supposera que $Pg = 0$ quitte à poser $h = g - Pg$. Notons $\hat{\lambda}_n = \lambda^*$ et énonçons un théorème limite portant sur $\hat{\lambda}_n$:

Théorème 2.2.5. *On suppose que $\mathbb{E}\|g(X)\|^2 < +\infty$ et Σ est définie positive. On suppose aussi que les hypothèses du théorème 2.2.2 sont satisfaites. Posons $\Sigma = Pgg^T$ et $\Sigma_n = \mathbb{P}_n gg^T$. Alors on a :*

1. $\max_{1\le i\le n} | < \hat{\lambda}_n, g(X_i) > | = o_p(1)$.
2. $\hat{\lambda}_n = \Sigma_n^{-1}\mathbb{P}_n g + o_p(\frac{1}{\sqrt{n}})$.
3. $\sqrt{n}\hat{\lambda}_n \Rightarrow \mathcal{N}(0, \Sigma^{-1})$.

Remarque 2.2.6. *On peut remplacer Σ_n par la variance empirique de g. En effet :*

$$\Sigma_n = Var_n(g) + (\mathbb{P}_n g)^T \mathbb{P}_n g = Var_n(g) + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Démonstration (du théorème).
Tout d'abord montrons que $\|\hat{\lambda}_n\| = O_p\left(\frac{1}{\sqrt{n}}\right)$. On écrit $\hat{\lambda}_n = \rho_n\theta_n$ où $\rho_n \ge 0$, $\|\theta_n\| = 1$. On pose pour tout $\lambda \in \mathbb{R}^m$

$$\phi(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{1 + \langle \lambda, g(X_i)\rangle}.$$

Par définition de $\hat{\lambda}_n$, on a $\phi(\hat{\lambda}_n) = 0$. Ainsi

$$0 = \|\phi(\hat{\lambda}_n)\| = \|\phi(\rho_n\theta_n)\| = | \langle \theta_n, \phi(\rho_n\theta_n)\rangle | .$$

Remarquons que

$$\begin{aligned} 0 = | \langle \theta_n, \phi(\rho_n\theta_n)\rangle | &= \frac{1}{n} | \langle \theta_n, \sum_{i=1}^{n}\left(1 - \frac{\rho_n\langle\theta_n, g(X_i)\rangle}{1 + \rho_n\langle\theta_n, g(X_i)\rangle}\right) g(X_i)\rangle | \\ &= \frac{1}{n} | \rho_n\langle\theta_n, \sum_{i=1}^{n} \frac{\langle\theta_n, g(X_i)\rangle g(X_i)}{1 + \rho_n\langle\theta_n, g(X_i)\rangle} - \langle\theta_n, \sum_{i=1}^{n} g(X_i)\rangle | \\ &\ge \frac{\rho_n}{n} | \theta_n^T \sum_{i=1}^{n} \frac{g(X_i)g(X_i)^T}{1 + \rho_n\langle\theta_n, g(X_i)\rangle}\theta_n | - \|\mathbb{P}_n g\| \end{aligned}$$

car $\|\theta_n\| = 1$. Ainsi

$$\|\mathbb{P}_n g\| \ge \frac{\rho_n}{n}\theta_n^T \sum_{i=1}^{n} \frac{g(X_i)g(X_i)^T}{| 1 + \rho_n\langle\theta_n, g(X_i)\rangle |}\theta_n.$$

Posons $Z_n = \max_{1\le i\le n}\|g(X_i)\|$. Puisque

$$\begin{aligned} | 1 + \rho_n\langle\theta_n, g(X_i)\rangle | &\le 1 + \rho_n \max_{1\le i\le n} | \langle\theta_n, g(X_i)\rangle | \\ &\le 1 + \rho_n Z_n \end{aligned}$$

on a

$$\|\mathbb{P}_n g\| \geq \frac{\rho_n}{n(1+\rho_n Z_n)} \sum_{i=1}^{n} \theta_n^T g(X_i) g(X_i)^T \theta_n = \frac{\rho_n}{1+\rho_n Z_n} \theta^T \Sigma_n \theta_n.$$

De plus $\theta_n^T \Sigma_n \theta_n \geq \inf_{\|\theta\|=1} \theta^T \Sigma_n \theta = \sigma_n$ *où* σ_n *est la valeur propre minimale de* Σ_n. *Ainsi*

$$\|\mathbb{P}_n g\| \geq \frac{\rho_n}{1+\rho_n Z_n} \sigma_n.$$

De plus, nous avons le résultat suivant concernant σ_n.

Lemme 2.2.7. *Notons* σ *(resp.* σ_n*) la valeur propre minimale de* Σ *(resp.* Σ_n*). Alors :*

$$\sigma_n \xrightarrow[n\to+\infty]{p.s.} \sigma > 0.$$

Démonstration (du lemme).
Posons $\gamma_n(\theta) = \langle \Sigma_n \theta, \theta \rangle$ *et* $\psi : C^0(\mathbb{R}^m) \to \mathbb{R}$ *définie par* $\psi(f) = \inf_{\|\theta\|=1} f(\theta)$. *Montrons que* ψ *est continue. En effet*

$$\begin{aligned} |\psi(f_1) - \psi(f_2)| &= |\inf_{\|\theta\|=1} f_1(\theta) - \inf_{\|\theta\|=1} f_2(\theta)| \\ &\leq \|f_1 - f_2\|_\infty. \end{aligned}$$

Ainsi ψ *est* 1*-lipschitzienne donc continue. De plus, puisque* $\|\Sigma_n - \Sigma\| \xrightarrow[n\to+\infty]{p.s.} 0$ *et par continuité du produit scalaire, on a*

$$\gamma_n \xrightarrow[n\to+\infty]{\|\cdot\|_\infty,\ p.s.} \gamma.$$

En utilisant la continuité de ψ, *on a que*

$$\psi(\gamma_n) \xrightarrow[n\to+\infty]{p.s.} \psi(\gamma).$$

Donc

$$\sigma_n \xrightarrow[n\to+\infty]{p.s.} \sigma.$$

□

On a

$$\frac{\rho_n}{1+\rho_n Z_n} \leq \frac{\|\mathbb{P}_n g\|}{\sigma_n} \Longleftrightarrow \rho_n \left(1 - \frac{Z_n \|\mathbb{P}_n g\|}{\sigma_n}\right) \leq \frac{\|\mathbb{P}_n g\|}{\sigma_n}.$$

Il nous suffit de montrer que $Z_n \|\mathbb{P}_n g\|$ *tend en probabilité vers* 0 *pour en déduire que* $p_n = O_p\left(\frac{1}{\sqrt{n}}\right)$. *Pour cela, nous allons utiliser le lemme d'Owen (1990) [31].*

Lemme 2.2.8. *Soit* $(Y_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires i.i.d. à valeurs dans* $\mathbb{R}_+$ *et* $Z_n = \max_{1\le i\le n} Y_i$. *Si* $\mathbb{E}Y^2 < +\infty$ *alors*

$$Z_n = o_p(\sqrt{n}),$$
$$\frac{1}{n}\sum_{i=1}^{n} Y_i^3 = o_p(\sqrt{n}).$$

Puisque $\mathbb{E}\|g(X)\|^2 < +\infty$, *on en déduit par le lemme que* $Z_n = o_p(\sqrt{n})$. *D'où*

$$Z_n\|\mathbb{P}_n g\| = \frac{Z_n}{\sqrt{n}}\sqrt{n}\|\mathbb{P}_n g\| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Puisque $\rho_n = \|\hat{\lambda}_n\|$, *on obtient que* $\|\hat{\lambda}_n\| = O_p\left(\frac{1}{\sqrt{n}}\right)$. *Ainsi*

$$\begin{aligned}\max_{1\le i\le n} |\langle\hat{\lambda}_n, g(X_i)\rangle| &\le \|\hat{\lambda}_n\| \max_{1\le i\le n}\|g(X_i)\| \\ &= \|\hat{\lambda}_n\| Z_n \\ &= O_p\left(\frac{1}{\sqrt{n}}\right) o_p(\sqrt{n}) \\ &= o_p(1).\end{aligned}$$

Nous avons donc prouvé la première assertion du théorème à savoir que

$$\max_{1\le i\le n} |\langle\hat{\lambda}_n, g(X_i)\rangle| = o_p(1).$$

Concernant la seconde assertion, écrivons

$$\begin{aligned}0 = \phi(\hat{\lambda}_n) &= \frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{1+\langle\hat{\lambda}_n, g(X_i)\rangle} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \langle\hat{\lambda}_n, g(X_i)\rangle + \frac{\langle\hat{\lambda}_n, g(X_i)\rangle^2}{1+\langle\hat{\lambda}_n, g(X_i)\rangle}\right) g(X_i) \\ &= \mathbb{P}_n g - \left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)g(X_i)^T\right)\hat{\lambda}_n + \frac{1}{n}\sum_{i=1}^{n}\frac{\langle\hat{\lambda}_n, g(X_i)\rangle^2}{1+\langle\hat{\lambda}_n, g(X_i)\rangle} g(X_i).\end{aligned}$$

D'où

$$\Sigma_n\hat{\lambda}_n = \mathbb{P}_n g + \frac{1}{n}\sum_{i=1}^{n}\frac{\langle\hat{\lambda}_n, g(X_i)\rangle^2}{1+\langle\hat{\lambda}_n, g(X_i)\rangle} g(X_i).$$

En utilisant le fait que

$$\frac{1}{1+\max_{1\le i\le n}|\langle\hat{\lambda}_n, g(X_i)\rangle|} \le \frac{1}{1+\langle\hat{\lambda}_n, g(X_i)\rangle} \le \frac{1}{1-\max_{1\le i\le n}|\langle\hat{\lambda}_n, g(X_i)\rangle|}$$

on a que $\max_{1\le i\le n}\frac{1}{|1+\langle\hat{\lambda}_n,g(X_i)\rangle|}=1+o_p(1)$ *et*

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\langle\hat{\lambda}_n,g(X_i)\rangle^2}{1+\langle\hat{\lambda}_n,g(X_i)\rangle}g(X_i)\right\| &\le \frac{1}{n}\|\hat{\lambda}_n\|^2\sum_{i=1}^{n}\frac{\|g(X_i)\|^3}{|\,1+\langle\hat{\lambda}_n,g(X_i)\rangle\,|}\\ &\le (1+o_p(1))\|\hat{\lambda}_n\|^2\left(\frac{1}{n}\sum_{i=1}^{n}\|g(X_i)\|^3\right)\\ &\le (1+o_p(1))O_p\left(\frac{1}{n}\right)o_p(\sqrt{n})=o_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

car $\|\hat{\lambda}_n\|^2=O_p\left(\frac{1}{n}\right)$ *et par le lemme d'Owen ([31])* $\frac{1}{n}\sum_{i=1}^{n}\|g(X_i)\|^3=o_p(\sqrt{n})$. *Ainsi puisque*

$$\hat{\lambda}_n=\Sigma_n^{-1}\mathbb{P}_n g+o_p\left(\frac{1}{\sqrt{n}}\right)$$

on obtient que

$$\sqrt{n}\hat{\lambda}_n\Rightarrow\mathcal{N}(0,\Sigma^{-1}).$$

□

2.2.2 Second problème d'optimisation

Dans cette sous-section, on conserve le même cadre que dans la sous-section précédente. On s'intéresse au **second problème d'optimisation** suivant

$$\begin{cases}\min_{\mathbf{q}}\sum_{i=1}^{n}q_i\log(q_i)\\ \sum_{i=1}^{n}q_i=1,\\ \forall i\in[\![1,n]\!]\ ,\ q_i>0,\\ n\sum_{i=1}^{n}q_i(g(X_i)-Pg)=0.\end{cases}$$

Posons

$$\forall q\in]0,1[^n,\ \varphi(q)=\sum_{i=1}^{n}q_i\log q_i.$$

Rappelons les notations suivantes

$$\begin{aligned}&\mathscr{C}_n=\left\{q\in[0,1]^n,\ \sum_{i=1}^{n}q_i=1,\ \sum_{i=1}^{n}q_i(g(X_i)-Pg)=0\right\},\\ &\forall k\in[\![1,m]\!],\ M_k(X)=(g_k(X_1),\cdots,g_k(X_n))^T.\end{aligned}$$

Le théorème suivant énonce l'existence et l'unicité de la solution du **second problème d'optimisation**

Théorème 2.2.9. *Supposons que Pg appartienne à l'enveloppe convexe de $\{g(X_1), \cdots, g(X_n)\}$ et*

$$\dim Vect\left((1, \cdots, 1)^T, M_1(X), \cdots, M_m(X)\right) = m+1. \tag{2.4}$$

Alors le second problème d'optimisation admet une unique solution $q^* = (q_1^*, \cdots, q_n^*) \in \mathscr{C}_n$. De plus pour tout $i \in [\![1, n]\!]$

$$q_i^* = \frac{\exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)}$$

pour un unique $\lambda^* \in \mathbb{R}^m$.

Démonstration.

Existence : *Puisque $\mathscr{C}_n$ est compact (démontration du théorème 2.2.2) et φ est continue, il existe $q^* \in \mathscr{C}_n$ tel que*

$$\sum_{i=1}^n q_i^* \log q_i^* = \min_{q \in \mathscr{C}_n} \sum_{i=1}^n q_i \log q_i.$$

Unicité : *Par l'hypothèse 2.4 et que q^* est un maximum global de φ sur $\mathscr{C}_n$, on peut donc appliquer le théorème des extrema liés. Ainsi il existe $(\lambda^*, \mu) \in \mathbb{R}^m \times \mathbb{R}$ tel que*

$$D\varphi(q^*) = \mu(1, \cdots, 1)^T + \sum_{j=1}^m \lambda_j^* \left(g_j(X_1) - Pg_j, \cdots, g_j(X_n) - Pg_j\right)^T.$$

Ainsi pour tout $i \in [\![1, n]\!]$

$$\log q_i^* + 1 = \mu + \langle \lambda^*, g(X_i) - Pg\rangle.$$

Du fait que $\sum_{i=1}^n q_i^* = 1$, on en déduit que

$$\mu = 1 - \log\left(\sum_{i=1}^n \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\right),$$
$$q_i^* = \frac{\exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)}.$$

En injectant cette expression, on a

$$\begin{aligned}\varphi(q^*) &= \frac{\sum_{i=1}^n \langle \lambda^*, g(X_i) - Pg\rangle \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)} - \log\left(\sum_{i=1}^n \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\right) \\ &= \frac{\langle \lambda^*, \sum_{i=1}^n (g(X_i) - Pg) \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\rangle}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)} - \log\left(\sum_{i=1}^n \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\right).\end{aligned}$$

Puisque $\sum_{i=1}^n q_i^* \left(g(X_i) - Pg\right) = 0$, *on a*

$$\frac{\sum_{i=1}^n (g(X_i) - Pg) \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)} = 0.$$

Ainsi

$$\varphi(q^*) = -\log\left(\sum_{i=1}^n \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\right).$$

Posons la fonction $F : \mathbb{R}^m \to \mathbb{R}$ *définie par*

$$\forall \lambda \in \mathbb{R}^m,\ F(\lambda) = -\log\left(\sum_{i=1}^n \exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)\right).$$

Par dualité convexe, cela revient à maximiser la fonction F *sur* $\mathbb{R}^m$. *Cette fonction est différentiable et concave. Ainsi* $\nabla F(\lambda) = 0$ *est équivalent à* $\lambda \in arg\max_{x\in\mathbb{R}^m} F(x)$. *Dans notre cas*

$$\nabla F(\lambda) = \frac{-\sum_{i=1}^n (g(X_i) - Pg) \exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)}.$$

D'où $\nabla F(\lambda^*) = 0$. *Il reste à montrer que* F *est strictement concave. Pour tout* $j, k \in [\![1, n]\!]$

$$\begin{aligned}\frac{\partial^2 F}{\partial\lambda_j \partial\lambda_k} &= \frac{-\sum_{i=1}^n (g_j(X_i) - Pg_j)(g_k(X_i) - Pg_k) \exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)} \\ &\quad + \frac{\sum_{i=1}^n (g_j(X_i) - Pg_j) \exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)} \frac{\sum_{i=1}^n (g_k(X_i) - Pg_k) \exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)}.\end{aligned}$$

En notant pour tout $\lambda \in \mathbb{R}^m$

$$\forall i \in [\![1, n]\!],\ \ q_i(\lambda) = \frac{\exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)},$$

$$Q_n(\lambda) = \sum_{i=1}^n q_i(\lambda)\delta_{X_i}.$$

Notons que pour tout $\lambda \in \mathbb{R}^m$, $Q_n(\lambda)$ *est une mesure de probabilité. La matrice hessienne de* F *est*

$$Hess_F = -Var_{Q_n(\lambda)}(g - Pg) = -Var_{Q_n(\lambda)}(g) < 0.$$

En utilisant l'hypothèse 2.4, on peut en déduire que la matrice hessienne de F *est définie négative. Ainsi* F *est strictement concave. On peut conclure à l'unicité de la solution.* □

De cette démonstration, on en déduit qu'il suffit de maximiser sur $\mathbb{R}^m$ la fonction suivante

$$\begin{array}{rcccl} F_n & : & \mathbb{R}^m & \to & \mathbb{R} \\ & & \lambda & \mapsto & -\log\left(\sum_{i=1}^n \exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)\right) \end{array}$$

afin de trouver λ^*.

Notons $\hat{\lambda}_n := \lambda^*$. On supposera aussi dorénavant que $Pg = 0$ quitte à poser $h = g - Pg$. Ainsi il est possible d'obtenir un théorème limite pour $\hat{\lambda}_n$:

Théorème 2.2.10. *On suppose que* $\mathbb{E}\|g(X)\|^2 < +\infty$ *et* Σ *est définie positive. On suppose aussi que les hypothèses du théorème 2.2.9 sont satisfaites. Posons* $\Sigma = Pgg^T$ *et* $\Sigma_n = \mathbb{P}_n gg^T$. *Alors on a :*

1. $\max_{1\le i\le n} | < \hat{\lambda}_n, g(X_i) > | = o_p(1)$.
2. $\hat{\lambda}_n = -\Sigma_n^{-1}\mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right)$.
3. $\sqrt{n}\hat{\lambda}_n \Rightarrow \mathcal{N}(0, \Sigma^{-1})$.

Remarque 2.2.11. *De même on peut remplacer* Σ_n *par la variance empirique de* g.

Démonstration (du théorème).
Tout d'abord montrons que $\|\hat{\lambda}_n\| = O_p\left(\frac{1}{\sqrt{n}}\right)$. *On écrit* $\hat{\lambda}_n = \rho_n\theta_n$ *où* $\rho_n \ge 0$, $\|\theta_n\| = 1$. *On pose pour tout* $\lambda \in \mathbb{R}^m$

$$\varphi(\lambda) = \sum_{i=1}^n q_i(\lambda) g(X_i)$$

où

$$q_i(\lambda) = \frac{\exp\left(\langle\lambda, g(X_i)\rangle\right)}{\sum_{j=1}^n \exp\left(\langle\lambda, g(X_j)\rangle\right)}.$$

Par définition de $\hat{\lambda}_n$, *on a* $\varphi(\hat{\lambda}_n) = 0$. *Ainsi*

$$0 = \|\phi(\hat{\lambda}_n)\| = \|\phi(\rho_n\theta_n)\| = |\langle\theta_n, \varphi(\rho_n\theta_n)\rangle| .$$

En notant $S_n(\lambda) = \sum_{j=1}^n \exp\left(\langle\lambda, g(X_j)\rangle\right) > 0$, *on a*

$$\frac{\theta_n^T}{S_n(\hat{\lambda}_n)} \sum_{i=1}^n \exp\left(\rho_n\theta_n^T g(X_i)\right) g(X_i) = 0.$$

En utilisant la formule de Taylor-Lagrange à l'ordre 1 *on a que pour tout* $i \in [\![1, n]\!]$, *il existe* r_i *compris entre* 0 *et* $\theta_n^T g(X_i)$ *tel que*

$$\exp\left(\rho_n\theta_n^T g(X_i)\right) = 1 + \rho_n\theta_n^T g(X_i)\exp(\rho_n r_i).$$

Notons S^{m-1} *la sphère unité dans* $\mathbb{R}^m$. *Ainsi*

$$\begin{aligned}
\frac{\theta_n^T}{S_n(\hat{\lambda}_n)} \sum_{i=1}^n \exp\left(\rho_n\theta_n^T g(X_i)\right) g(X_i) &= \frac{n\theta_n^T}{S_n(\hat{\lambda}_n)}\mathbb{P}_n g + \frac{n\rho_n}{S_n(\hat{\lambda}_n)}\frac{1}{n}\sum_{i=1}^n (\theta_n^T g(X_i))^2 \exp(\rho_n r_i) \\
&\ge \frac{n\theta_n^T}{S_n(\hat{\lambda}_n)}\mathbb{P}_n g + \frac{n\rho_n}{S_n(\hat{\lambda}_n)}\frac{1}{n}\sum_{i=1}^n (\theta_n^T g(X_i))^2 1_{\theta_n^T g(X_i)\ge 0} \\
&\ge \frac{n\theta_n^T}{S_n(\hat{\lambda}_n)}\mathbb{P}_n g + \frac{n\rho_n}{S_n(\hat{\lambda}_n)}\inf_{\theta\in S^{m-1}}\frac{1}{n}\sum_{i=1}^n (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\ge 0}
\end{aligned}$$

car $r_i \geq 0$ *lorsque* $\theta_n^T g(X_i) \geq 0$. *D'où*

$$\frac{n\theta_n^T}{S_n(\hat{\lambda}_n)}\mathbb{P}_n g + \frac{n\rho_n}{S_n(\hat{\lambda}_n)} \inf_{\theta \in S^{m-1}} \frac{1}{n}\sum_{i=1}^{n} (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0} \leq 0.$$

En multipliant par $\frac{S_n(\hat{\lambda}_n)}{n}$ *et en remarquant que* $\|\theta_n\| = 1$, *on obtient que*

$$\rho_n \inf_{\theta \in S^{m-1}} \frac{1}{n}\sum_{i=1}^{n} (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0} \leq \|\mathbb{P}_n g\|.$$

Il nous suffit de montrer que $\inf_{\theta \in S^{m-1}} \frac{1}{n}\sum_{i=1}^{n} (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0}$ *tend en probabilité vers une constante strictement positive afin d'en déduire que* $\rho_n = O_p\left(\frac{1}{\sqrt{n}}\right)$. *Pour cela, nous avons besoin du lemme technique suivant.*

Lemme 2.2.12. *On a :*

$$\inf_{\theta \in S^{m-1}} \frac{1}{n}\sum_{i=1}^{n} (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0} \xrightarrow[n\to+\infty]{p.s.} \inf_{\theta \in S^{m-1}} P(\theta^T g)^2 1_{\theta^T g\geq 0} > 0.$$

Démonstration (du lemme).
Posons pour tout $\theta \in S^{m-1}$

$$\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0}.$$

Montrons dans un premier temps que γ_n *est continue. On munit* S^{m-1} *de la topologie induite de* $\mathbb{R}^m$. *Définissons pour tout* $i \in [\![1, n]\!]$

$$\begin{array}{rccc} h_i & : & S^{m-1} & \to & \mathbb{R} \\ & & \theta & \mapsto & \theta^T g(X_i). \end{array}$$

Posons $l : \mathbb{R} \to \mathbb{R}$ *définie par* $l(x) = x^2 1_{x\geq 0}$ *avec* $x \in \mathbb{R}$. *Remarquons que pour tout* $i \in [\![1, n]\!]$ *les fonctions* h_i *et* l *sont continues et*

$$\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l \circ h_i(\theta).$$

Ainsi pour tout $n \in \mathbb{N}^*$, γ_n *est continue sur* S^{m-1}.

Posons $\psi : C^0(S^{m-1}) \to \mathbb{R}$ *définie par* $\psi(f) = \inf_{\|\theta\|=1} f(\theta)$. *Montrons que* ψ *est continue. En effet*

$$\begin{aligned} |\psi(f_1) - \psi(f_2)| &= |\inf f_1 - \inf f_2| \\ &\leq \|f_1 - f_2\|_\infty. \end{aligned}$$

Ainsi ψ *est 1-lipschitzienne donc continue. En remarquant que*

$$\psi(\gamma_n) = \inf_{\theta \in S^{m-1}} \frac{1}{n}\sum_{i=1}^{n} (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0}$$

il suffit de montrer que γ_n converge p.s. uniformément vers γ où pour tout $\theta \in S^{m-1}$

$$\gamma(\theta) = P(\theta^T g)^2 1_{\theta^T g \geq 0}$$

Pour cela remarquons que pour tout $\theta \in S^{m-1}$

$$|\gamma_n(\theta) - \gamma(\theta)| = |(\mathbb{P}_n - P)(\theta^T g)^2 1_{\theta^T g \geq 0})|.$$

On pose la classe de fonctions

$$\mathscr{F} = \left\{ h_\theta : x \mapsto \theta^T g(x))^2 1_{\theta^T g(x) \geq 0} \, , \, \theta \in S^{m-1} \right\}.$$

Puisque la sphère S^{m-1} est compacte, pour tout $x \in \mathscr{X}$ l'application $\theta \mapsto h_\theta(x)$ est continue et

$$\sup_{\theta \in S^{m-1}} |h_\theta| \leq \|g\|^2 \in L^1(P),$$

on en déduit que la classe $\mathscr{F}$ est P-Glivenko-Cantelli. D'où

$$\gamma_n \xrightarrow[n \to +\infty]{\|\cdot\|_\infty, \, p.s.} \gamma.$$

En utilisant la continuité de ψ, on a que

$$\psi(\gamma_n) \xrightarrow[n \to +\infty]{p.s.} \psi(\gamma).$$

On peut conclure que

$$\inf_{\theta \in S^{m-1}} \frac{1}{n} \sum_{i=1}^n (\theta^T g(X_i))^2 1_{\theta^T g(X_i) \geq 0} \xrightarrow[n \to +\infty]{p.s.} \inf_{\theta \in S^{m-1}} P(\theta^T g)^2 1_{\theta^T g \geq 0}.$$

Enfin démontrons que

$$\inf_{\theta \in S^{m-1}} P(\theta^T g)^2 1_{\theta^T g \geq 0} > 0.$$

Par le théorème de continuité sous l'intégrale, puisque $P\|g\|^2 < +\infty$, l'application $\theta \mapsto P(\theta^T g)^2 1_{\theta^T g \geq 0}$ est continue. Par compacité de la sphère, il existe $\theta_* \in S^{m-1}$ tel que

$$\inf_{\theta \in S^{m-1}} P(\theta^T g)^2 1_{\theta^T g \geq 0} = P(\theta_*^T g)^2 1_{\theta_*^T g \geq 0}.$$

Raisonnons par l'absurde et supposons que $P(\theta_*^T g)^2 1_{\theta_*^T g \geq 0} = 0$. Ainsi P-p.s. x, $\theta_*^T g(x) \leq 0$. Autrement dit

$$\mathbb{P}\left(\theta_*^T g(X) > 0\right) = 0.$$

Remarquons que

$$0 = \theta_*^T P g = P \theta_*^T g = P(\theta_*^T g) 1_{\theta_*^T g \leq 0}.$$

Ainsi P-p.s. x, $\theta_*^T g(x) = 0$. En notant que Σ est une matrice définie positive, on obtient une contradiction

$$0 = P(\theta_*^T g)^2 = \theta_*^T P g g^T \theta_* = \theta_*^T \Sigma \theta_* > 0.$$

□

Puisque $\rho_n = \|\hat{\lambda}_n\|$*, on en déduit par le lemme que*

$$\|\hat{\lambda}_n\| = O_p\left(\frac{1}{\sqrt{n}}\right),$$
$$\max_{1\le i\le n} |\langle \hat{\lambda}_n, g(X_i)\rangle| \le \|\hat{\lambda}_n\| \max_{1\le i\le n} \|g(X_i)\|.$$

Il nous suffit de démontrer que $\max_{1\le i\le n}\|g(X_i)\| = o_p(\sqrt{n})$*. Pour cela, nous allons utiliser le lemme d'Owen (1990)*

Lemme 2.2.13. *Soit* $(Y_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires i.i.d. à valeurs dans* $\mathbb{R}_+$ *et* $Z_n = \max_{1\le i\le n} Y_i$*. Si* $\mathbb{E}Y^2 < +\infty$ *alors*

$$Z_n = o_p(\sqrt{n}),$$
$$\frac{1}{n}\sum_{i=1}^n Y_i^3 = o_p(\sqrt{n}).$$

Puisque $\mathbb{E}\|g(X)\|^2 < +\infty$*, on en déduit par le lemme d'Owen que* $\max_{1\le i\le n}\|g(X_i)\| = o_p(\sqrt{n})$*. D'où la première assertion*

$$\max_{1\le i\le n} |\langle \hat{\lambda}_n, g(X_i)\rangle| = o_p(1).$$

Concernant la seconde assertion, appliquons la formule de Taylor-Lagrange à l'ordre 2*. Plus précisément pour tout* $i \in [\![1, n]\!]$*, il existe* s_i *compris entre* 0 *et* $\theta_n^T g(X_i)$ *tel que*

$$\begin{aligned}\exp\left(\rho_n\theta_n^T g(X_i)\right) &= 1 + \rho_n\theta_n^T g(X_i) + \rho_n^2\left(\theta_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right)\\ &= 1 + \hat{\lambda}_n^T g(X_i) + \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right).\end{aligned}$$

Ainsi

$$\begin{aligned}0 &= \frac{1}{S_n(\hat{\lambda}_n)}\sum_{i=1}^n \exp\left(\hat{\lambda}_n^T g(X_i)\right) g(X_i)\\ &= \frac{1}{S_n(\hat{\lambda}_n)}\sum_{i=1}^n g(X_i) + \frac{1}{S_n(\hat{\lambda}_n)}\left(\sum_{i=1}^n g(X_i)g(X_i)^T\right)\hat{\lambda}_n + \frac{1}{S_n(\hat{\lambda}_n)}\sum_{i=1}^n \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i).\end{aligned}$$

En multipliant deux deux côtés par $\frac{S_n(\hat{\lambda}_n)}{n}$*, on obtient*

$$\begin{aligned}0 &= \mathbb{P}_n g + \Sigma_n\hat{\lambda}_n + \frac{1}{n}\sum_{i=1}^n \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i)\\ \iff -\Sigma_n\hat{\lambda}_n &= \mathbb{P}_n g + \frac{1}{n}\sum_{i=1}^n \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i).\end{aligned}$$

Pour conclure il suffit de montrer que

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i) \right\| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

Par le lemme d'Owen et par la première assertion, on a

$$\exp\left(\rho_n s_i\right) \leq \exp\left(\rho_n |\theta_n^T g(X_i)|\right) \leq \exp\left(\max_{1\leq i\leq n} |\hat{\lambda}_n^T g(X_i)|\right) = 1 + o_p(1) \tag{2.5}$$

$$\frac{1}{n}\sum_{i=1}^{n} \|g(X_i)\|^3 = o_p(\sqrt{n}) \tag{2.6}$$

$$\|\hat{\lambda}_n\|^2 = O_p\left(\frac{1}{n}\right). \tag{2.7}$$

Ainsi

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i) \right\| &\leq \exp\left(\max_{1\leq i\leq n} |\hat{\lambda}_n^T g(X_i)|\right) \|\hat{\lambda}_n\|^2 \frac{1}{n}\sum_{i=1}^{n} \|g(X_i)\|^3 \\ &\leq (1+o_p(1)) O_p\left(\frac{1}{n}\right) o_p(\sqrt{n}) = o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

D'où

$$\hat{\lambda}_n = -\Sigma_n^{-1}\mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right).$$

On peut alors en déduire que

$$\sqrt{n}\hat{\lambda}_n \Rightarrow \mathcal{N}(0, \Sigma^{-1}).$$

□

2.3 Approximation commune de la solution de ces deux problèmes d'optimisation et définition de la mesure empirique informée

Soit $X_1, \cdots, X_n$ un échantillon de taille $n \in \mathbb{N}^*$ *i.i.d.* de loi P inconnue. On suppose qu'on dispose d'une information auxiliaire forte I donnée par une collection finie d'espérance Pg où $g : \mathcal{X} \to \mathbb{R}^m$ est une fonction mesurable. Quitte à poser $h = g - Pg$, on supposera que $Pg = 0$. On note $\mathcal{P}^I$ la sous variété des mesures de probabilités (à poids strictement positifs) à support dans l'échantillon vérifiant l'information auxiliaire I. Dans la section précédente, nous avons étudié séparément les deux problèmes d'optimisation

$$arg \min_{Q\in\mathcal{P}^I} KL(\mathbb{P}_n||Q), \tag{2.8}$$

$$arg \min_{Q\in\mathcal{P}^I} KL(Q||\mathbb{P}_n). \tag{2.9}$$

On suppose que les conditions d'existence et d'unicité de la solution sont satisfaites. Ainsi par le théorème 2.2.2, la solution du **premier problème d'optimisation** est donnée par

$$\forall i \in [\![1,n]\!],\ q_i^{(1)} = \frac{1}{n}\frac{1}{1+\langle \hat{\lambda}_n^{(1)}, g(X_i)\rangle}$$

Par le théorème 2.2.9, la solution du **second problème d'optimisation** est

$$\forall i \in [\![1,n]\!],\ q_i^{(2)} = \frac{\exp\left(\langle \hat{\lambda}_n^{(2)}, g(X_i)\rangle\right)}{S_n(\hat{\lambda}_n^{(2)})},$$
$$S_n(\hat{\lambda}_n^{(2)}) = \sum_{k=1}^{n} \exp\left(\langle \hat{\lambda}_n^{(2)}, g(X_k)\rangle\right).$$

On notera

$$\mathbb{P}_n^{(1)} = \sum_{i=1}^{n} q_i^{(1)} \delta_{X_i}, \tag{2.10}$$
$$\mathbb{P}_n^{(2)} = \sum_{i=1}^{n} q_i^{(2)} \delta_{X_i}. \tag{2.11}$$

On remarque que la solution de chacun de ces deux problèmes respectifs n'admet pas de forme explicite puisque $\hat{\lambda}_n^{(1)}$ et $\hat{\lambda}_n^{(2)}$ s'obtiennent par optimisation. De plus il est difficile de privilégier une solution par rapport à une autre puisqu'on ne sait pas si $\mathscr{P}^I$ est autoparallèle (voir le théorème 1.3.1). Il est donc nécessaire d'obtenir une **approximation commune** de ces poids.

Proposition 2.3.1. *Supposons que $\Sigma = Var_P g$ est inversible. Notons Σ_n la variance empirique de g. Alors pour tout $j \in [\![1,2]\!]$ et pour tout $i \in [\![1,n]\!]$*

$$q_i^{(j)} = p_i + \varepsilon_{i,n}^{(j)}.$$

avec

$$\max_{1\le i\le n} |\varepsilon_{i,n}^{(j)}| = o_p\left(\frac{1}{n}\right),$$
$$p_i = \frac{1}{n}\left(1 - g(X_i)^T \Sigma_n^{-1} \mathbb{P}_n g + (\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g\right).$$

De plus

$$\sum_{i=1}^{n} p_i = 1,$$
$$\sum_{i=1}^{n} p_i g(X_i) = 0.$$

Démonstration.
Traitons tout d'abord le cas $j = 1$. Pour tout $i \in [\![1, n]\!]$

$$q_i^{(1)} = \frac{1}{n} \frac{1}{1 + \langle \hat{\lambda}_n^{(1)}, g(X_i) \rangle}.$$

Par le théorème 2.2.5, on a

$$\max_{1 \le k \le n} |\langle \hat{\lambda}_n^{(1)}, g(X_k) \rangle| = o_p(1),$$
$$\hat{\lambda}_n^{(1)} = \Sigma_n^{-1} \mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Ainsi

$$\begin{aligned} q_i^{(1)} &= \frac{1}{n}\left(1 - \langle \hat{\lambda}_n^{(1)}, g(X_i) \rangle + o_p(1)\right) \\ &= \frac{1}{n}\left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle - \langle o_p\left(\frac{1}{\sqrt{n}}\right), g(X_i) \rangle + o_p(1)\right). \end{aligned}$$

Puisque $\max_{1 \le k \le n} \|g(X_k)\| = o_p(\sqrt{n})$*, on en déduit que*

$$\begin{aligned} q_i^{(1)} &= \frac{1}{n}\left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle + o_p(1)\right) \\ &= \frac{1}{n}\left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle\right) + o_p\left(\frac{1}{n}\right). \end{aligned}$$

De plus puisque $\frac{1}{n}(\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g = o_p\left(\frac{1}{n}\right)$ *alors*

$$q_i^{(1)} = p_i + o_p\left(\frac{1}{n}\right).$$

Traitons le cas $j = 2$. Pour tout $i \in [\![1, n]\!]$

$$q_i^{(2)} = \frac{\exp\left(\langle \hat{\lambda}_n^{(2)}, g(X_i) \rangle\right)}{S_n(\hat{\lambda}_n^{(2)})},$$
$$\text{avec } S_n(\hat{\lambda}_n^{(2)}) = \sum_{k=1}^{n} \exp\left(\langle \hat{\lambda}_n^{(2)}, g(X_k) \rangle\right).$$

Par le théorème 2.2.10, on a

$$\max_{1 \le k \le n} |\langle \hat{\lambda}_n^{(2)}, g(X_k) \rangle| = o_p(1),$$
$$\hat{\lambda}_n^{(2)} = -\Sigma_n^{-1} \mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Ainsi

$$\begin{aligned}
q_i^{(2)} &= \frac{1}{n}\frac{n}{S_n(\hat{\lambda}_n^{(2)})}\exp\left(\langle\hat{\lambda}_n^{(2)}, g(X_i)\rangle\right)\\
&= \frac{1}{n}\frac{n}{S_n(\hat{\lambda}_n^{(2)})}\left(1+\langle\hat{\lambda}_n^{(2)}, g(X_i)\rangle + o_p(1)\right)\\
&= \frac{1}{n}\frac{n}{S_n(\hat{\lambda}_n^{(2)})}\left(1-\langle\Sigma_n^{-1}\mathbb{P}_n g, g(X_i)\rangle + \langle o_p\left(\frac{1}{\sqrt{n}}\right), g(X_i)\rangle + o_p(1)\right)\\
&= \frac{1}{n}\frac{n}{S_n(\hat{\lambda}_n^{(2)})}\left(1-\langle\Sigma_n^{-1}\mathbb{P}_n g, g(X_i)\rangle + o_p(1)\right).
\end{aligned}$$

Montrons que $\frac{n}{S_n(\hat{\lambda}_n^{(2)})} = 1 + o_p(1)$. *En utilisant l'inégalité* $e^x \geq 1 + x$ *pour tout* $x \in \mathbb{R}$, *on a*

$$1 + \langle\hat{\lambda}_n^{(2)}, \mathbb{P}_n g\rangle \leq \frac{S_n(\hat{\lambda}_n^{(2)})}{n} \leq \exp\left(\max_{1\leq k\leq n}|\langle\hat{\lambda}_n^{(2)}, g(X_k)\rangle|\right).$$

Puisque les deux termes d'encadrement tendent en probabilité vers 1, *on en déduit que* $\frac{n}{S_n(\hat{\lambda}_n^{(2)})} = 1 + o_p(1)$. *D'où*

$$\begin{aligned}
q_i^{(2)} &= \frac{1}{n}(1+o_p(1))\left(1-\langle\Sigma_n^{-1}\mathbb{P}_n g, g(X_i)\rangle + o_p(1)\right)\\
&= \frac{1}{n}\left(1-\langle\Sigma_n^{-1}\mathbb{P}_n g, g(X_i)\rangle + o_p(1)\right).
\end{aligned}$$

De même puisque $\frac{1}{n}(\mathbb{P}_n g)^T\Sigma_n^{-1}\mathbb{P}_n g = o_p\left(\frac{1}{n}\right)$ *alors*

$$q_i^{(2)} = p_i + o_p\left(\frac{1}{n}\right).$$

Enfin remarquons que

$$\sum_{i=1}^n p_i = 1 - (\mathbb{P}_n g)^T\Sigma_n^{-1}\mathbb{P}_n g + (\mathbb{P}_n g)^T\Sigma_n^{-1}\mathbb{P}_n g = 1$$

$$\sum_{i=1}^n p_i g(X_i) = \mathbb{P}_n g - \mathbb{P}_n g g^T\Sigma_n^{-1}\mathbb{P}_n g + \mathbb{P}_n g(\mathbb{P}_n g)^T\Sigma_n^{-1}\mathbb{P}_n g = \mathbb{P}_n g - \mathbb{P}_n g = 0$$

car $\Sigma_n = \mathbb{P}_n g g^T - \mathbb{P}_n g(\mathbb{P}_n g)^T$. □

On peut à présent définir la mesure empirique informée.

Definition 2.3.1. La **mesure empirique informée** est définie par

$$\mathbb{P}_n^I = \sum_{i=1}^n p_i\delta_{X_i}$$

où pour tout $i \in [\![1, n]\!]$,

$$p_i = \frac{1}{n}\left(1 - g(X_i)^T\Sigma_n^{-1}\mathbb{P}_n g + (\mathbb{P}_n g)^T\Sigma_n^{-1}\mathbb{P}_n g\right).$$

Remarquons que pour toute fonction mesurable $f : \mathscr{X} \to \mathbb{R}$

$$\mathbb{P}_n^I f = \mathbb{P}_n f - cov_n(g, f)^T \Sigma_n^{-1} \mathbb{P}_n g. \tag{2.12}$$

On retrouve alors l'estimateur adaptatif de la mesure avec information auxiliaire étudiée par *Mickael Albertus* [4]. Cette dernière est un cas particulier des travaux de *Tarima* et *Pavlov* [37].

Remarque 2.3.2. *La mesure avec information auxiliaire étudiée par Mickael Albertus [4] est obtenue en cherchant $\mathbb{P}_n^I f$ de la forme*

$$\mathbb{P}_n^I f = \mathbb{P}_n f + x_*^T \mathbb{P}_n g$$

où $x_* \in \mathbb{R}^m$ est le minimiseur de l'application

$$x \in \mathbb{R}^m \mapsto Var\left(\mathbb{P}_n f + x^T \mathbb{P}_n g\right).$$

La solution à ce problème de minimisation est donnée par $x_*^T = -cov_P(g, f)^T \left(Var_P g\right)^{-1}$. Du fait que P n'est pas connue, x^* n'est pas calculable. Ainsi on utilise l'estimateur empirique de x^* qui est $x_n^T = -cov_n(g, f)^T \left(Var_n g\right)^{-1}$. On en déduit l'estimateur adaptatif de la mesure avec information auxiliaire donné par

$$\mathbb{P}_n^I f = \mathbb{P}_n f + x_n^T \mathbb{P}_n g = \mathbb{P}_n f - cov_n(g, f)^T \Sigma_n^{-1} \mathbb{P}_n g.$$

Exemple 2.3.1. Illustrons graphiquement la répartition des poids sur l'échantillon entre ces quatre mesures $\mathbb{P}_n^{(1)}$, $\mathbb{P}_n^{(2)}$, $\mathbb{P}_n^I$ et $\mathbb{P}_n$. Pour cela, on génère $n = 500$ variables *i.i.d.* de loi $P = \mathcal{N}(0, 1)$ et on injecte l'information auxiliaire I donnée par le moment d'ordre 1 et le moment d'ordre 2. On trace la courbe associée à la répartition des poids de $\mathbb{P}_n^I$, $\mathbb{P}_n^{(1)}$, $\mathbb{P}_n^{(2)}$ et $\mathbb{P}_n$ respectivement. Nous réalisons cette expérience deux fois (graphique de gauche et de droite de la Figure 2.1). On remarque que la répartition des poids est similaire entre ces trois mesures $\mathbb{P}_n^{(1)}$, $\mathbb{P}_n^{(2)}$, $\mathbb{P}_n^I$. De plus, on observe que la répartition des poids est aléatoire. Une étude sur cette répartition est faite à la sous-section 2.4.2.

La proposition suivante établit que la mesure empirique informée $\mathbb{P}_n^I$ est une **mesure de probabilité** presque sûrement à partir d'un certain rang.

Proposition 2.3.3. *Supposons qu'il existe $\varepsilon > 0$ tel que $P\|g\|^{4+\varepsilon} < +\infty$. Alors presque sûrement à partir d'un certain rang*

$$\mathbb{P}_n^I = \sum_{i=1}^n p_i \delta_{X_i}$$

est une mesure de probabilité.

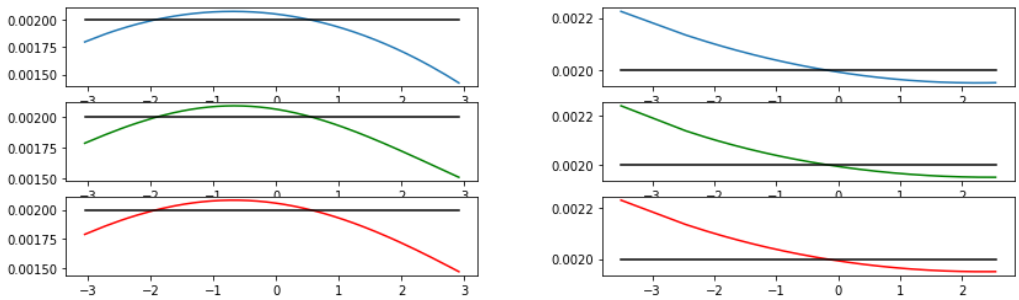

FIGURE 2.1 – Comparaison entre $\mathbb{P}_n^I$ (en bleue), $\mathbb{P}_n^{(1)}$ (en verte), $\mathbb{P}_n^{(2)}$ (en rouge) et $\mathbb{P}_n$ (en noire). Nous réalisons cette expérience deux fois (graphique de gauche et de droite)

Démonstration (de la proposition).
Par la proposition 2.3.1, il suffit de montrer que presque sûrement à partir d'un certain rang $\min_{1\le i\le n} p_i > 0$. *Pour cela nous devons généraliser le lemme d'Owen [31] :*

Lemme 2.3.4. *Soit* $(Y_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires i.i.d. à valeurs dans* $\mathbb{R}_+$ *et* $Z_n = \max_{1\le i\le n} Y_i$. *Si* $\mathbb{E}Y_1^s < +\infty$ *avec* $s > 0$ *alors*

$$Z_n = o_{p.s.}(n^{1/s}),$$
$$\frac{1}{n}\sum_{i=1}^{n} Y_i^{s+1} = o_{p.s.}(n^{1/s}).$$

Démonstration (du lemme).
Puisque $\mathbb{E}Y^s < +\infty$, *on a :*

$$\sum_{n\ge 1} \mathbb{P}(Y_1^s > n) < +\infty.$$

De plus $\mathbb{P}(Y_1^s > n) = \mathbb{P}(Y_n^s > n)$ *car la suite est i.i.d. et* $\mathbb{P}(Y_n^s > n) = \mathbb{P}(Y_n > n^{1/s})$ *car les* Y_i *sont positives, on obtient*

$$\sum_{n\ge 1} \mathbb{P}(Y_n > n^{1/s}) < +\infty.$$

Par le lemme de Borel-Cantelli

$$\mathbb{P}\left(\limsup\{Y_n > n^{1/s}\}\right) = 0.$$

Ainsi p.s. $Y_n > n^{1/s}$ *est vérifié que pour un nombre fini de* n *(ce nombre dépend de* $\omega \in \Omega$*). Ainsi p.s.* $Z_n > n^{1/s}$ *est vérifié que pour un nombre fini de* n. *Le même argument s'applique pour* $Z_n > An^{1/s}$ *pour un* $A > 0$ *quelconque. D'où*

$$\limsup_{n\to+\infty} \frac{Z_n}{n^{1/s}} \le A \ \ p.s..$$

Cette inégalité est vérifiée p.s. pour tout A appartenant à un ensemble dénombrable. Prenons $A_m = \frac{1}{m}$ pour $m \in \mathbb{N}^*$. On a alors p.s.

$$0 \le \liminf_{n\to+\infty} \frac{Z_n}{n^{1/s}} \le \limsup_{n\to+\infty} \frac{Z_n}{n^{1/s}} \le \frac{1}{m}.$$

En faisant $m \to +\infty$, on obtient

$$\lim_{n\to+\infty} \frac{Z_n}{n^{1/s}} := \liminf_{n\to+\infty} \frac{Z_n}{n^{1/s}} = \limsup_{n\to+\infty} \frac{Z_n}{n^{1/s}} = 0.$$

D'où $Z_n = o_{p.s.}(n^{1/s})$. Concernant la seconde assertion, il suffit de remarquer que

$$0 \le \frac{1}{n}\sum_{i=1}^{n} Y_i^{s+1} \le \frac{Z_n}{n}\sum_{i=1}^{n} Y_i^{s} = o_{p.s.}(n^{1/s})$$

par la loi forte des grands nombres. □

Puisque $P\|g\|^{4+\varepsilon} < +\infty$ pour un certain $\varepsilon > 0$, on a par le lemme précédent

$$\frac{\max_{1\le i\le n} \|g(X_i)\|}{n^{\frac{1}{4+\varepsilon}}} \xrightarrow[n\to+\infty]{p.s.} 0.$$

Remarquons que

$$\begin{aligned}
\min_{1\le i\le n} np_i &= 1 - \max_{1\le i\le n} g(X_i)^T \Sigma_n^{-1} \mathbb{P}_n g + (\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g, \\
|\max_{1\le i\le n} g(X_i)^T \Sigma_n^{-1} \mathbb{P}_n g| &\le \max_{1\le i\le n} |g(X_i)^T \Sigma_n^{-1} \mathbb{P}_n g| \\
&\le \max_{1\le i\le n} \|g(X_i)\| \|\Sigma_n^{-1}\| \|\mathbb{P}_n g\|.
\end{aligned}$$

Il suffit de montrer que

$$n^{\frac{1}{4+\varepsilon}} \|\mathbb{P}_n g\| \xrightarrow[n\to+\infty]{p.s.} 0.$$

Pour cela, nous avons besoin d'un théorème de Wellner et Van der Vaart [38] :

Lemme 2.3.5 (W et VdV)**.** *Soit $\mathscr{F}$ une classe P-mesurable ayant une fonction enveloppe mesurable F. Alors pour $p \ge 1$*

$$\|\|\alpha_n\|^*_{\mathscr{F}}\|_{P,p} \lesssim J(1,\mathscr{F})\|F\|_{P,2\vee p}$$

où

$$J(1,\mathscr{F}) = \sup_{Q \text{probabilité discrète}} \int_0^1 \sqrt{1 + \log N\left(\epsilon\|F\|_{Q,2}, \mathscr{F}, L^2(Q)\right)}\, d\epsilon.$$

L'inégalité est vraie à une constante multiplicative près qui dépend seulement de p.

Appliquons ce théorème dans un cadre plus général que sous l'hypothèse $\mathscr{H}$. Plus précisément on l'applique à $\mathscr{F}_j = \{g_j\}$ pour $j \in [\![1, m]\!]$. On souhaite démontrer que pour $s > 4$

$$n^{\frac{1}{s}} \|\mathbb{P}_n g\| \xrightarrow[n\to+\infty]{p.s.} 0.$$

Remarquons que $J(1, \mathscr{F}_j) = 1$. Fixons $\alpha > \frac{2s}{s-2}$ et on suppose que $P\|g\|^\alpha < +\infty$. Notons $\alpha_n = \sqrt{n}(\mathbb{P}_n - P)$. De plus, remarquons que

$$\begin{aligned}
\mathbb{P}(n^{1/s}\|\mathbb{P}_n g\| > \epsilon) &\leq \sum_{j=1}^{m} \mathbb{P}(n^{1/s} \mid \mathbb{P}_n g_j \mid > \epsilon) \\
&= \sum_{j=1}^{m} \mathbb{P}(\mid \alpha_n(g_j) \mid > \frac{\sqrt{n}}{n^{1/s}}\epsilon) \\
&= \sum_{j=1}^{m} \mathbb{P}(\mid \alpha_n(g_j) \mid^\alpha > (\frac{\sqrt{n}}{n^{1/s}}\epsilon)^\alpha) \\
&\leq \sum_{j=1}^{m} \frac{\mathbb{E}(\mid \alpha_n(g_j) \mid^\alpha)}{n^{\alpha(1/2-1/s)}\epsilon^\alpha} \\
&\lesssim \sum_{j=1}^{m} \frac{\|g_j\|^\alpha_{P,2\vee\alpha}}{n^{\alpha\frac{s-2}{2s}}\epsilon^\alpha}.
\end{aligned}$$

Puisque $\alpha\frac{s-2}{2s} > 1$, cette quantité est sommable en n. On peut donc utiliser le lemme de Borel-Cantelli pour conclure que $n^{1/s}\|\mathbb{P}_n g\| \xrightarrow[n\to+\infty]{p.s.} 0$.

Dans notre cadre $s = 4 + \varepsilon$. Posons la fonction définie pour tout $x \in]2, +\infty[$ par

$$\varphi(x) = \frac{2x}{x-2}.$$

Par une étude de fonction, il est possible de montrer que pour tout $x > 4$, $\varphi(x) < x$ et $\varphi(4) = 4$. Puisque $s > 4$, on a le résultat souhaité

$$n^{\frac{1}{4+\varepsilon}} \|\mathbb{P}_n g\| \xrightarrow[n\to+\infty]{p.s.} 0.$$

□

De plus, les poids $(p_i)_{i\in[\![1,n]\!]}$ sont très proches de $\frac{1}{n}$ comme le montre la proposition suivante.

Proposition 2.3.6. *Soit $\alpha \in]0, 1[$ et supposons que $\mathbb{E}\left(\|g(X)\|^{\frac{1}{1-\alpha}}\right) < +\infty$. Alors*

$$n^\alpha \left(\max_{1\leq i\leq n} p_i - \frac{1}{n}\right) \xrightarrow[n\to+\infty]{p.s.} 0.$$

Démonstration.
Tout d'abord remarquons que

$$n^{\alpha}\left(\max_{1\le i\le n} p_i - \frac{1}{n}\right) = \max_{1\le i\le n} n^{\alpha}\left(p_i - \frac{1}{n}\right).$$

Puisque pour tout $i \in [\![1, n]\!]$,

$$p_i = \frac{1}{n}\left(1 - g(X_i)^T \Sigma_n^{-1}\mathbb{P}_n g + (\mathbb{P}_n g)^T \Sigma_n^{-1}\mathbb{P}_n g\right)$$

alors

$$\left|n^{\alpha}\left(p_i - \frac{1}{n}\right)\right| = \frac{1}{n^{1-\alpha}}\left|(\mathbb{P}_n g)^T \Sigma_n^{-1}\mathbb{P}_n g - g(X_i)^T \Sigma_n^{-1}\mathbb{P}_n g\right|.$$

Observons que $(\mathbb{P}_n g)^T \Sigma_n^{-1}\mathbb{P}_n g \xrightarrow[n\to+\infty]{p.s.} 0$ *et par l'inégalité de Cauchy-Schwarz, on a*

$$|\, g(X_i)^T \Sigma_n^{-1}\mathbb{P}_n g \,| \le \|\Sigma_n^{-1}\|\|\mathbb{P}_n g\|\|g(X_i)\| \le \|\Sigma_n^{-1}\|\|\mathbb{P}_n g\| \max_{1\le i\le n}\|g(X_i)\|.$$

Puisque $\mathbb{E}\left(\|g(X)\|^{\frac{1}{1-\alpha}}\right) < +\infty$ *alors par le lemme 2.3.4 on obtient que*

$$\max_{1\le i\le n}\|g(X_i)\| = o_{p.s.}\left(n^{1-\alpha}\right).$$

Ainsi on en déduit que

$$\max_{1\le i\le n}\left|\, n^{\alpha}\left(p_i - \frac{1}{n}\right)\right| \le \frac{1}{n^{1-\alpha}}\,|\,(\mathbb{P}_n g)^T \Sigma_n^{-1}\mathbb{P}_n g\,| + \|\Sigma_n^{-1}\|\|\mathbb{P}_n g\|\frac{\max_{1\le i\le n}\|g(X_i)\|}{n^{1-\alpha}} \xrightarrow[n\to+\infty]{p.s.} 0.$$

□

2.4 Généralisation à une information auxiliaire forte donnée par une fonctionnelle de la mesure P et étude de la répartition des poids sur l'échantillon

2.4.1 Généralisation à une information auxiliaire forte donnée par une fonctionnelle de la mesure P

La généralisation à une information auxiliaire forte donnée par une fonctionnelle de la mesure P a été étudiée par *Mickael Albertus* dans [4]. Dans cette section, nous en donnons une présentation avec quelques modifications en mettant en lumière la structure de mesure. Soit $X_1, \cdots, X_n$ un échantillon de taille $n \in \mathbb{N}^*$ *i.i.d.* à valeurs dans $\mathscr{X}$ et de loi P inconnue. Dans cette section l'information auxiliaire est représentée par la connaissance $g_1(P), \cdots, g_m(P)$ avec $m \in \mathbb{N}^*$. Posons $g = (g_1, \cdots, g_m)^T$. Afin de prendre en compte ce nouveau type d'information auxiliaire, il est nécessaire de supposer quelques conditions de régularité pour l'information auxiliaire $g(P)$.

Hypothèses 2.4.1. Soit $\mathscr{F}$ une classe de fonctions de carré intégrable.

1. Notons $\mathscr{H} = \{g_1, \cdots, g_m\}$ une classe de fonctions définies sur $l^\infty(\mathscr{F})$ à valeurs réelles. Soit $t_0 > 0$. Pour tout $h \in \mathscr{H}$, $h(P)$ est défini et pour tout $Q \in l^\infty(\mathscr{F})$ vérifiant $\|Q\|_{\mathscr{F}} < t_0$,

$$h(P+Q) = h(P) + Q \circ \varphi_h(P) + R_h(Q)$$

   où $\varphi_h : l^\infty(\mathscr{F}) \to Vect(\mathscr{F}_h)$ et $\mathscr{F}_h$ est un ensemble **fini** de fonctions définies sur $\mathscr{X}$ et à valeurs réelles. Puisque $\mathscr{H}$ est finie et quitte à agrandir $\mathscr{F}$, on peut supposer que pour tout $h \in \mathscr{H}$, $\mathscr{F}_h \subset \mathscr{F}$. En munissant $Vect(\mathscr{F}_h)$ de la topologie de la convergence simple et $l^\infty(\mathscr{F})$ muni de sa norme $\|\cdot\|_{\mathscr{F}}$, l'application φ_g est continue. Enfin $R_h : l^\infty(\mathscr{F}) \to \mathbb{R}$ est une application vérifiant sur une boule centrée $\mathscr{B}$ en l'origine, $|R_h(\cdot)| \leq \|\cdot\|_{\mathscr{F}}^q$ pour un certain $q > 1$.
2. Pour tout $h \in \mathscr{H}$, $Var(\varphi_h(P))$ existe et est inversible.

Remarque 2.4.1. *Toute condition faite sur la classe de fonctions $\mathscr{F}$ doit aussi être vérifiée sur $\bigcup_{h\in\mathscr{H}} \mathscr{F}_h$. Notons que puisque $\mathscr{H}$ est finie, $\bigcup_{h\in\mathscr{H}} \mathscr{F}_h$ est une classe Glivenko-Cantelli et Donsker dès qu'elle est incluse respectivement dans $L^1(P)$ et $L^2(P)$.*

La classe de fonctions $\mathscr{H}$ **contient donc l'information auxiliaire**.

Exemple 2.4.1. Donnons quelques **exemples** :

- **Espérance** : Si $g(P) = Pf$ pour une certaine fonction f alors

$$\mathscr{F}_g = \{f\},$$
$$\varphi_g(P) = f.$$

- **Variance** : Si $g(P) = Var_P f$ pour une certaine fonction f alors

$$\mathscr{F}_g = \{f, f^2\},$$
$$\varphi_g(P) = f^2 - 2fPf.$$

- **Covariance** : Si $g(P) = cov_P(f_1, f_2)$ pour deux fonctions intégrables $f_1, f_2 : \mathscr{X} \to \mathbb{R}$ alors

$$\mathscr{F}_g = \{f_1, f_2, f_1 f_2\},$$
$$\varphi_g(P) = f_1 f_2 - f_1 P f_2 - f_2 P f_1.$$

- **Inverse de l'espérance** : Si $g(P) = \frac{1}{Pf}$ pour une certaine fonction f alors

$$\mathscr{F}_g = \{f\},$$
$$\varphi_g(P) = \frac{-f}{(Pf)^2}.$$

- **Espérance conditionnelle** :Si $g(P) = \frac{Pf1_A}{P(A)}$ pour une certaine fonction f et pour $A \subset \mathscr{X}$ alors

$$\mathscr{F}_g = \{1_A, f1_A\},$$
$$\varphi_g(P) = \frac{(f - P(f|A))1_A}{P(A)}.$$

On peut définir la mesure empirique informée en généralisant (2.12) de la manière suivante

$$\mathbb{P}_n^I(f) = \mathbb{P}_n f + \Lambda_n(f) \cdot (g(\mathbb{P}_n) - g(P))$$

où $f : \mathscr{X} \to \mathbb{R}$ est une fonction intégrable et

$$\Lambda_n(f) = -cov_n(\varphi_g(\mathbb{P}_n), f)^T Var_n(\varphi_g(\mathbb{P}_n))^{-1}.$$

Posons $\Sigma_{1,n}(f) = cov_n(\varphi_g(\mathbb{P}_n), f)^T$ et $\Sigma_{2,n} = Var_n(\varphi_g(\mathbb{P}_n))$. Remarquons que $\mathbb{P}_n^I$ est toujours une mesure. En effet

$$\begin{aligned}
\mathbb{P}_n^I(f) &= \mathbb{P}_n f - cov_n(\varphi_g(\mathbb{P}_n), f)^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - g(P)) \\
&= \mathbb{P}_n f - \mathbb{P}_n(f\varphi_g(\mathbb{P}_n))^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - g(P)) + \mathbb{P}_n(\varphi_g(\mathbb{P}_n))^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - g(P))\mathbb{P}_n(f) \\
&= \sum_{i=1}^n \frac{1}{n}\left(1 - \varphi_g(\mathbb{P}_n)(X_i)^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - g(P)) + \mathbb{P}_n(\varphi_g(\mathbb{P}_n))^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - g(P))\right) f(X_i).
\end{aligned}$$

Ainsi on peut en déduire la définition de la mesure empirique informée par $g(P)$.

Definition 2.4.1. Soit I une information auxiliaire donnée par $g(P)$. On appelle **mesure empirique informée** par I la mesure définie par

$$\mathbb{P}_n^I = \sum_{i=1}^n p_i \delta_{X_i}$$

où pour tout $i \in [\![1, n]\!]$

$$p_i = \frac{1}{n}\left(1 - \varphi_g(\mathbb{P}_n)(X_i)^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - g(P)) + \mathbb{P}_n(\varphi_g(\mathbb{P}_n))^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - g(P))\right).$$

Remarque 2.4.2.
- *Remarquons que dans le cas où l'information auxiliaire I est donnée par des espérances, la mesure empirique informée coïncide avec la mesure empirique donnée dans la définition 2.3.1.*
- *Cette mesure est une mesure de probabilité presque sûrement à partir d'un certain rang.*
- *Afin d'exploiter au mieux l'information auxiliaire, nous pourrons utiliser le processus (qui est presque une mesure) défini comme suit,*

$$h(\mathbb{P}_n^{I,*}) = h(\mathbb{P}_n^I)1_{h\notin\{g_1,\cdots,g_m\}} + h(P)1_{h\in\{g_1,\cdots,g_m\}}.$$

où h est une fonction définie sur $l^\infty(\mathscr{F})$.

2.4.2 Étude de la répartition des poids sur l'échantillon

On observe un échantillon $X_1,\cdots,X_n$ de taille $n\in\mathbb{N}^*$ *i.i.d.* de loi P inconnue. On suppose qu'on dispose d'une information auxiliaire forte I donnée par la fonctionnelle $g(P)$. Les poids $(p_i)_{i\in[\![1,n]\!]}$ sont données pour tout $i\in[\![1,n]\!]$,

$$p_i = \frac{1}{n}\left(1-\varphi_g(\mathbb{P}_n)(X_i)^T\Sigma_{2,n}^{-1}(g[\mathbb{P}_n]-g[P]) + \mathbb{P}_n(\varphi_g(\mathbb{P}_n))^T\Sigma_{2,n}^{-1}(g[\mathbb{P}_n]-g[P])\right).$$

On pose

$$\begin{aligned}\psi_n(y) &= \langle A_n, y\rangle + B_n,\\ \varphi_n(x) &= \langle A_n, \varphi_g(\mathbb{P}_n)(x)\rangle + B_n = \psi_n\circ\varphi_g(\mathbb{P}_n)(x)\end{aligned}$$

où

$$\begin{aligned}A_n &= -\frac{1}{n}\Sigma_n^{-1}(g[\mathbb{P}_n]-g[P]),\\ B_n &= \frac{1}{n}\left(1+\langle\Sigma_n^{-1}\mathbb{P}_n\varphi_g(\mathbb{P}_n), g[\mathbb{P}_n]-g[P]\rangle\right).\end{aligned}$$

Par conséquent, on remarque que pour tout $i\in[\![1,n]\!]$,

$$p_i = \varphi_n(X_i) = \psi_n\circ\varphi_g(\mathbb{P}_n)(X_i).$$

Ainsi **les poids sont modifiés par une transformation affine composée par la fonction** $\varphi_g(\mathbb{P}_n)$. Dans le cas où $g(P) = Ph$ pour une certaine fonction h intégrable alors $\varphi_g(P) = h$. Dans ce cas

$$\begin{aligned}A_n &= -\frac{1}{n}\Sigma_n^{-1}(\mathbb{P}_n h - Ph),\\ B_n &= \frac{1}{n}\left(1+\langle\Sigma_n^{-1}\mathbb{P}_n h, \mathbb{P}_n h - Ph\rangle\right),\\ \varphi_n(x) &= \psi_n\circ h(x),\ x\in\mathscr{X}.\end{aligned}$$

Ainsi lorsque h est à valeurs réelles, on peut remarquer que si h est monotone alors les poids associés à l'échantillon ordonné sont aussi ordonnés. La courbe associée est la courbe représentative de la fonction φ_n. De plus, on peut aussi remarquer que p.s. pour n assez grand $B_n > 0$ pour une information de type Ph. Ainsi si $h(x) = x$ alors la courbe représentative est une fonction affine.

Exemple 2.4.2. Faisons quelques simulations afin d'illustrer cela. On génère $n = 500$ variables aléatoires *i.i.d.* de loi $P = \mathcal{N}(0,1)$. On suppose qu'on dispose d'une information auxiliaire forte I. Après avoir ordonné l'échantillon dans l'ordre croissant, on trace la courbe représentative φ_n. On réalise l'expérience deux fois pour chaque h. Dans la Figure 2.2, la fonction est $h(x) = x$ et on observe que la répartition des poids est linéaire mais le sens de variation est aléatoire. Dans la Figure 2.3, la fonction est $h(x) = x^2$ et on remarque la répartition des poids est répartie selon une parabole qui est tournée aléatoirement vers le haut ou vers le bas. Enfin dans la Figure 2.4, la fonction est $h(x) = (x, x^2)$ et on observe que la répartion des poids est donnée selon une portion d'une parabole (transformation affine de h).

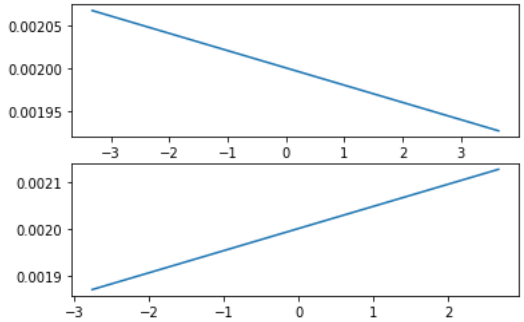

FIGURE 2.2 – On injecte l'information auxiliaire donnée par Ph avec $h(x) = x$ pour tout $x \in \mathbb{R}$

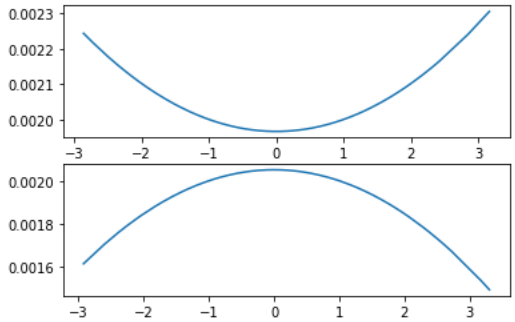

FIGURE 2.3 – On injecte l'information auxiliaire donnée par Ph avec $h(x) = x^2$ pour tout $x \in \mathbb{R}$

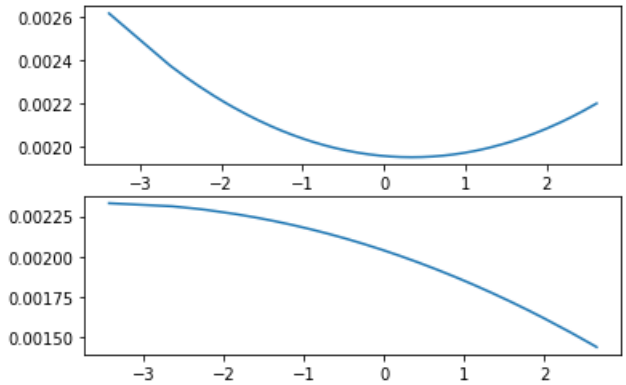

FIGURE 2.4 – On injecte l'information auxiliaire donnée par Ph avec $h(x) = (x, x^2)$ pour tout $x \in \mathbb{R}$

3

Résultats asymptotiques pour l'injection d'information auxiliaire forte

Dans ce chapitre, nous établissons des résultats asymptotiques du type Glivenko-Cantelli et Donsker pour la mesure empirique informée dans le cas d'une information auxiliaire forte. De plus, nous quantifions le gain apporté par l'information auxiliaire en s'intéressant à la diminution de la variance asymptotique.

3.1 Résultats du type Glivenko-Cantelli et Donsker sous des hypothèses minimales dans le cas d'information auxiliaire donnée par des espérances

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite *i.i.d.* à valeurs dans $\mathscr{X}$ et de loi P inconnue. On suppose qu'on dispose d'une information auxiliaire donnée par Pg avec $g = (g_1, \cdots, g_m) : \mathscr{X} \to \mathbb{R}^m$ mesurable. Pour tout $n \in \mathbb{N}^*$, on note $\mathbb{P}_n^I$ la mesure empirique informée (2.3.1) et $\alpha_n^I = \sqrt{n}\left(\mathbb{P}_n^I - P\right)$ le processus empirique informé. Dans cette section, on souhaite obtenir des résultats de type Glivenko-Cantelli (GC) et Donsker sous des hypothèses minimales.

Notons que pour une classe de fonctions $\mathscr{F}$, $\|\mathbb{P}_n^I - P\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} |\mathbb{P}_n f - Pf|$ n'est pas forcément mesurable. Ainsi on peut prendre le majorant mesurable minimal $\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^*$. On pourra consulter à ce propos le chapitre 1 de l'ouvrage [38].

Le résultat suivant énonce un résultat du type **Glivenko-Cantelli pour la mesure empirique informée**.

Théorème 3.1.1. *Soit $\mathscr{F}$ une classe P-Glivenko-Cantelli ayant une enveloppe $F \in L^2(P)$. On dispose d'une information auxiliaire donnée par le vecteur Pg. Alors*

$$\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^* \xrightarrow[n\to+\infty]{p.s.} 0$$

avec $\mathbb{P}_n^I$ définie comme dans la définition 2.3.1.

Démonstration.
Remarquons que pour tout $f \in \mathscr{F}$,

$$\begin{aligned}\|\Lambda_n(f)\| &= \|cov_n(g,f)^T\Sigma_n^{-1}\| \\ &\leq \sqrt{m}\sqrt{Var_n f}\|\Sigma_n^{-1}\| \max_{1\leq i\leq m}\sqrt{Var_n g_i} \\ &\leq \sqrt{m}\|F\|_{L^2(\mathbb{P}_n)}\|\Sigma_n^{-1}\| \max_{1\leq i\leq m}\sqrt{Var_n g_i}.\end{aligned}$$

D'où

$$\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^* \leq \|\mathbb{P}_n - P\|_{\mathscr{F}}^* + \|\mathbb{P}_n g\| \sup_{f\in\mathscr{F}}\|\Lambda_n(f)\| \xrightarrow[n\to+\infty]{p.s.} 0.$$

□

Nous avons aussi un résultat de type **Donsker**.

Théorème 3.1.2. *Soit $\mathscr{F}$ une classe P-Donsker. On suppose qu'il existe une enveloppe $F \in L^2(P)$. Alors*

$$\alpha_n^I \Rightarrow G_I \ \ dans \ \ l^\infty(\mathscr{F})$$

où $G_I(f) = G(f) - cov(g,f)^T\Sigma^{-1}G(g)$ avec $f \in \mathscr{F}$, $\Sigma = Var_P g$ et G est un P-pont brownien standard ayant presque sûrement des trajectoires continues pour la semi-distance $\rho^2(f_1,f_2) = Var_P\left(f_1 - f_2\right)$ pour tous $f_1, f_2 \in L^2(P)$.

Démonstration.
Tout d'abord remarquons que

$$\alpha_n^I(f) = \alpha_n(f) - cov(g,f)^T\Sigma^{-1}\alpha_n(g) + o_p(1).$$

En effet

$$\begin{aligned}\left(cov_n(g,f)^T\Sigma_n^{-1} - cov(g,f)^T\Sigma^{-1}\right)\alpha_n(g) &= \left(cov_n(g,f)^T - cov(g,f)^T\right)\Sigma_n^{-1}\alpha_n(g) \\ &\quad + cov(g,f)^T\left(\Sigma_n^{-1} - \Sigma^{-1}\right)\alpha_n(g).\end{aligned}$$

Le second terme se traite de la manière suivante

$$\begin{aligned}|cov_P(g,f)^T(\Sigma_n^{-1} - \Sigma^{-1})\alpha_n(g)| &\leq \sqrt{m}\sqrt{Var_P f}\max_{1\leq i\leq m}\sqrt{Var_P g_i}\|\Sigma_n^{-1} - \Sigma^{-1}\|\|\alpha_n(g)\| \\ &\leq \sqrt{m}\max_{1\leq i\leq m}\sqrt{Var_P g_i}\|F\|_{L^2(P)}\|\Sigma_n^{-1} - \Sigma^{-1}\|\|\alpha_n(g)\| = o_p(1).\end{aligned}$$

Le premier terme de la manière suivante

$$\begin{aligned}|(cov_n(g,f)^T - cov_P(g,f)^T)\Sigma_n^{-1}\alpha_n(g)| &\leq \|\mathbb{P}_n fg - Pfg - \mathbb{P}_n f\mathbb{P}_n g\|\|\Sigma_n^{-1}\|\|\alpha_n(g)\| \\ &\leq \sqrt{m}\max_{1\leq j\leq m}\|\mathbb{P}_n - P\|_{\mathscr{F}_j}^*\|\Sigma_n^{-1}\|\|\alpha_n(g)\| \\ &\quad + |\mathbb{P}_n F|\|\mathbb{P}_n g\|\|\Sigma_n^{-1}\|\|\alpha_n(g)\|\end{aligned}$$

où pour tout $j \in [\![1,m]\!]$, $\mathscr{F}_j = \{g_j f, f \in \mathscr{F}\}$. Montrons que pour tout $j \in [\![1,m]\!]$, $\mathscr{F}_j$ est P-Glivenko-Cantelli. Fixons $j \in [\![1,m]\!]$ quelconque. Puisque g est intégrable sous P alors $\{g_j\}$ est P-Glivenko-Cantelli. De plus par les hypothèses faites sur $\mathscr{F}$, on en déduit que $\mathscr{F}$ est P-Donsker et donc P-Glivenko-Cantelli. Posons $\phi(x,y) = xy$ avec $x, y \in \mathbb{R}$. Remarquons que ϕ est continue et $\mathscr{F}_j = \phi(\{g_j\}, \mathscr{F})$. On conserve donc le caractère P-GC pour $\mathscr{F}_j$. D'où puisque j est quelconque,

$$\max_{1\le j\le m} \|\mathbb{P}_n - P\|^*_{\mathscr{F}_j} = o_{p.s.}(1).$$

Ainsi

$$\left(\sup_{f\in\mathscr{F}} |(cov_n(g,f)^T\Sigma_n^{-1} - cov(g,f)^T\Sigma^{-1})\alpha_n(g)|\right)^* = o_p(1).$$

Posons pour tout $f \in \mathscr{F}$,

$$W_n(f) = \alpha_n(f) - cov(g,f)^T\Sigma^{-1}\alpha_n(g)$$

et montrons qu'il converge en loi vers le processus souhaité. Tout d'abord par le théorème central limite, on a

$$\forall f \in \mathscr{F}\ ,\ \begin{pmatrix}\alpha_n(f)\\ \alpha_n(g)\end{pmatrix} \Rightarrow \begin{pmatrix}G(f)\\ G(g)\end{pmatrix}.$$

En posant l'application continue $\psi_f(a,b) = (a, cov(f,g)^T\Sigma^{-1}b)$, on a

$$\forall f \in \mathscr{F}\ ,\ \tilde{\mathbb{Y}}_n(f) = \begin{pmatrix}\alpha_n(f)\\ cov(g,f)^T\Sigma^{-1}\alpha_n(g)\end{pmatrix} \Rightarrow \tilde{\mathbb{Y}}(f) = \begin{pmatrix}G(f)\\ cov(g,f)^T\Sigma^{-1}G(g)\end{pmatrix}$$

En fait pour tout $k \in \mathbb{N}^*$ on a la convergence multivariée pour $f = (f_1, ..., f_k) \in \mathscr{F}^k$ en posant l'application continue

$$\tilde{\psi}_f(a_1, ..., a_{k+1}) = \left(\psi_{f_1}(a_1, a_{k+1}), ..., \psi_{f_k}(a_k, a_{k+1})\right)$$

et en utilisant le théorème central limite multivarié pour $(f_1, ..., f_k, g)$. Finalement on peut en déduire que

$$\left(W_n(f_1), \cdots, W_n(f_k)\right)^T \Rightarrow \left(G(f_1) - cov_P(g,f_1)^T\Sigma^{-1}G(g), \cdots, G(f_k) - cov_P(g,f_k)^T\Sigma^{-1}G(g)\right)^T.$$

Montrons que $W_n \Rightarrow G_I$ dans $l^\infty(\mathscr{F})$. Pour cela, on va appliquer le lemme technique suivant.

Lemme 3.1.3. *Soit $X_n : \Omega_n \to l^\infty(T)$ une suite d'applications. Alors ces deux assertions sont équivalentes :*

- *On a*
  1. *Pour tout $(t_1, ..., t_k) \in T^k$, $(X_n(t_1), ..., X_n(t_k))$ converges en loi vers un vecteur aléatoire dans $\mathbb{R}^k$ pour tout $k \in \mathbb{N}^*$.*
  2. *Il existe une semi-distance ρ telle que (T,ρ) est précompact et pour tout $\epsilon > 0$,*

$$\lim_{\delta\to 0}\limsup_n \mathbb{P}^*\left(\sup_{\rho(s,t)<\delta} |X_n(s) - X_n(t)| > \epsilon\right) = 0.$$

- *Il existe $X : \Omega \to l^\infty(T)$ un processus stochastique mesurable et tendu tel que*

$$X_n \Rightarrow X \ \ dans\ l^\infty(T).$$

Tout d'abord vérifions que p.s. $W_n \in l^\infty(\mathscr{F})$,

$$\begin{aligned}\|W_n\|_{\mathscr{F}} &\leq 2\max\left(\|\alpha_n\|_{\mathscr{F}}, \sup_{f\in\mathscr{F}} |(Pfg)^T\Sigma^{-1}\alpha_n(g)|\right)\\ &\leq 2\max\left(\|\alpha_n\|_{\mathscr{F}}, \sqrt{m}\max_{1\leq i\leq m}\sqrt{Var_P g_i}\|F\|_{L^2(P)}\|\Sigma^{-1}\|\|\alpha_n(g)\|\right) < +\infty.\end{aligned}$$

Par ce qui précède, le point 1. est vérifié (convergence fini-dimensionnelles). Il nous manque à vérifier le point 2. Pour cela, prenons la semi-distance de $L^2(P)$, $\rho^2(f_1, f_2) = Var_P\left(f_1 - f_2\right)$ pour tout $f_1, f_2 \in L^2(P)$. Puisque $\mathscr{F}$ est P-Donsker, $(\mathscr{F}, \rho)$ est précompact. Calculons maintenant

$$\begin{aligned}&\limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\ W_n(f_1) - W_n(f_2)\ | > 2\epsilon\right)\\ &\leq \limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} \max\left(|\alpha_n(f_1) - \alpha_n(f_2)|, |(cov_P(f_1 - f_2)g)^T\Sigma^{-1}\alpha_n(g))|\right) > \epsilon\right)\\ &\leq \limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1) - \alpha_n(f_2)| > \epsilon\right)\\ &+ \limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} \sqrt{m}\max_{1\leq i\leq m}\sqrt{Var_P g_i}\ \rho(f_1, f_2)\|\Sigma^{-1}\|\|\alpha_n(g)\| > \epsilon\right)\\ &\leq \limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1) - \alpha_n(f_2)| > \epsilon\right)\\ &+ \limsup_n \mathbb{P}^*\left(\delta\sqrt{m}\max_{1\leq i\leq m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|\alpha_n(g)\| > \epsilon\right).\end{aligned}$$

Concernant le premier terme, puisque $\mathscr{F}$ est P-Donsker alors

$$\lim_{\delta\to 0}\limsup_n \mathbb{P}^*(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1) - \alpha_n(f_2)| > \epsilon) = 0.$$

Concernant le second terme, en utilisant le fait que

$$\|\alpha_n(g)\| \Rightarrow \|G(g)\|$$

on a par le théorème du portemanteau

$$\begin{aligned}&\limsup_n \mathbb{P}^*\left(\delta\sqrt{m}\max_{1\leq i\leq m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|\alpha_n(g) > \epsilon\right)\\ &\leq \limsup_n \mathbb{P}^*\left(\delta\sqrt{m}\max_{1\leq i\leq m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|\alpha_n(g) \geq \epsilon\right)\\ &\leq \mathbb{P}^*\left(\delta\sqrt{m}\max_{1\leq i\leq m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|G(g)\| \geq \epsilon\right).\end{aligned}$$

Puisque $\delta\sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|G(g)\| = o_p(1)$, *on en déduit que*

$$\lim_{\delta\to 0}\limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\, W_n(f_1) - W_n(f_2)\,| > 2\epsilon\right) = 0.$$

Ainsi $W_n \Rightarrow G_I$ *dans* $l^\infty(\mathscr{F})$. □

Enfin énonçons une proposition montrant que **la variance du processus limite** G_I **est plus petite que celle de** G.

Proposition 3.1.4. *Soit* $\mathscr{F}$ *une classe* P*-Donsker. Alors pour tout* $f \in \mathscr{F}$, *la variance du processus limite est*

$$Var(G_I(f)) = Var_P f - cov_P(g,f)^T\Sigma^{-1}cov_P(g,f)$$

et $Var(G_I(f)) \le Var(G(f))$. *De plus,* $Var(G_I(f)) < Var(G(f))$ *dès que* $cov_P(g,f) \neq 0$.

Démonstration.
Calculons la variance de $G_I(f)$ *pour* $f \in \mathscr{F}$,

$$\begin{aligned}
Var(G_I(f)) &= Var(G(f) - cov_P(g,f)^T\Sigma^{-1}G(g)) \\
&= Var(G(f)) + Var\left(cov_P(g,f)^T\Sigma^{-1}G(g)\right) - 2cov\left(G(f), cov_P(g,f)^T\Sigma^{-1}G(g)\right) \\
&= Var_P f + cov_P(g,f)^T\Sigma^{-1}Var_P g\Sigma^{-1}cov_P(g,f) - 2cov_P(g,f)^T\Sigma^{-1}cov_P(g,f) \\
&= Var_P f - cov_P(g,f)^T\Sigma^{-1}cov_P(g,f).
\end{aligned}$$

Puisque $\Sigma^{-1} > 0$, *on en déduit que* $cov_P(g,f)^T\Sigma^{-1}cov_P(g,f) \ge 0$. *Ainsi*

$$Var(G_I(f)) \le Var(G(f)).$$

□

Lorsqu'on injecte plusieurs informations auxiliaires décorrélées I, la variance limite du processus limite informé par I peut s'écrire en fonction des variances limites des processus informés par chacune de ces informations auxiliaires. Plus précisément, supposons que nous observons $n \in \mathbb{N}^*$ réalisations de $X = (Z_1, \cdots, Z_m)$ avec $m \in \mathbb{N}^*$ et pour tout $k \in [\![1,m]\!]$, Z_k est une variable aléatoire. Pour tout $k \in [\![1,n]\!]$, les variables Z_k sont à valeurs dans $\mathscr{Z}_k \subset \mathbb{R}^{m_k}$ avec $m_k \in \mathbb{N}^*$. Notons $P = \mathbb{P}^X$ la loi de X et $P_k = \mathbb{P}^{Z_k}$ la loi de Z_k pour tout $k \in [\![1,n]\!]$. De plus, on dispose d'une information auxiliaire sur chacune de ces m variables notée I_k pour tout $k \in [\![1,n]\!]$. On note I l'information auxiliaire donnée par $I_1, \cdots, I_m$. Pour chaque $k \in [\![1,n]\!]$, l'information auxiliaire I_k est donnée par $P_k\widetilde{g}_k = \mathbb{E}(\widetilde{g}_k(Z_k))$. On suppose que $\widetilde{g}_1(Z_1), \cdots, \widetilde{g}_m(Z_m)$ sont décorrélées deux à deux. Posons pour tout $k \in [\![1,m]\!]$, $g_k(x) := \widetilde{g}_k(x_k)$ avec $x = (x_1, \cdots, x_m) \in \mathscr{Z}_1 \times \cdots \times \mathscr{Z}_m$ et $g = (g_1, \cdots, g_m)$. Fixons $f \in L^2(P)$. Par le théorème 3.1.2, la variance du pont brownien informé par I est

$$Var\left(G_I(f)\right) = Var_P f - cov_P(f,g)\left(Var_P g\right)^{-1} cov_P(g,f).$$

Puisque $\widetilde{g}_1(Z_1), \cdots, \widetilde{g}_m(Z_m)$ sont décorrélées deux à deux, la matrice de covariance est

$$Var_P g = \begin{pmatrix} Var_P g_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Var_P g_m \end{pmatrix}.$$

Ainsi la matrice inverse de la covariance est

$$Var_P g = \begin{pmatrix} \left(Var_P g_1\right)^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left(Var_P g_m\right)^{-1} \end{pmatrix}.$$

Ainsi on en déduit que

$$\begin{aligned} Var\left(G_I(f)\right) &= Var_P f - \sum_{k=1}^{m} cov_P(f, g_k)\left(Var_P g_k\right)^{-1} cov_P(g_k, f) \\ &= \sum_{k=1}^{m} Var_P\left(G_{I_k}(f)\right) - (m-1) Var_P f. \end{aligned}$$

3.2 Extension des résultats asymptotiques à des fonctionnelles

Dans cette section, on généralise les résultats asymptotiques de la section précédente à des fonctionnelles de la mesure P de type $g_1(P), \cdots, g_m(P)$ avec $m \in \mathbb{N}^*$. On souhaite obtenir des résultats asymptotiques sous des hypothèses minimales. Les démonstrations de ces résultats sont similaires à ceux de la section précédente. Rappelons les notations principales de la section 2.4,

$$\begin{aligned} &\mathbb{P}_n^I(f) = \mathbb{P}_n f + \Lambda_n(f) \cdot (g(\mathbb{P}_n) - g(P)), \\ &\alpha_n^I(f) = \sqrt{n}(\mathbb{P}_n f - Pf) + \Lambda_n(f) \cdot \sqrt{n}(g(\mathbb{P}_n) - g(P)) \\ &\Lambda_n(f) = -\Sigma_{1,n}(f)\Sigma_{2,n}^{-1} = -cov_n(\varphi_g(\mathbb{P}_n), f)^T Var_n(\varphi_g[\mathbb{P}_n])^{-1} \end{aligned}$$

où f est un fonction de carré intégrable. Énonçons un lemme technique portant sur Λ_n.

Lemme 3.2.1. *Soit $\mathscr{F}$ une classe P-Glivenko-Cantelli ayant une enveloppe $F \in L^2(P)$. On a*

i) $\Sigma_{1,n} \xrightarrow[n\to+\infty]{p.s., \|\cdot\|_{\mathscr{F}}} \Sigma_1 : f \mapsto cov_P(\varphi_g(P), f)^T$.

ii) $\Sigma_{2,n}^{-1} \xrightarrow[n\to+\infty]{p.s.} \Sigma_2^{-1} = Var(\varphi_g(P))^{-1}$.

iii) $\Lambda_n \xrightarrow[n\to+\infty]{p.s., \|\cdot\|_{\mathscr{F}}} \Lambda = -\Sigma_1(\cdot)\Sigma_2^{-1}$.

Démonstration.
Tout d'abord puisque $\mathscr{F}$ est une classe P-Glivenko-Cantelli, alors $\varphi_g(\mathbb{P}_n) \xrightarrow[n\to+\infty]{p.s.} \varphi_g(P)$. Par définition

de l'application φ_g, on a

$$\varphi_g(P) = \sum_{i=1}^{k} \lambda_i f_i,$$
$$\varphi_g(\mathbb{P}_n) = \sum_{i=1}^{k} \lambda_{i,n} f_i,$$
$$\lambda_{i,n} \xrightarrow[n\to+\infty]{p.s.} \lambda_i.$$

Commençons par i). *Soit $f \in \mathscr{F}$, alors*

$$\begin{aligned}
&\|\Sigma_{1,n}(f) - \Sigma_1(f)\| \\
&\leq \|\sum_{i=1}^{k} \lambda_{i,n} cov_n(f_i, f) - \sum_{i=1}^{k} \lambda_i cov(f_i, f)\| \\
&\leq \|\sum_{i=1}^{k} (\lambda_{i,n} - \lambda_i) cov_n(f_i, f)\| + \|\sum_{i=1}^{k} \lambda_i (cov_n(f_i, f) - cov(f_i, f))\| \\
&\leq k\|F\|_{L^2(P)} \max_{1\leq i\leq k} \|\lambda_{i,n} - \lambda_i\| + \sum_{i=1}^{k} \|\lambda_i\| \max_{1\leq i\leq k} |cov_n(f_i, f) - cov(f_i, f)|.
\end{aligned}$$

Par ce qui précède, le premier terme tend p.s. vers 0. Concernant le second terme, il se traite de la manière suivante

$$\begin{aligned}
\max_{1\leq i\leq k} |(cov_n(f_i, f) - cov(f_i, f)| &\leq \max_{1\leq i\leq k} |\mathbb{P}_n f f_i - P f f_i - \mathbb{P}_n f(\mathbb{P}_n f_i - P f_i| \\
&\leq \max_{1\leq i\leq k} \|\mathbb{P}_n - P\|_{\mathscr{F}_i} + |\mathbb{P}_n F| \max_{1\leq i\leq k} |\mathbb{P}_n f_i - P f_i|
\end{aligned}$$

où $\mathscr{F}_i = \{f_i f, f \in \mathscr{F}\}$ avec $i \in [\![1, k]\!]$. Montrons que pour tout $j \in [\![1, k]\!]$, $\mathscr{F}_j$ est P-Glivenko-Cantelli. Fixons $j \in [\![1, k]\!]$. Puisque f_j est intégrable sous P alors $\{f_j\}$ est P-GC. De plus $\mathscr{F}$ est P-Glivenko-Cantelli. Posons $\phi(x, y) = xy$ avec $x, y \in \mathbb{R}$. Remarquons que l'application ϕ est continue et $\mathscr{F}_j = \phi(\{g_j\}, \mathscr{F})$. On conserve donc le caractère P-Glivenko-Cantelli pour $\mathscr{F}_j$. D'où

$$\max_{1\leq j\leq m} \|\mathbb{P}_n - P\|^*_{\mathscr{F}_j} = o_{p.s.}(1).$$

Il est clair que le second terme tend p.s. vers 0. *Ainsi*

$$\left(\sup_{f\in\mathscr{F}} \max_{1\leq i\leq k} |(cov_n(f_i, f) - cov(f_i, f)|\right)^* = o_{p.s.}(1).$$

D'où

$$\left(\sup_{f\in\mathscr{F}} \|\Sigma_{1,n}(f) - \Sigma_1(f)\|\right)^* = o_{p.s.}(1).$$

Concernant la seconde assertion

$$\begin{aligned}
\|Var_n(\varphi_g(\mathbb{P}_n)) - Var(\varphi_g(P))\| &= \|cov_n(\varphi_g(\mathbb{P}_n), \varphi_g(\mathbb{P}_n)) - cov(\varphi_g(P), \varphi_g(P))\| \\
&= \|\sum_{1\leq i,j\leq k} \lambda_{i,n} \lambda_{j,n}^T cov_n(f_i, f_j) - \sum_{1\leq i,j\leq k} \lambda_i \lambda_j^T cov(f_i, f_j)\|.
\end{aligned}$$

En faisant le même découpage que précédemment et en utilisant le fait que $\|\lambda_{i,n}\lambda_{j,n}^T - \lambda_i\lambda_j^T\| \xrightarrow[n\to+\infty]{p.s.} 0$, *on obtient*

$$\|Var_n(\varphi_g(\mathbb{P}_n)) - Var(\varphi_g(P))\| \xrightarrow[n\to+\infty]{p.s.} 0.$$

Enfin en utilisant le fait que $Var(\varphi_g(P))$ *est inversible et que l'application déterminant est continue, on en déduit que p.s. à partir d'un certain rang* $Var_n(\varphi_g(\mathbb{P}_n))$ *est inversible. Puisque l'application* $A \mapsto A^{-1}$ *est continue dans* $GL_m(\mathbb{R})$, *alors on obtient le résultat souhaité à savoir que*

$$\Sigma_{2,n}^{-1} = Var_n(\varphi_g(\mathbb{P}_n)) \xrightarrow[n\to+\infty]{p.s.} \Sigma_2^{-1} = Var(\varphi_g(P))^{-1}.$$

Concernant la troisième assertion, elle se déduit des deux précédentes par découpage. En effet

$$\begin{aligned}\sup_{f\in\mathscr{F}} \|\Lambda_n(f) - \Lambda(f)\| &\leq \sup_{f\in\mathscr{F}} \|\Sigma_{1,n}(f) - \Sigma_1(f)\| \|\Sigma_{2,n}^{-1}\| + \|\Sigma_{2,n}^{-1} - \Sigma_2^{-1}\| \sup_{f\in\mathscr{F}} \|\Sigma_1(f)\| \\ &\xrightarrow[n\to+\infty]{p.s.} 0.\end{aligned}$$

□

Énonçons un résultat du type **Glivenko-Cantelli**.

Théorème 3.2.2. *Soit* $\mathscr{F}$ *une classe* P*-Glivenko-Cantelli ayant une enveloppe* $F \in L^2(P)$. *On dispose d'une information auxiliaire donnée par* $g(P)$. *Alors*

$$\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^* \xrightarrow[n\to+\infty]{p.s.} 0$$

avec $\mathbb{P}_n^I$ *définie comme dans la définition 2.4.1.*

Démonstration.
Remarquons que pour tout $f \in \mathscr{F}$,

$$\begin{aligned}\|\Lambda_n(f)\| &= \|cov_n(\varphi_g(\mathbb{P}_n), f)^T \Sigma_{2,n}^{-1}\| \\ &\leq \sqrt{m}\sqrt{Var_n f}\|\Sigma_{2,n}^{-1}\| \max_{1\leq i\leq m} \sqrt{Var_n \varphi_{g_i}(\mathbb{P}_n)} \\ &\leq \sqrt{m}\|F\|_{L^2(\mathbb{P}_n)}\|\Sigma_{2,n}^{-1}\| \max_{1\leq i\leq m} \sqrt{Var_n \varphi_{g_i}(\mathbb{P}_n)}.\end{aligned}$$

De plus, observons que

$$\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^* \leq \|\mathbb{P}_n - P\|_{\mathscr{F}}^* + \|g(\mathbb{P}_n) - g(P)\| \sup_{f\in\mathscr{F}} \|\Lambda_n(f)\|.$$

Par le lemme technique 3.2.1, on a

$$\|\Sigma_{2,n}^{-1}\| \max_{1\leq i\leq m} \sqrt{Var_n \varphi_{g_i}(\mathbb{P}_n)} \xrightarrow[n\to+\infty]{p.s.} \|\Sigma^{-1}\| \max_{1\leq i\leq m} \sqrt{Var_P \varphi_{g_i}(P)}).$$

Ainsi il suffit de montrer que $g(\mathbb{P}_n) \xrightarrow{p.s.} g(P)$. *En effet pour tout* $i \in [\![1, m]\!]$

$$g_i(\mathbb{P}_n) = g_i(P + \mathbb{P}_n - P) = g_i(P) + (\mathbb{P}_n - P) \circ \varphi_{g_i}(P) + R_{g_i}(\mathbb{P}_n - P).$$

Puisque $\varphi_{g_i}(P) \in L^1(P)$, *on en déduit que* $(\mathbb{P}_n - P) \circ \varphi_{g_i}(P)$ *tend p.s. vers* 0 *lorsque* $n \to +\infty$. *De plus il existe* $q > 1$ *tel que*

$$R_{g_i}(\mathbb{P}_n - P) \leq (\|\mathbb{P}_n - P\|^*_{\mathscr{F}})^q \xrightarrow[n \to +\infty]{p.s.} 0.$$

□

Le théorème suivant est un résultat du type **Donsker** .

Théorème 3.2.3. *Soit* $\mathscr{F}$ *une classe* P*-Donsker. On suppose qu'il existe une enveloppe* $F \in L^2(P)$. *Alors*

$$\alpha_n^I \Rightarrow G_I \ \ dans \ \ l^\infty(\mathscr{F})$$

où $G_I(f) = G(f) - cov(\varphi_g(P), f)^T \Sigma_2^{-1} G(\varphi_g(P))$ *et* G *est un* P*-pont brownien standard ayant presque sûrement des trajectoires continues pour la semi-distance* $\rho^2(f_1, f_2) = Var_P\left(f_1 - f_2\right)$ *pour tous* $f_1, f_2 \in L^2(P)$.

Démonstration.
Fixons $f \in \mathscr{F}$. *Tout d'abord remarquons que par le lemme technique*

$$\alpha_n^I(f) = \alpha_n(f) + \Lambda_n(f)\sqrt{n}(g(\mathbb{P}_n) - g(P)) + o_p(1).$$

De plus, puisque

$$\sqrt{n}(g(\mathbb{P}_n) - g(P)) = \alpha_n(\varphi_g(P)) + \sqrt{n} R_g(\mathbb{P}_n - P)$$
$$\sqrt{n}\|R_g(\mathbb{P}_n - P)\|_\infty \leq \sqrt{n}\|\mathbb{P}_n - P\|^q_{\mathscr{F}} \xrightarrow[n \to +\infty]{p.s.} 0$$

on a

$$\alpha_n^I(f) = \alpha_n(f) + \Lambda_n(f)\alpha_n(\varphi_g(P)) + o_p(1).$$

De la même manière que dans la démonstration du théorème 3.1.2, on a

$$\alpha_n^I(f) = \alpha_n(f) + \Lambda(f)\alpha_n(\varphi_g(P)) + o_p(1).$$

Posons pour tout $f \in \mathscr{F}$

$$W_n(f) = \alpha_n(f) - cov(\varphi_g(P), f)^T \Sigma_2^{-1} \alpha_n(\varphi_g(P))$$

et montrons qu'il converge en loi vers le processus souhaité. Par le théorème central limite multivarié et en utilisant la stabilité de la convergence en loi par une application continue, on a de manière similaire que dans la démonstration du théorème 3.1.2

$$\left(W_n(f_1), \cdots, W_n(f_k)\right)^T \Rightarrow \left(G_I(f_1), \cdots, G_I(f_k)\right)^T.$$

pour tout $k \in \mathbb{N}^*$, $f_1, \cdots, f_k \in \mathscr{F}$ *et pour tout* $j \in [\![1, k]\!]$ *:*

$$G_I(f_j) = G_I(f_j) - cov_P(\varphi_g(P), f_j)^T \Sigma^{-1} G(\varphi_g(P)).$$

Montrons que $W_n \Rightarrow G_I$ *dans* $l^\infty(\mathscr{F})$. *Pour cela, on va appliquer le lemme technique suivant.*

Lemme 3.2.4. *Soit* $X_n : \Omega_n \to l^\infty(T)$ *une suite d'applications. Alors ces deux assertions sont équivalentes*

- *On a*
  1. *Pour tout* $(t_1, ..., t_k) \in T^k$, $(X_n(t_1), ..., X_n(t_k))$ *converges en loi vers un vecteur aléatoire dans* $\mathbb{R}^k$ *pour tout* $k \in \mathbb{N}^*$.
  2. *Il existe une semi-distance* ρ *telle que* (T, ρ) *is précompact et pour tout* $\epsilon > 0$ *:*

$$\lim_{\delta \to 0} \limsup_n \mathbb{P}^*(\sup_{\rho(s,t)<\delta} |X_n(s) - X_n(t)| > \epsilon) = 0.$$

- *Il existe* $X : \Omega \to l^\infty(T)$ *un processus stochastique mesurable et tendu tel que :*

$$X_n \Rightarrow X \ \text{ dans } l^\infty(T).$$

Tout d'abord vérifions que p.s. $\tilde{\mathbb{Y}}_n \in l^\infty(\mathscr{F})$

$$\begin{aligned} p.s.\ \|\tilde{\mathbb{Y}}_n\| &= \max\left(\|\alpha_n\|_{\mathscr{F}}, \sup_{f \in \mathscr{F}} |cov_P(\varphi_g(P), f)^T \Sigma_2^{-1} \alpha_n(\varphi_g(P))| \right) \\ &\leq \max\left(\|\alpha_n\|_{\mathscr{F}}, \sqrt{m} \|F\|_{L^2(P)} \max_{1 \leq i \leq m} \sqrt{Var_P \varphi_{g_i}(P)} \|\Sigma_2^{-1}\| \|\alpha_n(\varphi_g(P))\| \right) \\ &< +\infty. \end{aligned}$$

Par ce qui précède, le point 1. est vérifié (convergence fini-dimensionnelles). Il nous manque à vérifier le point 2. Pour cela, prenons la distance de $L^2(P)$, $\rho(f_1, f_2) = Var_P\left(f_1 - f_2\right)$ *for all* $f_1, f_2 \in L^2(P)$.

Puisque $\mathscr{F}$ est P-Donsker, $(\mathscr{F},\rho)$ est précompact. Calculons maintenant

$$\begin{aligned}
&\limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |W_n(f_1)-W_n(f_2)|>\epsilon\right)\\
&=\mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} \max\left(|\alpha_n(f_1)-\alpha_n(f_2)|, |cov_P(\varphi_g(P), f_1-f_2)^T\Sigma_2^{-1}\alpha_n(\varphi_g(P))|\right)>\epsilon\right)\\
&\leq \limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1)-\alpha_n(f_2)|>\epsilon\right)\\
&+\limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} \sqrt{m}\rho(f_1,f_2)\max_{1\leq i\leq m}\sqrt{Var_P\varphi_{g_i}(P)}\|\Sigma_2^{-1}\|\|\alpha_n(\varphi_g(P))\|>\epsilon\right)\\
&\leq \limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1)-\alpha_n(f_2)|>\epsilon\right)\\
&+\limsup_n \mathbb{P}^*\left(\sqrt{m}\delta\max_{1\leq i\leq m}\sqrt{Var_P\varphi_{g_i}(P)}\|\Sigma_2^{-1}\|\|\alpha_n(\varphi_g(P))\|>\epsilon\right).
\end{aligned}$$

Concernant le premier terme, puisque $\mathscr{F}$ est P-Donsker

$$\lim_{\delta\to 0}\limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1)-\alpha_n(f_2)|>\epsilon\right)=0.$$

Concernant le second terme, en utilisant le fait que

$$\|\alpha_n(\varphi_g(P))\|\Rightarrow\|G(\varphi_g(P))\|$$

on a

$$\begin{aligned}
&\limsup_n \mathbb{P}^*\left(\sqrt{m}\delta\max_{1\leq i\leq m}\sqrt{Var_P\varphi_{g_i}(P)}\|\Sigma_2^{-1}\|\|\alpha_n(\varphi_g(P))\|>\epsilon\right)\\
&\leq \limsup_n \mathbb{P}^*\left(\sqrt{m}\delta\max_{1\leq i\leq m}\sqrt{Var_P\varphi_{g_i}(P)}\|\Sigma_2^{-1}\|\|\alpha_n(\varphi_g(P)))\|\geq\epsilon\right)\\
&\leq \mathbb{P}^*\left(\sqrt{m}\delta\max_{1\leq i\leq m}\sqrt{Var_P\varphi_{g_i}(P)}\|\Sigma_2^{-1}\|\|G(\varphi_g(P))\|\geq\epsilon\right)
\end{aligned}$$

par le lemme de porte-manteau. Puisque

$$\sqrt{m}\delta\max_{1\leq i\leq m}\sqrt{Var_P\varphi_{g_i}(P)}\|\Sigma_2^{-1}\|\|G(\varphi_g(P)))\xrightarrow[\delta\to 0]{proba}0.$$

on en déduit que

$$\lim_{\delta\to 0}\limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |W_n(f_1)-W_n(f_2)|>\epsilon\right)=0.$$

Ainsi $W_n\Rightarrow G_I$ dans $l^\infty(\mathscr{F})$. $\square$

4

Injection d'information auxiliaire faible et résultats asymptotiques

Dans ce chapitre, nous souhaitons injecter une information auxiliaire plus affaiblie et de différentes natures. Ces dernières peuvent être des estimations d'un paramètre issues de sources indépendantes, des ajouts de données indépendantes, un choix entre plusieurs valeurs avec une mesure de préférence associée. Nous étudions chacune de ces informations auxiliaires et nous établissons des résultats asymptotiques pour ces mesures empiriques informées.

4.1 Information auxiliaire faible issue de sources indépendantes à l'échantillon initial

4.1.1 Information auxiliaire faible donnée par des estimations issues de sources indépendantes

On observe un échantillon $X_1, \cdots, X_n$ *i.i.d.* de taille $n \in \mathbb{N}^*$ et de loi P. On suppose qu'on a accès à des estimations auxiliaires de certains paramètres basées sur des échantillon indépendants de taille différentes. Cependant nous n'avons pas accès aux bases de données indépendantes mais seulement à des estimations issues de ces bases de données. Plus formellement, on a $K \geq 1$ échantillons indépendants $Y_{k,1}, \cdots, Y_{k,m_k}$ où $m_k \in \mathbb{N}^*$ est la taille du k-ième échantillon et $(Y_{k,i})_{1 \leq i \leq m_k}$ sont *i.i.d.* de loi P. Notons que pour tout $k \in [\![1, K]\!]$, m_k est une fonction de n. Pour tout $k \in [\![1, K]\!]$, on a accès aux estimations suivantes $g_{k,1}(\mathbb{P}_{m_k}), \cdots, g_{k,r_k}(\mathbb{P}_{m_k})$ avec $r_k \in \mathbb{N}^*$, on notera $g_k = (g_{k,1}, \cdots, g_{k,r_k}) : Q \mapsto g_k(Q)$. On pose

$$g = (g_k)^T_{k \in [\![1,K]\!]}$$

l'ensemble des informations auxiliaires de taille $m = \sum_{k=1}^K r_k$.

Remarque 4.1.1. *Dans le cas où l'information auxiliaire est donnée par des* ***espérances*** $Pg_1, \cdots, Pg_K$ *alors il est possible de relâcher l'hypothèse de la loi des données auxiliaires en supposant que* $Q := \mathbb{P}^Y$ *vérifie* $Qg_k = Pg_k$ *pour tout* $k \in [\![1, K]\!]$.

Ainsi il y a $\sum_{k=1}^{K} r_k$ informations auxiliaires. Et notons

$$\varphi_g(P) = \left(\varphi_{g_k}(P)\right)_{k\in[\![1,K]\!]}$$

les différentielles de chaque information auxiliaire. Enfin on fera l'hypothèse suivante

$$\forall k \in [\![1,K]\!] \ , \ \frac{n}{m_k} \xrightarrow[n\to+\infty]{} c_k \in \overline{\mathbb{R}}_+.$$

On utilise les mêmes hypothèses que dans la section 2.4.

Hypothèses 4.1.1. De plus, on suppose que ces informations auxiliaires vérifient les conditions de régularité suivantes pour une classe de fonctions $\mathscr{F}$:

1. Notons $\mathscr{H} = \{g_1, \cdots, g_m\}$ une classe de fonctions définies sur $l^\infty(\mathscr{F})$ à valeurs réelles. Soit $t_0 > 0$. Pour tout $h \in \mathscr{H}$, $h(P)$ est défini et pour tout $Q \in l^\infty(\mathscr{F})$ vérifiant $\|Q\|_{\mathscr{F}} < t_0$,

$$h(P+Q) = h(P) + Q \circ \varphi_h(P) + R_h(Q)$$

   où $\varphi_h : l^\infty(\mathscr{F}) \to Vect(\mathscr{F}_h)$ et $\mathscr{F}_h$ est un ensemble **fini** de fonctions définies sur $\mathscr{X}$ et à valeurs réelles. Puisque $\mathscr{H}$ est finie et quitte à agrandir $\mathscr{F}$, on peut supposer que pour tout $h \in \mathscr{H}$, $\mathscr{F}_h \subset \mathscr{F}$. En munissant $Vect(\mathscr{F}_h)$ de la topologie de la convergence simple et $l^\infty(\mathscr{F})$ muni de sa norme $\|\cdot\|_{\mathscr{F}}$, l'application φ_g est continue. Enfin $R_h : l^\infty(\mathscr{F}) \to \mathbb{R}$ est une application vérifiant sur une boule centrée $\mathscr{B}$ en l'origine, $|R_h(\cdot)| \leq \|\cdot\|_{\mathscr{F}}^q$ pour un certain $q > 1$.
2. Pour tout $h \in \mathscr{H}$, $Var(\varphi_h(P))$ existe et est inversible.

Posons

$$I_{K,n} = \begin{pmatrix} g_1(\mathbb{P}_{m_1}) \\ \vdots \\ g_K(\mathbb{P}_{m_K}) \end{pmatrix}.$$

On cherche une mesure empirique informée de la forme

$$\mathbb{P}_n^I f = \mathbb{P}_n f + \Lambda_n^n(f) \cdot (g(\mathbb{P}_n) - I_{K,n}).$$

avec $f : \mathscr{X} \to \mathbb{R}$ une fonction mesurable. Cela permettra de généraliser la relation (2.12) à une information auxiliaire faible. Notons

$$\tilde{I}_{n,K} = \begin{pmatrix} \mathbb{P}_{m_1}\varphi_{g_1}(P) \\ \vdots \\ \mathbb{P}_{m_K}\varphi_{g_K}(P) \end{pmatrix}.$$

En suivant la même approche que dans la remarque 2.3.2, on s'intéresse à un certain problème de minimisation permettant d'obtenir à terme $\Lambda_n^n(f)$.

Proposition 4.1.2. *Posons*

$$C_g = \begin{pmatrix} \frac{Var_P \varphi_{g_1}(P)}{m_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{Var_P \varphi_{g_K}(P)}{m_K} \end{pmatrix}$$

et supposons que $Var_P \varphi_g(P) + nC_g$ est inversible. Alors le paramètre $\Lambda^n(f)$ qui minimise pour toute fonction intégrable $f : \mathscr{X} \to \mathbb{R}$ l'application suivante

$$x \in \mathbb{R}^m \mapsto Var\left(\mathbb{P}_n f + x^T \cdot (\mathbb{P}_n \varphi_g(P) - \tilde{I}_{K,n})\right)$$

est

$$\Lambda^n(f)^T = -cov_P(f, \varphi_g(P)) \left(Var_P \varphi_g(P) + nC_g\right)^{-1}.$$

De plus

$$\boldsymbol{C}_g = \lim_{n \to +\infty} nC_g = \begin{pmatrix} c_1 Var_P \varphi_{g_1}(P) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & c_K Var_P \varphi_{g_K}(P) \end{pmatrix}.$$

Remarque 4.1.3. *Dans le contexte de la remarque 4.1.1 et en notant $Q := \mathbb{P}^Y$, on a*

$$C_g = \begin{pmatrix} \frac{Var_Q g_1}{m_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{Var_Q g_K}{m_K} \end{pmatrix},$$

$$\boldsymbol{C}_g = \begin{pmatrix} c_1 Var_Q g_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & c_K Var_Q g_K \end{pmatrix}.$$

Démonstration.
Fixons $f : \mathscr{X} \to \mathbb{R}$ une fonction intégrable et soit $x \in \mathbb{R}^m$. Alors

$$\begin{aligned} & Var\left(\mathbb{P}_n f + x^T \cdot (\mathbb{P}_n \varphi_g(P) - \tilde{I}_{K,n})\right) \\ &= \frac{1}{n}\left(Var_P f + x^T (Var_P \varphi_g(P)) x + 2cov(f, \varphi_g(P)) x\right) + x^T C_g x \\ &= \frac{1}{n}\left(Var_P f + x^T \left(Var_P \varphi_g(P) + nC_g\right) x + 2cov(f, \varphi_g(P)) x\right). \end{aligned}$$

On pose pour tout $x \in \mathbb{R}^m$

$$\psi(x) = Var_P f + x^T \left(Var_P \varphi_g(P) + nC_g\right) x + 2cov(f, \varphi_g(P))x.$$

La différentielle de ψ *en* x *est alors*

$$D\psi(x) = 2\left(Var_P \varphi_g(P) + nC_g\right) x + 2cov(f, \varphi_g(P))^T.$$

Puisque $Var_P \varphi_g(P) + nC_g$ *est inversible alors*

$$x = -\left(Var_P \varphi_g(P) + nC_g\right)^{-1} cov(f, \varphi_g(P))^T.$$

Ainsi

$$\Lambda^n(f) = -cov(f, \varphi_g(P))\left(Var_P \varphi_g(P) + nC_g\right)^{-1}.$$

□

Puisque nous ne connaissons pas P, Λ^n n'est pas calculable pour une fonction intégrable $f : \mathscr{X} \to \mathbb{R}$ donnée. Ainsi on remplace Λ^n par une **estimation empirique** donnée par

$$\Lambda_n^n(f) = -cov_n(f, \varphi_g(\mathbb{P}_n))\left(Var_n \varphi_g(\mathbb{P}_n) + nC_{g,n}\right)^{-1}$$

où

$$C_{g,n} = \begin{pmatrix} \frac{Var_n \varphi_{g_1}(\mathbb{P}_n)}{m_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{Var_n \varphi_{g_K}(\mathbb{P}_n)}{m_K} \end{pmatrix},$$

$$\mathbf{C}_{g,n} = nC_{g,n}.$$

On pose

$$\Sigma_{1,n}(f) = cov_n(f, \varphi_g(\mathbb{P}_n)),$$
$$\Sigma_{2,n} = Var_n \varphi_g(\mathbb{P}_n) + nC_{g,n}.$$

Ainsi $\Lambda_n^n(f) = -\Sigma_{1,n}(f)\Sigma_{2,n}^{-1}$.

Definition 4.1.1. La **mesure empirique informée** $\mathbb{P}_n^I$ associée à l'information auxiliaire faible donnée par des estimations issues de sources indépendantes est définie par

$$\mathbb{P}_n^I := \sum_{i=1}^n q_i \delta_{X_i}$$

où pour tout $i \in [\![1, n]\!]$,

$$q_i = \frac{1}{n}\left(1 - \varphi_g(\mathbb{P}_n)(X_i)^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - I_{K,n}) + \mathbb{P}_n(\varphi_g(\mathbb{P}_n))^T \Sigma_{2,n}^{-1}(g(\mathbb{P}_n) - I_{K,n})\right).$$

Énonçons à présent un lemme technique utile portant sur Λ_n^n.

Lemme 4.1.4. *Soit $\mathscr{F}$ une classe P-Glivenko-Cantelli ayant une enveloppe $F \in L^2(P)$. On suppose que $Var(\varphi_g(P)) + \boldsymbol{C}_g$ est inversible. On a :*

i) $\Sigma_{1,n} \xrightarrow[n\to+\infty]{p.s.,\ \|\cdot\|_{\mathscr{F}}} \Sigma_1 : f \mapsto cov_P(\varphi_g(P), f)^T$.

ii) $\Sigma_{2,n}^{-1} \xrightarrow[n\to+\infty]{p.s.} \Sigma_2^{-1} = \left(Var(\varphi_g(P)) + \boldsymbol{C}_g\right)^{-1}$.

iii) $\Lambda_n^n \xrightarrow[n\to+\infty]{p.s.,\ \|\cdot\|_{\mathscr{F}}} \Lambda = -\Sigma_1(\cdot)\Sigma_2^{-1}$.

Démonstration.
L'assertion $i)$ a déjà été prouvée dans le lemme 3.2.1. Il nous reste à démontrer l'assertion $ii)$. Tout d'abord puisque pour tout $k \in [\![1, K]\!]$, $\frac{n}{m_k} \xrightarrow[n\to+\infty]{} c_k$ et $Var_n(\varphi_{g_k}(\mathbb{P}_n)) \xrightarrow[n\to+\infty]{p.s.} Var(\varphi_g(P))$, on en déduit que

$$Var_n\varphi_g(\mathbb{P}_n) + \boldsymbol{C}_{g,n} \xrightarrow[n\to+\infty]{p.s.} Var(\varphi_g(P)) + \boldsymbol{C}_g.$$

Enfin en utilisant le fait que $Var(\varphi_g(P)) + \boldsymbol{C}_g$ est inversible et que l'application déterminant est continue, on en déduit que p.s. à partir d'un certain rang $Var_n(\varphi_g(\mathbb{P}_n)) + C_{g,n}$ est inversible. Puisque l'application $A \mapsto A^{-1}$ est continue dans $GL_m(\mathbb{R})$, alors on a le résultat souhaité

$$\Sigma_{2,n}^{-1} = Var_n\varphi_g(\mathbb{P}_n) + \boldsymbol{C}_{g,n} \xrightarrow[n\to+\infty]{p.s.} \Sigma_2^{-1} = Var(\varphi_g(P)) + \boldsymbol{C}_g.$$

***Pour** $iii)$. Ce dernier se déduit des deux précédents par découpage, en effet*

$$\begin{aligned}\sup_{f\in\mathscr{F}} \|\Lambda_n(f) - \Lambda(f)\| &\le \sup_{f\in\mathscr{F}} \|\Sigma_{1,n}(f) - \Sigma_1(f)\| \|\Sigma_{2,n}^{-1}\| + \|\Sigma_{2,n}^{-1} - \Sigma_2^{-1}\| \sup_{f\in\mathscr{F}} \|\Sigma_1(f)\| \\ &\xrightarrow[n\to+\infty]{p.s.} 0.\end{aligned}$$

□

On en déduit un résultat du type **Glivenko-Cantelli** pour la mesure empirique informée.

Théorème 4.1.5. *Soit $\mathscr{F}$ une classe P-Glivenko-Cantelli ayant une enveloppe $F \in L^2(P)$. On dispose d'une information auxiliaire faible donnée par des estimations issues de sources indépendantes. Alors*

$$\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^* \xrightarrow[n\to+\infty]{p.s.} 0$$

avec $\mathbb{P}_n^I$ définie comme dans la définition 4.1.1.

Démonstration.
Notons $m = \sum_{k=1}^{K} r_k$ le nombre de fonctions auxiliaires et on réindexe ces dernières de 1 à m. Remarquons que pour tout $f \in \mathscr{F}$,

$$\begin{aligned}
\|\Lambda_n(f)\| &= \|cov_n(\varphi_g(\mathbb{P}_n), f)^T \Sigma_{2,n}^{-1}\| \\
&\leq \sqrt{Var_n f}\|\Sigma_{2,n}^{-1}\| \max_{1\leq i\leq m} \sqrt{Var_n \varphi_{g_i}(\mathbb{P}_n)} \\
&\leq \sqrt{m}\|f\|_{L^2(\mathbb{P}_n)}\|\Sigma_{2,n}^{-1}\| \max_{1\leq i\leq m} \sqrt{Var_n \varphi_{g_i}(\mathbb{P}_n)} \\
&\leq \sqrt{m}\|F\|_{L^2(\mathbb{P}_n)}\|\Sigma_{2,n}^{-1}\| \max_{1\leq i\leq m} \sqrt{Var_n \varphi_{g_i}(\mathbb{P}_n)}.
\end{aligned}$$

D'où

$$\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^* \leq \|\mathbb{P}_n - P\|_{\mathscr{F}}^* + \|g(\mathbb{P}_n) - I_{K,n}\| \sup_{f\in\mathscr{F}} \|\Lambda_n(f)\| \xrightarrow[n\to\infty]{p.s.} 0$$

par le lemme technique précédent 4.1.4 et que $g(\mathbb{P}_n) - I_{K,n} \xrightarrow[n\to\infty]{p.s.} 0$. □

On souhaite maintenant obtenir un résultat du type Donsker. Tout d'abord remarquons que par le lemme technique 4.1.4, il suffit d'étudier la convergence en loi du processus

$$\alpha_n(f) + \Lambda(f)\cdot\sqrt{n}(g(\mathbb{P}_n) - I_{K,n}).$$

Par les conditions de régularité portant sur les informations auxiliaires, on a

$$\begin{aligned}
g_k(\mathbb{P}_n) - g_k(\mathbb{P}_{m_k}) &= (\mathbb{P}_n - P)(\varphi_{g_k}(P)) + R_{g_k}(\mathbb{P}_n - P) \\
&\quad - (\mathbb{P}_{m_k} - P)(\varphi_{g_k}(P)) - R_{g_k}(\mathbb{P}_{m_k} - P) \\
&= \frac{\alpha_n(\varphi_{g_k}(P))}{\sqrt{n}} - \frac{\alpha_{m_k}(\varphi_{g_k}(P))}{\sqrt{m_k}} + R_{g_k}(\mathbb{P}_n - P) - R_{g_k}(\mathbb{P}_{m_k} - P)
\end{aligned}$$

en remarquant que le reste tend p.s. vers 0 pour une classe P-Glivenko-Cantelli. Ainsi il suffit d'étudier la convergence en loi du processus sur une classe de fonctions $\mathscr{F}$,

$$\alpha_n^I(f) = \alpha_n(f) + \Lambda(f)\cdot\sqrt{n}\left(\frac{\alpha_n(\varphi_g(P))}{\sqrt{n}} - \begin{pmatrix} \frac{\alpha_{m_1}(\varphi_{g_1}(P))}{\sqrt{m_1}} \\ \vdots \\ \frac{\alpha_{m_K}(\varphi_{g_K}(P))}{\sqrt{m_K}} \end{pmatrix}\right) + o_p(1)$$

où $f \in \mathscr{F}$. Énonçons à présent un théorème du type **Donsker** pour la mesure empirique informée.

Théorème 4.1.6. *Si $\mathscr{F}$ est une classe P-Donsker alors*

$$\alpha_n^I \Rightarrow G_I \ \ dans \ \ l^\infty(\mathscr{F})$$

où

$$G_I(f) = G(f) + \Lambda(f) \cdot \left(G(\varphi_g(P)) - \begin{pmatrix} \sqrt{c_1} G_1(\varphi_{g_1}(P)) \\ \vdots \\ \sqrt{c_K} G_K(\varphi_{g_K}(P)) \end{pmatrix} \right)$$

et $G, G_1, \cdots, G_K$ sont des P-ponts browniens indépendants deux à deux.

Remarque 4.1.7. *Dans le contexte de la remarque 4.1.1, $G_1, \cdots, G_K$ sont des Q-ponts browniens indépendants deux à deux où $Q := \mathbb{P}^Y$.*

Démonstration.
Par ce qui précède, il suffit d'étudier la convergence en loi du processus suivant

$$\begin{aligned}
\mathbb{Z}_n(f) &= \alpha_n(f) + \Lambda(f) \cdot \sqrt{n} \left(\frac{\alpha_n(\varphi_g(P))}{\sqrt{n}} - \begin{pmatrix} \frac{\alpha_{m_1}(\varphi_{g_1}(P))}{\sqrt{m_1}} \\ \vdots \\ \frac{\alpha_{m_K}(\varphi_{g_k}(P))}{\sqrt{m_K}} \end{pmatrix} \right) \\
&= \alpha_n(f) + \Lambda(f) \cdot \alpha_n(\varphi_g(P)) - \Lambda(f) \begin{pmatrix} \sqrt{\frac{n}{m_1}} \alpha_{m_1}(\varphi_{g_1}(P)) \\ \vdots \\ \sqrt{\frac{n}{m_K}} \alpha_{m_K}(\varphi_{g_1}(P)) \end{pmatrix} \\
&= \mathbb{X}_n(f) + \mathbb{Y}_n(f).
\end{aligned}$$

En utilisant la même démonstration que dans le théorème 3.2.3, on a

$$\mathbb{X}_n(\cdot) = \alpha_n(\cdot) + \Lambda(\cdot) \cdot \alpha_n(\varphi_g(P)) \Rightarrow G + \Lambda(\cdot) \cdot G(\varphi_g(P)) \ \ dans \ \ l^\infty(\mathscr{F}).$$

Puisque $\mathbb{Y}_n$ est indépendant de $\mathbb{X}_n$ et

$$\mathbb{Y}_n = -\Lambda(\cdot) \begin{pmatrix} \sqrt{\frac{n}{m_1}} \alpha_{m_1}(\varphi_{g_1}(P)) \\ \vdots \\ \sqrt{\frac{n}{m_K}} \alpha_{m_K}(\varphi_{g_1}(P)) \end{pmatrix} \Rightarrow -\Lambda(\cdot) \begin{pmatrix} \sqrt{c_1} G_1(\varphi_{g_1}(P)) \\ \vdots \\ \sqrt{c_K} G_K(\varphi_{g_K}(P)) \end{pmatrix} \ \ dans \ \ l^\infty(\mathscr{F}),$$

on obtient que

$$\mathbb{Z}_n \Rightarrow G_I \ \ dans \ \ l^\infty(\mathscr{F}).$$

□

De plus nous avons une **décroissance de la variance du processus limite** .

Proposition 4.1.8. *Soit $\mathscr{F}$ une classe P-Donsker. Alors la variance du processus limite G_I est donnée pour tout $f \in \mathscr{F}$ par*

$$Var(G_I(f)) = Var_P f - cov(f, \varphi_g(P)) \left(Var(\varphi_g(P)) + \boldsymbol{C}_g\right)^{-1} cov(f, \varphi_g(P))^T.$$

De plus

$$Var(G_I(f)) \leq Var_P f.$$

Remarque 4.1.9. *Dans le contexte de la remarque 4.1.1, la variance du processus limite pour $f \in \mathscr{F}$ est*

$$Var(G_I(f)) = Var_P f - cov(f, g) \left(Var(g) + \boldsymbol{C}_g\right)^{-1} cov(f, g)^T$$

où $\boldsymbol{C}_g$ est donnée dans la remarque 4.1.3.

Démonstration.
Calculons la variance du processus limite. Soit $f \in \mathscr{F}$. Remarquons que

$$Var(G_I(f)) = Var\left(G(f) + \Lambda(f) \cdot G(\varphi_g(P))\right) + Var\left(\Lambda(f) \cdot \left(\begin{pmatrix} \sqrt{c_1} G_1(\varphi_{g_1}(P)) \\ \vdots \\ \sqrt{c_K} G_K(\varphi_{g_K}(P)) \end{pmatrix}\right)\right).$$

En développant la variance, le premier terme s'écrit

$$Var\left(G(f) + \Lambda(f) \cdot G(\varphi_g(P))\right) = Var_P f + \Lambda(f) Var_P \varphi_g(P) \Lambda(f)^T + 2 cov_P(f, \varphi_g(P)) \Lambda(f)^T.$$

Puisque $G_1, \cdots, G_K$ sont indépendants alors le second terme est

$$Var\left(\begin{pmatrix} \sqrt{c_1} G_1(\varphi_{g_1}(P)) \\ \vdots \\ \sqrt{c_K} G_K(\varphi_{g_K}(P)) \end{pmatrix}\right) = \boldsymbol{C}_g.$$

Ainsi

$$Var(G_I(f)) = Var_P f + \Lambda(f) \left(Var_P \varphi_g(P) + \boldsymbol{C}_g\right) \Lambda(f)^T + 2 cov_P(f, \varphi_g(P)) \Lambda(f)^T.$$

Rappelons que $\Lambda(f) = -cov_P(f, \varphi_g(P)) \left(Var(\varphi_g(P)) + \boldsymbol{C}_g\right)^{-1}$. On obtient alors

$$Var(G_I(f)) = Var_P f - cov(f, \varphi_g(P)) \left(Var(\varphi_g(P)) + \boldsymbol{C}_g\right)^{-1} cov(f, \varphi_g(P))^T \leq Var_P f$$

car $Var(\varphi_g(P)) + \boldsymbol{C}_g$ est une matrice symétrique définie positive comme somme de deux matrices symétriques definies positives et $\boldsymbol{C}_g$ est une matrice diagonale par blocs où les blocs sont des matrices symétriques positives. □

Exemple 4.1.1. Dans cet exemple, nous nous intéressons au cas où nous avons qu'une estimation issue d'une source indépendante c'est à dire $K = 1$. Notons $c := c_1$ et $c_n = \frac{n}{m_1}$. Alors

$$\begin{aligned} Var(\varphi_g(P)) + \mathbf{C}_g &= Var(\varphi_g(P)) + cVar(\varphi_g(P)) \\ &= (1+c)Var(\varphi_g(P)) \end{aligned}$$

et

$$\begin{aligned} Var_n(\varphi_g(\mathbb{P}_n)) + \mathbf{C}_{g,n} &= Var_n(\varphi_g(\mathbb{P}_n)) + cVar_n(\varphi_g(\mathbb{P}_n)) \\ &= (1+c_n)Var_n(\varphi_g(\mathbb{P}_n)). \end{aligned}$$

On en déduit que

$$Var(G_I(f)) = Var_P f - \frac{1}{1+c} cov(f, \varphi_g(P)) \left(Var(\varphi_g(P))\right)^{-1} cov(f, \varphi_g(P))^T.$$

Ainsi pour tout $0 \leq c < +\infty$, on améliore l'estimation. Le cas $c = 0$ correspond au cas **totalement informé**. De plus, remarquons que le fait d'injecter des estimations basées sur un échantillon de taille $\frac{n}{4}$ améliore l'estimation puisque $c = 4$. Enfin dans le cas où $g(P) = Pg$, on remarque que $\mathbb{P}_n^I g \neq \mathbb{P}_{m_1} g$ lorsque $c_n \neq 0$ car

$$\begin{aligned} \mathbb{P}_n^I g &= \mathbb{P}_n g - \frac{1}{1+c_n} Var_n(\varphi_g(\mathbb{P}_n)) Var_n(\varphi_g(\mathbb{P}_n))^{-1} (\mathbb{P}_n g - \mathbb{P}_{m_1} g) \\ &= \frac{c_n}{1+c_n} \mathbb{P}_n g + \frac{1}{1+c_n} \mathbb{P}_{m_1} g. \end{aligned}$$

Ainsi $\mathbb{P}_n^I g$ est une combinaison convexe entre ces deux points. Ainsi plus c_n est proche de 0 (autrement dit plus la taille de la source auxiliaire est grande par rapport à n), plus $\mathbb{P}_n^I g$ sera proche de $\mathbb{P}_{m_1} g$.

4.1.2 Ajout de données auxiliaires à l'échantillon initial

On se place dans le même contexte que précédemment à la seule différence que cette fois-ci, **nous avons accès à l'ensemble des K-bases de données indépendantes** avec $K \in \mathbb{N}^*$. Par indépendance des K-échantillons et puisque toutes les variables aléatoires sont *i.i.d.* de loi P, on dispose d'un nouvel échantillon de taille $n + \sum_{i=1}^K m_i$ sous la forme

$$\{Z_1, \cdots, Z_{M_{K_n}}\} = \{X_1, \cdots, X_n, Y_{1,1}, \cdots, Y_{1,m_1}, \cdots, Y_{K,1}, \cdots, Y_{k,m_K}\}$$

où $M_{K,n} = n + \sum_{i=1}^K m_i$.

Definition 4.1.2. La **mesure empirique informée** associée à l'information auxiliaire faible apportée par cet ajout de données auxiliaires est définie par

$$\mathbb{P}_n^I = \mathbb{P}_{M_{K,n}} = \frac{1}{M_{K,n}} \sum_{i=1}^{M_{K,n}} \delta_{Z_i}.$$

Nous pouvons alors obtenir un résultat de type **Donsker** pour le processus empirique informé.

Théorème 4.1.10. *Notons $c_k = \lim_{n\to\infty} \frac{n}{m_k}$. Soient $\mathscr{F}$ une classe P-Donsker et $\mathbb{P}_n^I$ donnée dans la définition 4.1.2. On obtient que :*

- *Si pour tout $k \in [\![1, K]\!]$, $c_k > 0$ alors*

$$\sqrt{n}(\mathbb{P}_n^I - P) \Rightarrow \frac{G}{\sqrt{1 + \sum_{k=1}^K \frac{1}{c_k}}} \ \ dans \ \ l^\infty(\mathscr{F})$$

  où G est P-pont brownien.
- *S'il existe $k \in [\![1, K]\!]$ tel que $c_k = 0$ alors*

$$\sqrt{n}\|\mathbb{P}_n^I - P\|_{\mathscr{F}} \xrightarrow{\mathbb{P}} 0,$$
$$\sqrt{\sum_{i=1}^K m_i} \left(\mathbb{P}_n^I - P\right) \Rightarrow G \ \ dans \ \ l^\infty(\mathscr{F})$$

  où G est P-pont brownien. La vitesse de convergence est alors améliorée grâce à l'information auxiliaire donnée.

Démonstration.
La preuve se déduit rapidement en remarquant que

$$\sqrt{n}(\mathbb{P}_n^I - P) = \sqrt{\frac{n}{M_{K,n}}} \sqrt{M_{K,n}}(\mathbb{P}_{M_{K,n}} - P),$$
$$\sqrt{M_{K,n}}(\mathbb{P}_{M_{K,n}} - P) \Rightarrow G \ \ dans \ \ l^\infty(\mathscr{F}),$$
$$\sqrt{\frac{n}{M_{K,n}}} = \sqrt{\frac{n}{n + \sum_{i=1}^K m_i}} \xrightarrow{n\to\infty} \frac{1}{\sqrt{1 + \sum_{k=1}^K \frac{1}{c_k}}},$$
$$\sqrt{n + \sum_{i=1}^K m_i} = \sqrt{\sum_{i=1}^K m_i} \sqrt{\frac{n}{\sum_{i=1}^K m_i} + 1}.$$

□

Remarque 4.1.11. *On remarque alors que l'ajout de données auxiliaires permet au moins de réduire la variance et au mieux d'améliorer la vitesse de convergence.*

Nous pouvons à présent comparer la diminution de variance entre le fait d'injecter une estimation d'une fonctionnelle basée sur une source indépendante de l'échantillon initial et le fait d'utiliser une base de données indépendantes.

Exemple 4.1.2. Supposons que l'on dispose d'un échantillon indépendant de taille m (autrement dit $K = 1$) et on note $c = \lim_{n\to+\infty} \frac{n}{m} > 0$. Par la sous-section précédente 4.1.1, nous avons vu que lorsqu'on injecte une estimation de $g(P)$ basée sur l'échantillon indépendant, la variance du processus limite est pour toute fonction mesurable f

$$Var\ G_I(f) = Var_P f - cov(f, \varphi_g(P)) \left(Var(\varphi_g(P)) + \mathbf{C}_g\right)^{-1} cov(f, \varphi_g(P))^T.$$

Dans le cas où nous injectons des données auxiliaires, la variance du processus limite est donnée par

$$Var\left(\frac{G(f)}{\sqrt{1+\frac{1}{c}}}\right) = \frac{c}{1+c} Var_P f.$$

Ainsi la différence des variances asymptotiques est

$$\begin{aligned} &Var(G_I(f)) - Var\left(\frac{G(f)}{\sqrt{1+\frac{1}{c}}}\right) \\ &= \frac{1}{1+c}\left(Var_P f - cov(f, \varphi_g(P)) \left(Var(\varphi_g(P))\right)^{-1} cov(f, \varphi_g(P))^T\right) \geq 0. \end{aligned}$$

Par conséquent, on peut conclure par cet exemple qu'utiliser des données auxiliaires apportent plus d'information que de donner seulement une estimation basée sur ces données auxiliaires.

4.1.3 Un mélange des deux informations auxiliaires faibles précédentes

Dans cette sous-section, nous souhaitons incorporer simultanément les deux types d'information auxiliaire faible étudiés dans les sous-sections 4.1.1 et 4.1.2. Plus précisément, nous supposons que nous n'avons pas accès à l'ensemble des bases de données mais **seulement à certaines**. Concernant les bases de données restantes, nous avons seulement accès à des statistiques issues de ces bases de données. Dans cette sous-section, nous conservons les mêmes notations que dans les sous-sections 4.1.1 et 4.1.2. Notons $K \in \mathbb{N}^*$ le nombre de sources indépendantes. Quitte à réordonner, on suppose qu'on a accès aux $J \in [\![1, K-1]\!]$ premières bases de données et seulement à des statistiques

pour les bases de données restantes $[\![J+1,K]\!]$. En notant m_k la taille de la k-ième base de donnée pour $k \in [\![1,K]\!]$, nous avons à notre disposition $n+M_J$ données avec $M_J := \sum_{i=1}^{J} m_i$. L'échantillon sera noté $\{Z_1,\cdots,Z_{n+M_J}\}$ et la mesure empirique de l'échantillon est définie de la manière suivante

$$\mathbb{P}_{n+M_J} = \frac{1}{n+M_J}\sum_{i=1}^{n+M_J}\delta_{Z_i}.$$

De plus, nous notons $\mathbb{P}_{m_l}$ la mesure empirique associée à la l-ème base de données avec $l \in [\![J+1,K]\!]$ et $g = (g_{J+1},...,g_K)$ les fonctionnelles associées aux statistiques issues des bases de données restantes. De manière similaire à la sous-section 4.1.1, la mesure empirique informée peut s'écrire pour toute fonction mesurable de carré intégrable de la manière suivante

$$\mathbb{P}_n^I f = \mathbb{P}_{n+M_J} f + \Lambda_{n+M_J}^{n+M_J}(f)\cdot\left(g\left(\mathbb{P}_{n+M_J}\right) - \begin{pmatrix} g_{J+1}(\mathbb{P}_{m_{J+1}}) \\ \vdots \\ g_K(\mathbb{P}_{m_K})\end{pmatrix}\right)$$

où

$$\Lambda_{n+M_J}^{n+M_J}(f) = -cov_{n+M_J}(f,\varphi_g(\mathbb{P}_{n+M_J}))\left(Var_{n+M_J}\varphi_g(\mathbb{P}_{n+M_J}) + (n+M_j)C_{g,n+M_J}\right)^{-1},$$

$$C_{g,n+M_J} = \begin{pmatrix} \frac{Var_{n+M_J}\varphi_{g_{J+1}}(\mathbb{P}_{n+M_J})}{m_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{Var_{n+M_J}\varphi_{g_K}(\mathbb{P}_{n+M_J})}{m_K}\end{pmatrix}.$$

Notons

$$\Sigma_{1,n+M_J}(f) = cov_{n+M_J}(f,\varphi_g(\mathbb{P}_{n+M_J})),$$
$$\Sigma_{2,n+M_J} = Var_{n+M_J}\varphi_g(\mathbb{P}_{n+M_J}) + (n+M_j)C_{g,n+M_J}.$$

On peut alors en déduire la définition de la mesure empirique informée.

Definition 4.1.3. La **mesure empirique informée** $\mathbb{P}_n^I$ associée aux mélanges de ces deux types d'informations auxiliaire faible est définie par

$$\mathbb{P}_n^I := \sum_{i=1}^{n+M_J} q_i\delta_{Z_i}$$

où pour tout $i \in [\![1,n+M_J]\!]$,

$$q_i = \frac{1}{n+M_J}\left(1-\varphi_g\left(\mathbb{P}_{n+M_J}\right)(Z_i)\Sigma_{2,n+M_J}^{-1}\begin{pmatrix} g_{J+1}(\mathbb{P}_{m_{J+1}}) \\ \vdots \\ g_K(\mathbb{P}_{m_K})\end{pmatrix} + \mathbb{P}_{n+M_J}\varphi_g\left(\mathbb{P}_{n+M_J}\right)\Sigma_{2,n+M_J}^{-1}\begin{pmatrix} g_{J+1}(\mathbb{P}_{m_{J+1}}) \\ \vdots \\ g_K(\mathbb{P}_{m_K})\end{pmatrix}\right).$$

Comme nous avons pu le faire dans les sous-sections précédentes 4.1.1 et 4.1.2, on peut décomposer le processus empirique informé $\alpha_n^I = \sqrt{n}\left(\mathbb{P}_n^I - P\right)$ pour toute fonction f mesurable de carré intégrable de la manière suivante

$$\alpha_n^I(f) = \sqrt{\frac{n}{n+M_J}}\left(\alpha_{n+M_j}(f) + \Lambda(f)\cdot\left(\alpha_{n+M_j}(\varphi_g(P)) - \begin{pmatrix}\sqrt{\frac{n+M_J}{m_1}}\alpha_{m_1}(\varphi_{g_{J+1}}(P))\\ \vdots \\ \sqrt{\frac{n+M_J}{m_K}}\alpha_{m_K}(\varphi_{g_K}(P))\end{pmatrix}\right)\right)$$
$$+ o_p(1)$$

où

$$\Lambda(f) = -cov_P(f,\varphi_g(P))\left(Var_P\varphi_g(P) + \mathbf{C}_g\right)^{-1},$$
$$\mathbf{C}_g = \lim_n (n+M_J)C_{g,n+M_J} = \begin{pmatrix} w_{J+1}Var_P\varphi_{g_{J+1}}(P) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_K Var_P\varphi_{g_K}(P)\end{pmatrix}$$

et $w_l = \lim_n \frac{n+M_J}{m_l} = c_l\left(1+\sum_{i=1}^J \frac{1}{c_i}\right)$ pour $l \in [\![J+1,K]\!]$.

On en déduit le résultat du type **Donsker** dont la démonstration est similaire à celle du théorème 4.1.6.

Théorème 4.1.12. *Soit $\mathscr{F}$ est une classe P-Donsker. On suppose que pour tout $k \in [\![1,K]\!]$, $c_k \in \mathbb{R}_+^*$. Posons :*

$$w_l = \lim_n \frac{n+M_J}{m_l} = c_l\left(1+\sum_{i=1}^J \frac{1}{c_i}\right),\ l \in [\![J+1,K]\!],$$
$$\lambda_J = \lim_{n\to\infty}\sqrt{\frac{n}{n+M_J}} = \frac{1}{\sqrt{1+\sum_{k=1}^J \frac{1}{c_k}}}.$$

Alors

$$\alpha_n^I \Rightarrow G_I \ \ dans\ \ l^\infty(\mathscr{F})$$

où

$$G_I(f) = \lambda_J\left(G(f) + \Lambda(f)\cdot\left(G(\varphi_g(P)) - \begin{pmatrix}\sqrt{w_{J+1}}G_{J+1}(\varphi_{g_{J+1}}(P))\\ \vdots \\ \sqrt{w_K}G_K(\varphi_{g_K}(P))\end{pmatrix}\right)\right)$$

et $G, G_{J+1}, \cdots, G_K$ sont des P-ponts brownien indépendants deux à deux.

Remarque 4.1.13. *Puisque $\lambda_J \leq 1$, nous pouvons facilement remarquer que nous avons une diminution de variance asymptotique*

$$\begin{aligned} Var\ G_I(f) &= \lambda_J^2 \left(Var\left(G(f) + \Lambda(f) \cdot G(\varphi_g(P))\right) + Var\left(\Lambda(f) \cdot \left(\begin{pmatrix} \sqrt{w_{J+1}} G_{J+1}(\varphi_{g_{J+1}}(P)) \\ \vdots \\ \sqrt{w_K} G_K(\varphi_{g_K}(P)) \end{pmatrix}\right)\right)\right) \\ &\leq Var\left(G(f) + \Lambda(f) \cdot G(\varphi_g(P))\right) + Var\left(\Lambda(f) \cdot \left(\begin{pmatrix} \sqrt{w_{J+1}} G_{J+1}(\varphi_{g_{J+1}}(P)) \\ \vdots \\ \sqrt{w_K} G_K(\varphi_{g_K}(P)) \end{pmatrix}\right)\right) \\ &\leq Var_P f. \end{aligned}$$

4.2 Information auxiliaire donnée par une mesure des préférences d'un expert

4.2.1 Choix entre un nombre fini de valeurs et généralisation au cas infini

Le contexte est le suivant. Un expert nous informe que $g(P) \in \{\alpha_1, ..., \alpha_M\}$ et accorde une certaine préférence à chaque α_i pour $i \in [\![1, M]\!]$. Notons I cette information auxiliaire. On note I_0 le cas où on n'a aucune information et pour tout $k \in [\![1, M]\!]$, $I_k : g(P) = \alpha_k$ dans le cas où on suppose que $g(P) = \alpha_k$. Ainsi il existe un unique $k^* \in [\![1, M]\!]$ tel que $g(P) = \alpha_{k^*}$. Cependant cet indice est inconnu dans le cadre de notre problème de choix.

On modélise la préférence de l'expert en faisant une combinaison convexe des mesures empiriques informées $\left\{\mathbb{P}_n^{I_0}, \mathbb{P}_n^{I_1}, \cdots, \mathbb{P}_n^{I_M}\right\}$. Plus précisément, on définit la **mesure empirique informée** de la manière suivante

$$\mathbb{P}_n^I = \sum_{k=0}^{M} \beta_{k,n} \mathbb{P}_n^{I_k}$$

où $\mathbb{P}_n^{I_0} = \mathbb{P}_n$ et pour tout $(k, n) \in [\![1, M]\!] \times \mathbb{N}^*$, $\beta_{k,n} \in [0, 1]$ et $\sum_{k=0}^{M} \beta_{k,n} = 1$. De plus rappelons que pour tout $k \in [\![1, M]\!]$, $\mathbb{P}_n^{I_k}$ est la mesure empirique informée par I_k définie dans la définition 2.4.1.

Exemple 4.2.1. Donnons quelques exemples de choix des $\beta_{k,n}$.

- On suppose qu'il existe une variable aléatoire Z_n (généralement un estimateur empirique de la variance limite) telle que

$$T_{k^*,n} = \frac{\sqrt{n}\left(g(\mathbb{P}_n) - \alpha_{k^*}\right)}{Z_n} \Rightarrow \mathcal{N}(0, 1).$$

Alors en notant $z_{1-\alpha/2}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $|\mathcal{N}(0,1)|$ avec $\alpha \in]0,1[$, on peut poser pour tout $k \in [\![1,M]\!]$

$$\beta_{k,n} = \frac{\beta_k}{C_n} 1_{|T_{k,n}| \leq z_{1-\alpha/2}},$$
$$\beta_{0,n} = \frac{\beta_0}{C_n} \prod_{k=1}^{M} 1_{|T_{k,n}| > z_{1-\alpha/2}},$$

où $(\beta_k)_{k\in[\![0,M]\!]}$ représente les préférences de l'expert et C_n est la constante de normalisation c'est à dire

$$C_n = \beta_0 \prod_{k=1}^{M} 1_{|T_{k,n}| > z_{1-\alpha/2}} + \sum_{k=1}^{M} \beta_k 1_{|T_{k,n}| \leq z_{1-\alpha/2}}.$$

- Choix aléatoire : Soit Y_n une variable aléatoire à support dans $\{e_0, ..., e_M\}$ (la base canonique de $\mathbb{R}^{M+1}$) telle que pour tout $k \in [\![0,M]\!]$:

$$\mathbb{P}(Y_n = e_k) = a_{k,n}$$

avec $\sum_{k=0}^{M} a_{k,n} = 1$. Ainsi la mesure informée s'écrit

$$\mathbb{P}_n^I = \sum_{k=0}^{M} 1_{Y_n = e_k} \mathbb{P}_n^{I_k}.$$

Le théorème suivant établit la convergence en loi du processus empirique informé $\sqrt{n}(\mathbb{P}_n^I - P)$ où $\mathbb{P}_n^I = \sum_{k=0}^{M} \beta_{k,n} \mathbb{P}_n^{I_k}$.

Théorème 4.2.1. *Supposons que*

- *Pour tout $k \neq k^*$, $\sqrt{n}\beta_{k,n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0$,*
- $\beta_{k^*,n} \xrightarrow[n\to+\infty]{\mathbb{P}} 1$.

Alors pour toute classe de Donsker $\mathcal{F}$, on a

$$\sqrt{n}(\mathbb{P}_n^I - P) \Rightarrow G_{I_{k^*}} \quad \textit{dans } l^\infty(\mathcal{F})$$

où $G_{I_{k^*}}$ est le P-pont brownien informé par I_{k^*}.

Démonstration.

On peut écrire que

$$\sqrt{n}(\mathbb{P}_n^I - P) = \sum_{k=0}^{M} \beta_{k,n}\sqrt{n}(\mathbb{P}_n^{I_k} - P).$$

Remarquons que pour $k \neq k^*$ et pour tout $f \in L^2(P)$,

$$\mathbb{P}_n^{I_k} f - Pf \xrightarrow[n\to+\infty]{p.s.} -cov_P(f, \varphi_{g_k}(P)) Var_P\left(\varphi_{g_k}(P)\right)^{-1}\left(g(P) - \alpha_k\right).$$

En utilisant le fait que pour $k \neq k^*$, $\sqrt{n}\beta_{k,n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0$ et par le lemme technique 3.2.1, on en déduit que

$$\sum_{k=0, k\neq k^*}^{M} \beta_{k,n}\sqrt{n}(\mathbb{P}_n^{I_k} - P) = o_p(1).$$

Ainsi par le théorème 3.2.3, on a pour toute classe de Donsker $\mathscr{F}$

$$\sqrt{n}(\mathbb{P}_n^I - P) = \beta_{k^*,n}\sqrt{n}(\mathbb{P}_n^{I_{k^*}} - P) + o_P(1) \Rightarrow G_{I_{k^*}} \quad dans\ l^\infty(\mathscr{F}).$$

□

On peut généraliser ce procédé à un ensemble de choix infini. Plus précisément on suppose qu'on dispose d'une information auxiliaire faible concernant la fonctionnelle $g(P)$ donnée par un espace $\Theta \subset \mathbb{R}^k$ pour $k \geq 1$ et pour tout $n \in \mathbb{N}^*$ une fonction de poids positifs $(w_n)_{n\in\mathbb{N}^*}$ telle que

$$\int_\Theta w_n(\theta)\, d\mu(\theta) = 1 \tag{4.1}$$

où μ est une mesure σ-finie. Posons

$$\mathbb{P}_n^I := \sum_{i=1}^{n} \left(\int_\Theta q_i(\theta)\, w_n(\theta)\, d\mu(\theta)\right) \delta_{X_i}$$

où pour tout $\theta \in \Theta$ $(q_i(\theta))_{i\in[\![1,n]\!]}$ sont les poids de la mesure empirique informée par $g(P) = \theta$. Ainsi

$$\mathbb{P}_n^{I_\theta} = \sum_{i=1}^{n} q_i(\theta)\delta_{X_i}.$$

Fixons $f \in L^2(P)$ et remarquons que

$$\mathbb{P}_n^I f = \int_\Theta \mathbb{P}_n^{I_\theta} f\ w_n(\theta)\, d\mu(\theta),$$

$$\alpha_n^I(f) = \sqrt{n}\left(\mathbb{P}_n^I f - Pf\right) = \int_\Theta \sqrt{n}\left(\mathbb{P}_n^{I_\theta} f - Pf\right)\ w_n(\theta)\, d\mu(\theta).$$

Notons $\mathbb{E}_{w_n}(\theta) = \int_\Theta \theta\, w_n(\theta)\, d\mu(\theta)$ et observons

$$\sqrt{n}\int_\Theta (\theta - \theta_0)\, w_n(\theta)\, d\mu(\theta) = \sqrt{n}\left(\mathbb{E}_{w_n}(\theta) - \theta_0\right).$$

Le théorème suivant porte sur la convergence en loi de α_n^I.

Théorème 4.2.2. *Soit $\mathscr{F}$ une classe de fonctions P-Donsker. On suppose qu'il existe $\theta_0 \in \Theta$ tel que $g(P) = \theta_0$ et pour tout $m \in \mathbb{N}^*$, $f = (f_1, \cdots, f_m)^T \in \mathscr{F}^k$*

$$\left(\alpha_n(f)^T, \sqrt{n}\left(\mathbb{E}_{w_n}(\theta) - \theta_0\right)\right)^T \Rightarrow \left(G(f), V\right)$$

où V est une variable aléatoire à valeurs dans l'espace $\Theta \subset \mathbb{R}^k$. Alors

$$\alpha_n^I \Rightarrow \widetilde{G}_I \ \ dans\ l^\infty(\mathscr{F})$$

où pour tout $f \in \mathscr{F}$, $\widetilde{G}_I(f) = G_{I_{\theta_0}}(f) - cov_P(f, \varphi_g(P))\Sigma^{-1}V$. La variance asymptotique est alors

$$\begin{aligned} Var\ \widetilde{G}_I(f) = Var\ G_{I_{\theta_0}}(f) + cov_P(f, \varphi_g(P))\Sigma^{-1}\ VarV\ \Sigma^{-1} cov_P(f, \varphi_g(P))^T \\ -2cov(G_I(f), V)\Sigma^{-1} cov_P(f, \varphi_g(P))^T. \end{aligned}$$

Remarque 4.2.3. *Si V est presque sûrement constante alors pour tout $f \in \mathscr{F}$, $\widetilde{G}_I(f)$ a la même variance que $G_{I_{\theta_0}}$ mais a un biais persistant si V n'est pas nulle presque sûrement.*

Démonstration.
Posons pour tout $n \in \mathbb{N}^*$, $V_n = \sqrt{n}\left(\mathbb{E}_{w_n}(\theta) - \theta_0\right)$. Tout d'abord observons que

$$\alpha_n^{I_\theta}(f) = \alpha_n^{I_{\theta_0}}(f) - cov_n(f, \varphi_g(\mathbb{P}_n))\Sigma_n^{-1}\sqrt{n}(\theta - \theta_0).$$

Ainsi

$$\alpha_n^I(f) = \alpha_n^{I_{\theta_0}}(f) - cov_n(f, \varphi_g(\mathbb{P}_n))\Sigma_n^{-1}\sqrt{n}\int_\Theta (\theta - \theta_0)\, w_n(\theta)\, d\mu(\theta).$$

De la même manière que dans la démonstration du théorème 3.1.2 et 3.2.3 , on a

$$\alpha_n^I(f) = \alpha_n(f) - cov_P(\varphi_g(P), f)^T\Sigma^{-1}\alpha_n(\varphi_g(P)) - cov_P(\varphi_g(P), f)^T\Sigma^{-1}V_n + o_p(1).$$

Posons pour tout $f \in \mathscr{F}$

$$W_n(f) = \alpha_n(f) - cov_P(\varphi_g(P), f)^T\Sigma^{-1}\alpha_n(\varphi_g(P)) - cov_P(\varphi_g(P), f)^T\Sigma^{-1}V_n.$$

De manière similaire que dans la démonstration du théorème 3.1.2 et 3.2.3, on peut montrer que pour tout $f = (f_1, \cdots, f_k)^T \in \mathscr{F}$ avec $k \geq 1$

$$\left(W_n(f_1), \cdots, W_n(f_k)\right)^T \Rightarrow \left(\widetilde{G}_I(f_1), \cdots, \widetilde{G}_I(f_k)\right)^T$$

où pour tout $j \in [\![1, k]\!]$, $\widetilde{G}_I(f_j) = G_{I_{\theta_0}}(f) - cov_P(f, \varphi_g(P))\Sigma^{-1}V$. En appliquant la même démonstration que dans le théorème 3.2.3 et en utilisant le lemme technique suivant

Lemme 4.2.4. *Soit $X_n : \Omega_n \to l^\infty(T)$ une suite d'applications. Alors ces deux assertions sont équivalentes :*

- *On a*
  1. *Pour tout $(t_1, ..., t_k) \in T^k$, $(X_n(t_1), ..., X_n(t_k))$ converges en loi vers un vecteur aléatoire dans $\mathbb{R}^k$ pour tout $k \in \mathbb{N}^*$.*
  2. *Il existe une semi-distance ρ telle que (T, ρ) est précompact et pour tout $\epsilon > 0$*

$$\lim_{\delta \to 0} \limsup_n \mathbb{P}^*(\sup_{\rho(s,t)<\delta} |X_n(s) - X_n(t)| > \epsilon) = 0.$$

- *Il existe $X : \Omega \to l^\infty(T)$ un processus stochastique mesurable et tendu tel que*

$$X_n \Rightarrow X \ \ dans\ l^\infty(T).$$

on obtient la convergence en loi souhaitée à savoir que

$$\alpha_n^I \Rightarrow \widetilde{G}_I \ \ dans\ l^\infty(\mathscr{F}).$$

□

4.2.2 Information auxiliaire donnée par une densité de probabilité

Soit $r \in \mathbb{N}^*$. Dans cette sous section, on suppose que les conseils de l'expert sont modélisés par une **densité** pour chaque Pg_i où $i \in [\![1, r]\!]$ et $g = (g_1, \cdots, g_r)$. On note $h = (h_1, \cdots, h_r)$ les densités respectives et pour tout $n \in \mathbb{N}^*$ et pour tout $i \in [\![1, r]\!]$

$$J_{i,n} = \left[\mathbb{P}_n g_i - z_{1-\alpha_n/2}\sqrt{\frac{Var_{\mathbb{P}_n} g_i}{n}}, \mathbb{P}_n g_i + z_{1-\alpha_n/2}\sqrt{\frac{Var_{\mathbb{P}_n} g_i}{n}}\right]$$

où on choisit α_n telle que $z_{1-\alpha_n/2} = C \log n$ ou $z_{1-\alpha_n/2} = C \log\log n$ pour une certaine constante $C > 1$ déterminée par la loi du logarithme itéré. La loi du logarithme itéré nous permet d'affirmer que p.s. à partir d'un certain rang $Pg_i \in J_{i,n}$ pour tout $i \in [\![1, r]\!]$. Pour tout $n \in \mathbb{N}^*$, on injecte l'information auxiliaire suivante

$$I_n : Pg \in (arg \max_{x \in J_{i,n}} h_i(x))_{i \in [\![1,r]\!]}.$$

Plus précisément, on injecte un des maximiseurs (qui est souvent unique pour n assez grand). Pour cela, définissons pour tout $i \in [\![1, r]\!]$ l'ensemble $\mathscr{H}_i$ des maxima locaux de h_i. Dans le cas où h_i est suffisamment régulière cela correspond à l'ensemble

$$\mathscr{H}_i = \{x \in \mathscr{X},\ \nabla h_i(x) = 0 \ \ D^2 h_i(x) > 0\}.$$

Posons $\mathscr{K}_n = \{i \in [\![1, r]\!],\ J_{i,n} \cap \mathscr{H}_i \neq \emptyset\}$.

Definition 4.2.1. La **mesure empirique informée** est définie de la manière suivante :

1. Si $\mathscr{K}_n \neq \varnothing$, alors on sélectionne seulement les contraintes suivantes $g = (g_i)_{i\in\mathscr{K}_n}$ et on injecte cela. La mesure informée est
$$\mathbb{P}_n^I = \mathbb{P}_n^{I_n}.$$
2. Sinon on injecte aucune de ces contraintes g_i et la mesure informée est donc la mesure empirique $\mathbb{P}_n$.

Posons $\mathscr{K}^* = \{i \in [\![1,r]\!],\ \{Pg_i\} \cap \mathscr{H}_i = \{Pg_i\}\}$. Remarquons qu'on ne peut pas calculer cet ensemble car on n'a pas accès aux Pg. Pour cette mesure empirique informée, on a un résultat du type **Donsker**.

Théorème 4.2.5. *Si $\mathscr{K}^* \neq \varnothing$ alors*
$$\sqrt{n}(\mathbb{P}_n^I - P) \Rightarrow G_{I^*} \quad dans \quad l^\infty(\mathscr{F})$$
où I^* est donnée par $(Pg_i)_{i\in\mathscr{K}^*}$. Si $I^* = \varnothing$ (i.e $\mathscr{K}^* = \varnothing$) alors G_I est le P-Pont-Brownien standard.

Démonstration (du théorème).
Puisque p.s. à partir d'un certain rang pour tout $i \in [\![1,r]\!]$, $Pg_i \in J_{i,n}$ et $J_{i,n} \xrightarrow{p.s.} \{Pg_i\}$ alors p.s. il existe $N > 0$ tel que pour tout $n \geq N$ et pour tout $i \in [\![1,r]\!]$ on a $\mathscr{H}_i \cap J_{i,n} = \mathscr{H}_i \cap \{Pg_i\}$. En effet, cela vient du fait que l'ensemble des maxima locaux stricts $\mathscr{H}_i$ est au plus dénombrable pour une fonction réelle :

Lemme 4.2.6. *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction réelle. Alors l'ensemble des maxima locaux stricts $\mathscr{H}$ est au plus dénombrable.*

Démonstration (du lemme).
Posons $B_\delta = \{x \in \mathbb{R},\ \forall y \in \mathbb{R}\ |x-y| < \delta \Rightarrow f(y) < f(x)\}$. Montrons que B_δ est au plus dénombrable. Puisqu'il y a au plus un élément dans $B_\delta \cap [k\delta/2, (k+1)\delta/2]$ par définition de B_δ et que $\bigcup_{k\in\mathbb{Z}}[k\delta/2, (k+1)\delta/2] = \mathbb{R}$, alors

$B_\delta = \bigcup_{k\in\mathbb{Z}} B_\delta \cap [k\delta/2, (k+1)\delta/2]$ est au plus dénombrable comme réunion dénombrable d'ensembles au plus dénombrables. En notant que $\mathscr{H} = \bigcup_{n\in\mathbb{N}^*} B_{1/n}$, on en déduit que $\mathscr{H}$ est au plus dénombrable. □

D'où pour tout $n \geq N$, $\mathscr{K}_n = \mathscr{K}^*$ et donc $\mathbb{P}_n^I = \mathbb{P}_n^{I^*}$. Par le théorème 3, on a la convergence souhaitée. □

5

Concentration de la mesure empirique informée

Dans ce chapitre, on s'intéresse à la concentration de la mesure empirique informée $\mathbb{P}_n^I$ autour P. Nous montrons que le processus empirique informé est plus concentré autour de 0 *que le processus empirique classique pour une information auxiliaire générale. De plus, nous donnons une borne de concentration plus fine que la borne usuelle donnée par l'inégalité de Bernstein dans le cas d'une information auxiliaire forte donnée par des espérances.*

5.1 Concentration du processus empirique informé et application

L'objectif de cette section est de montrer que le processus empirique informé se concentre mieux que celui du processus empirique non informé pour n assez grand. De plus, nous montrons aussi que le supremum du processus limite informé est plus concentré que son analogue non informé. Tout d'abord précisons le cadre. Soit $X_1, \cdots, X_n$ un échantillon de taille $n \in \mathbb{N}^*$ *i.i.d.* de loi P inconnue. On suppose qu'on dispose d'une information auxiliaire I générale et d'une classe de fonctions $\mathscr{F} \subset L^2(P)$. On note α_n et G (resp. $\alpha_{n,I}$ et G_I) le processus empirique non informé et son processus limite indexés par $\mathscr{F}$ (resp. le processus empirique informé et son processus limite). Par définition du processus limite informé, on a l'inégalité suivante vérifiée pour tout $f \in \mathscr{F} \subset L^2(P)$

$$Var_P\left(G_I(f)\right) \leq Var_P\left(G(f)\right).$$

Le résultat suivant énonce que pour n suffisamment grand, $\alpha_n^I(f)$ est plus concentré autour de 0 que ne l'est $\alpha_n(f)$ pour f fixée.

Théorème 5.1.1. *Soit $\mathscr{F} \subset L^2(P)$ une classe de fonctions mesurables et de carré intégrable par rapport à P. Alors pour tout $f \in \mathscr{F}$ et $\lambda > 0$,*

$$\mathbb{P}\left(|G_I(f)| > \lambda\right) \leq \mathbb{P}\left(|G(f)| > \lambda\right).$$

De plus si $Var_P\left(G_I(f)\right) < Var_P\left(G(f)\right)$ *alors*

$$\mathbb{P}\left(|G_I(f)| > \lambda\right) < \mathbb{P}\left(|G(f)| > \lambda\right)$$

et il existe un rang $N > 0$ *tel que pour tout* $n \geq N$,

$$\mathbb{P}\left(|\alpha_{n,I}(f)| > \lambda\right) < \mathbb{P}\left(|\alpha_n(f)| > \lambda\right).$$

Démonstration.
Soient $\lambda \in \mathbb{R}_+^*$ *et* $f \in \mathscr{F}$ *une fonction mesurable de carré intégrable par rapport à* P. *Alors par convergence en loi, on a*

$$\mathbb{P}\left(|\alpha_{n,I}(f)| > \lambda\right) \xrightarrow[n\to+\infty]{} \mathbb{P}\left(|G_I(f)| > \lambda\right),$$
$$\mathbb{P}\left(|\alpha_n(f)| > \lambda\right) \xrightarrow[n\to+\infty]{} \mathbb{P}\left(|G(f)| > \lambda\right).$$

Notons $\sigma_1 := \sqrt{Var_P G_I(f)} < \sqrt{Var_P G(f)} =: \sigma_2$. *Remarquons que*

$$\mathbb{P}\left(|G_I(f)| > \lambda\right) = \mathbb{P}\left(|\mathscr{N}(0,1)| > \frac{\lambda}{\sigma_1}\right) < \mathbb{P}\left(|\mathscr{N}(0,1)| > \frac{\lambda}{\sigma_2}\right) = \mathbb{P}\left(|G(f)| > \lambda\right).$$

Ainsi il existe un rang $N > 0$ *à partir duquel pour tout* $n \geq N$,

$$\mathbb{P}\left(|\alpha_{n,I}(f)| > \lambda\right) < \mathbb{P}\left(|\alpha_n(f)| > \lambda\right).$$

□

Le théorème suivant établit que le supremum du processus limite informé pour une classe de fonctions est plus concentré que son analogue non informé.

Théorème 5.1.2. *Soit* $\mathscr{F} \subset L^2(P)$ *une classe* P*-Donsker. On suppose que* $\mathscr{F}$ *est séparable point par point. Alors pour tout* $\lambda > 0$

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} |G_I(f)| > \lambda\right) \leq \mathbb{P}\left(\sup_{f\in\mathscr{F}} |G(f)| > \lambda\right).$$

De plus, s'il existe $\lambda > 0$ *tel que*

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} |G_I(f)| > \lambda\right) < \mathbb{P}\left(\sup_{f\in\mathscr{F}} |G(f)| > \lambda\right)$$

alors il existe un rang $N > 0$ *tel que pour tout* $n \geq N$,

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} |\alpha_{n,I}(f)| > \lambda\right) < \mathbb{P}\left(\sup_{f\in\mathscr{F}} |\alpha_n(f)| > \lambda\right).$$

Démonstration.
Afin de démontrer le théorème, nous avons besoin du résultat suivant dû à Slepian, Fernique, Marcus, Shepp (on peut le retrouver dans [38] à la page 441*) :*

Lemme 5.1.3. *Soient X et Y deux processus gaussiens centrés et séparables indexés par un ensemble T. On suppose que*

$$\mathbb{E}(X_s - X_t)^2 \leq \mathbb{E}(Y_s - Y_t)^2 \quad \textit{pour tout } s, t \in T.$$

Alors pour tout $\lambda > 0$

$$\mathbb{P}\left(\sup_{t\in T} X_t > \lambda\right) \leq \mathbb{P}\left(\sup_{t\in T} Y_t > \lambda\right).$$

Posons $\widetilde{\mathscr{F}} = \mathscr{F} \cup (-\mathscr{F})$ la classe symétrisée de $\mathscr{F}$ et remarquons que

$$\begin{aligned}
\sup_{f\in\mathscr{F}} |G_I(f)| &= \sup_{f\in\mathscr{F}} \max\big(G_I(f), -G_I(f)\big) \\
&= \sup_{f\in\mathscr{F}} \max\big(G_I(f), G_I(-f)\big) \\
&= \sup_{f\in\widetilde{\mathscr{F}}} G_I(f).
\end{aligned}$$

De même $\sup_{f\in\mathscr{F}} |G(f)| = \sup_{f\in\widetilde{\mathscr{F}}} G(f)$. *Vérifions maintenant la condition. Soient $f, h \in \widetilde{\mathscr{F}}$,*

$$\begin{aligned}
\mathbb{E}\big(G_I(f) - G_I(h))\big)^2 &= \mathbb{E}\big(G_I(f-h))\big)^2 \\
&= Var\big(G_I(f-h)\big) \\
&\leq Var\big(G(f-h)\big) \\
&= \mathbb{E}\big(G(f) - G(h))\big)^2.
\end{aligned}$$

Par le lemme technique précédent 5.1.3, on obtient que

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathscr{F}} |G_I(f)| > \lambda\right) &= \mathbb{P}\left(\sup_{f\in\widetilde{\mathscr{F}}} G_I(f) > \lambda\right) \\
&\leq \mathbb{P}\left(\sup_{f\in\widetilde{\mathscr{F}}} G(f) > \lambda\right) \\
&= \mathbb{P}\left(\sup_{f\in\mathscr{F}} |G(f)| > \lambda\right).
\end{aligned}$$

Ainsi s'il existe $\lambda > 0$ tel que

$$\mathbb{P}\left(\sup_{f\in\widetilde{\mathscr{F}}} G_I(f) > \lambda\right) < \mathbb{P}\left(\sup_{f\in\widetilde{\mathscr{F}}} G(f) > \lambda\right)$$

et en utilisant le fait que l'application $X \mapsto \|X\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} |X(f)|$ *(pour* $X \in l^{\infty}(\mathscr{F})$*) est continue et par convergence en loi, alors il existe un rang* $N > 0$ *à partir duquel pour tout* $n \geq N$*,*

$$\mathbb{P}\left(\|\alpha_{n,I}\|_{\mathscr{F}} > \lambda\right) = \mathbb{P}\left(\sup_{f\in\mathscr{F}} |\alpha_{n,I}(f)| > \lambda\right) < \mathbb{P}\left(\sup_{f\in\mathscr{F}} |\alpha_n(f)| > \lambda\right) = \mathbb{P}\left(\|\alpha_n\|_{\mathscr{F}} > \lambda\right).$$

□

On peut en déduire un corollaire portant sur les fonctions quantiles des suprema.

Corollaire 5.1.4. *Soit* $\mathscr{F} \subset L^2(P)$ *une classe* P*-Donsker. On suppose que* $\mathscr{F}$ *est séparable point par point. Notons* F_1 *(respectivement* F_2*) la fonction de répartition de* $\sup_{f\in\mathscr{F}} |G(f)|$ *(respectivement* $\sup_{f\in\mathscr{F}} |G_I(f)|$*). Alors les fonctions quantiles sont aussi ordonnées*

$$\forall \alpha \in [0,1[, \quad F_2^{-1}(\alpha) \leq F_1^{-1}(\alpha)$$

où pour $i \in [\![1,2]\!]$*,* $F_i^{-1}(\alpha) = \inf\{t \in \mathbb{R}_+,\ F_i(t) \geq \alpha\}$*.*

Démonstration.
Remarquons que par le théorème 7, on a pour tout $t \in \mathbb{R}_+$*,*

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathscr{F}} |G_I(f)| > t\right) \leq \mathbb{P}\left(\sup_{f\in\mathscr{F}} |G(f)| > t\right)\\
&\Leftrightarrow 1 - F_2(t) \leq 1 - F_1(t)\\
&\Leftrightarrow F_1(t) \leq F_2(t).
\end{aligned}$$

Soit $\alpha \in [0,1[$*. Alors*

$$F_2\left(F_1^{-1}(\alpha)\right) \geq F_1\left(F_1^{-1}(\alpha)\right).$$

Puisque la fonction de répartition est continue à droite, on a

$$F_1\left(F_1^{-1}(\alpha)\right) \geq \alpha.$$

Ainsi

$$F_2^{-1}(\alpha) \leq F_1^{-1}(\alpha).$$

□

Ce corollaire a des applications intéressantes dont l'amélioration du test de Kolmogorov-Smirnov. On pourra consulter à ce sujet la thèse de *M. Albertus* [4] à la page 34 pour le cas particulier d'une information auxiliaire donnée par la connaissance des probabilités d'une partition.

5.2 Borne de concentration pour le processus empirique informé

Dans cette section, on souhaite établir une borne de concentration pour le processus empirique informé (dans le cas d'une information auxiliaire donnée par des espérances) qui soit meilleure asymptotiquement que la borne usuelle de type Bernstein. Le cadre est le suivant. Soit $X_1, \cdots, X_n$ un échantillon *i.i.d.* de taille $n \in \mathbb{N}^*$ et de loi P à valeurs dans $\mathscr{X}$. On suppose qu'on dispose d'une information auxiliaire I donnée par des espérances $Pg = \left(Pg_1, \cdots, Pg_m\right)^T$ où $g = (g_1, \cdots, g_m)^T$ est un vecteur de fonctions intégrables et $m \in \mathbb{N}^*$. Cette information auxiliaire I est donc une information auxiliaire forte. Sans perte de généralité, on supposera que $Pg = 0$. Notons pour tout $n \in \mathbb{N}^*$, $\alpha_{n,I}$ et α_n le processus empirique informé et non informé respectivement. Pour le processus empirique non informé, *l'inégalité de Bernstein* nous donne une borne de concentration exponentielle pour le processus empirique non informé :

Lemme 5.2.1. *Pour tout $x > 0$ et pour toute fonction f mesurable bornée sur le support de la variable aléatoire X notée $supp(X)$, on a*

$$\mathbb{P}\left(|\mathbb{P}_n f - Pf| > x\right) \leq 2\exp\left(-\frac{1}{4}\frac{nx^2}{Pf^2 + x\|f\|_\infty}\right) := \varphi_n(x),$$

$$\mathbb{P}\left(|\alpha_n(f)| > x\right) \leq 2\exp\left(-\frac{1}{4}\frac{x^2}{Pf^2 + \frac{x\|f\|_\infty}{\sqrt{n}}}\right) = \varphi_n\left(\frac{x}{\sqrt{n}}\right)$$

où $\|f\|_\infty = \sup_{x \in supp(X)} |f(x)|$.

Pour une fonction f mesurable bornée sur le support de X, on souhaite obtenir une borne de concentration pour $t > 0$

$$\mathbb{P}\left(|\mathbb{P}_n^I f - Pf| > t\right) \leq \varphi_{n,I}(t)$$

telle que pour tout $\lambda > 0$

$$\frac{\varphi_{n,I}\left(\frac{\lambda}{\sqrt{n}}\right)}{\varphi_n\left(\frac{\lambda}{\sqrt{n}}\right)} \xrightarrow[n \to +\infty]{} c^* < 1.$$

Cela permettrait d'en déduire que pour n suffisamment grand

$$\varphi_{n,I}\left(\frac{\lambda}{\sqrt{n}}\right) < \varphi_n\left(\frac{\lambda}{\sqrt{n}}\right).$$

Ainsi la borne de concentration $\varphi_{n,I}$ serait meilleure que la borne usuelle φ_n. Par le théorème 5.1.1 et 7, nous avons montré que le processus empirique informé se concentre asymptotiquement plus vite que le processus empirique non informé. Notons pour toute fonction f mesurable bornée sur

le support de X,

$$\begin{aligned}\Lambda_n(f) &= -cov_n(g,f)^T(Var_n(g))^{-1},\\ \Lambda(f) &= -cov_P(g,f)^T(Var_P(g))^{-1},\\ \mathbb{P}_n^I f &= \mathbb{P}_n f + \Lambda_n(f)\cdot\mathbb{P}_n g.\end{aligned}$$

La proposition suivante nous donne une borne de concentration pour le processus empirique informé.

Proposition 5.2.2. *Soit $f \in L^2(P)$ une fonction mesurable bornée sur le support de X. On suppose que*

- *g est bornée c'est à dire que $\|g\| = \max_{1\le i\le m}|g_i|$ est bornée sur le support de X.*
- *$cov(f,g) \neq 0$.*

Alors pour tout $\varepsilon \in]0,1[$ et pour tout $\delta \in \left]0,\frac{1}{4}\right[$, on a pour tout $t \in \mathbb{R}_+^*$

$$\begin{aligned}\varphi_{n,I}(t) = {} & 2\exp\left(-\frac{1}{4}\frac{nt^2(1-\varepsilon)^2}{P(f+\Lambda(f)\cdot g)^2 + \frac{t(1-\varepsilon)}{m}\|f+\Lambda(f)\cdot g\|_\infty}\right) + 2\sum_{j=1}^m \exp\left(-\frac{1}{4\sqrt{m}}\frac{\varepsilon t n^{1-2\delta}}{\sqrt{m}Pg_j^2 + \sqrt{\varepsilon t}\frac{\|g_j\|_\infty}{n^\delta}}\right)\\ & + 2\sum_{j=1}^m \exp\left(-\frac{1}{16}\frac{n\eta^2}{P(fg_j)^2 + \frac{\eta}{2}\|fg_j\|_\infty}\right) + 2\sum_{j=1}^m \exp\left(-\frac{1}{8}\frac{\sqrt{n}\eta}{Pg_j^2 + \sqrt{\frac{\eta}{2}}\frac{\|g_j\|_\infty}{n^{\frac{1}{4}}}}\right)\\ & + 2\exp\left(-\frac{1}{4}\frac{n\left(\sqrt{\frac{\eta}{2}}n^{\frac{1}{4}} - |Pf|\right)^2}{Pf^2 + \left(\sqrt{\frac{\eta}{2}}n^{\frac{1}{4}} - |Pf|\right)\|f\|_\infty}\right) + 4\sum_{j=1}^m \exp\left(-\frac{1}{8m}\frac{n\eta}{Pg_j^2 + \sqrt{\frac{\eta}{2m}}\|g_j\|_\infty}\right)\\ & + 2\sum_{1\le i\le j\le m} \exp\left(-\frac{1}{16m^2}\frac{n\eta^2}{P(g_ig_j)^2 + \frac{\eta}{2m}\|g_ig_j\|_\infty}\right)\end{aligned}$$

où

$$\eta := \eta_n(t) = \min\left(\frac{1}{2\|(Var_P(g))^{-1}\|}, \frac{\sqrt{\varepsilon t}n^\delta}{2\sqrt{m}\|(Var_P(g))^{-1}\|\left(1+\|(Var_P(g))^{-1}\||cov(g,f)|\right)}\right).$$

et $\|\cdot\|$ est la norme subordonnée associée à la norme infinie $|\cdot|$ de $\mathbb{R}^m$.

Remarque 5.2.3. *Remarquons que la borne de concentration $\varphi_{n,I}$ dépend des paramètres $\varepsilon \in]0,1[$ et $\delta \in \left]0,\frac{1}{2}\right[$.*

Démonstration (de la proposition).

Fixons $\varepsilon \in \left]0, 1-\sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}}\right[$, $\delta \in \left]0, \frac{1}{4}\right[$ *et* $t \in \mathbb{R}_+^*$. *Rappelons que* m *est fixé. On notera* $|\cdot|$ *et* $|\cdot|_2$ *la norme infinie et euclidienne respectivement dans un espace vectoriel de dimension finie. Remarquons que*

$$\begin{aligned}
\mathbb{P}_n^I f - Pf &= \frac{1}{\sqrt{n}}\left(\alpha_n(f) + \Lambda_n(f)\alpha_n(g)\right) \\
&= \frac{1}{\sqrt{n}}\left(\alpha_n\left(f + \Lambda(f)\cdot g\right) + (\Lambda_n(f) - \Lambda(f))\cdot \alpha_n(g)\right) \\
&= \frac{1}{\sqrt{n}}\alpha_n\left(f + \Lambda(f) g\right) + \frac{1}{\sqrt{n}}(\Lambda_n(f) - \Lambda(f))\cdot \alpha_n(g).
\end{aligned}$$

Ainsi en utilisant l'inégalité de Cauchy-Schwarz et le fait que $|\cdot|_2 \leq \sqrt{m}|\cdot|$, *on a*

$$\begin{aligned}
\mathbb{P}\left(|\mathbb{P}_n^I f - Pf| > mt\right) &\leq \mathbb{P}\left(|\alpha_n\left(f + \Lambda(f)\cdot g\right)| > mt(1-\varepsilon)\sqrt{n}\right) + \mathbb{P}\left(|\left(\Lambda_n(f) - \Lambda(f)\right)\cdot \alpha_n(g)| > \varepsilon m t\sqrt{n}\right) \\
&\leq \mathbb{P}\left(|\alpha_n\left(f + \Lambda(f)\cdot g\right)| > mt(1-\varepsilon)\sqrt{n}\right) + \mathbb{P}\left(|\Lambda_n(f) - \Lambda(f)||\alpha_n(g)| > \varepsilon t\sqrt{n}\right).
\end{aligned}$$

Le premier terme se traite directement en utilisant la proposition 5.2.1. Concernant le second terme

$$\begin{aligned}
\mathbb{P}\left(|\Lambda_n(f) - \Lambda(f)||\alpha_n(g)| > \varepsilon t\sqrt{n}\right) &\leq \mathbb{P}\left(|\Lambda_n(f) - \Lambda(f)| > \sqrt{\varepsilon t} n^{\delta}\right) + \mathbb{P}\left(|\alpha_n(g)| > \sqrt{\varepsilon t} n^{\frac{1}{2}-\delta}\right) \\
&\leq \mathbb{P}\left(|\Lambda_n(f) - \Lambda(f)| > \sqrt{\varepsilon t} n^{\delta}\right) + \sum_{j=1}^{m} \mathbb{P}\left(|\alpha_n(g_j)| > \sqrt{\varepsilon t} n^{\frac{1}{2}-\delta}\right).
\end{aligned}$$

Ainsi

$$\mathbb{P}\left(|\mathbb{P}_n^I f - Pf| > mt\right) \leq (1) + (2) + (3).$$

Le troisième terme se traite aussi par la proposition 5.2.1. Il nous reste à traire le second terme à savoir $\mathbb{P}\left(|\Lambda_n(f) - \Lambda(f)| > \sqrt{\varepsilon t} n^{\delta}\right)$. *Pour cela, nous avons besoin d'un second lemme technique.*

Lemme 5.2.4. *Soit* $(y, B) \in \mathbb{R}^m \times \mathscr{S}_m^{++}(\mathbb{R})$ *où* $\mathscr{S}_m^{++}(\mathbb{R})$ *est l'ensemble des matrices symétriques définies positives. On note* $|\cdot|$ *une norme sur* $\mathbb{R}^m$ *et* $\|\cdot\|$ *sa norme d'opérateur associée. Posons pour tout* $\lambda > 0$

$$\eta = \min\left(\frac{1}{2\|B^{-1}\|}, \frac{\lambda}{2\|B^{-1}\|(1 + \|B^{-1}\||y|)}\right),$$

alors on a

$$\forall (x, A) \in \mathbb{R}^m \times \mathscr{S}_m^{++}(\mathbb{R}),\ \left(\|(x, A) - (y, B)\|_\infty \leq \eta \implies \left|x^T A^{-1} - y^T B^{-1}\right| \leq \lambda\right).$$

Démonstration (du lemme).
Soit $(x, A) \in \mathbb{R}^k \times \mathscr{S}_k^{++}(\mathbb{R})$. *Tout d'abord remarquons que*

$$\begin{aligned}
\left|x^T A^{-1} - y^T B^{-1}\right| &= \left|x^T A^{-1} - y^T A^{-1} + y^T A^{-1} - y^T B^{-1}\right| \\
&\leq \|A^{-1}\||x - y| + |y|\|A^{-1} - B^{-1}\|.
\end{aligned}$$

Or

$$\begin{aligned}\|A^{-1}-B^{-1}\| &= \|B^{-1}(B-A)A^{-1}\| \\ &\leq \|B^{-1}\|\|A^{-1}\|\|A-B\|.\end{aligned}$$

Il nous suffit donc de borner $\|A^{-1}\|$. *Pour tout* $z \in \mathbb{R}^k$,

$$|z| = |B^{-1}Bz| \leq \|B^{-1}\||Bz|.$$

Ainsi

$$|Bz| \geq \frac{|z|}{\|B^{-1}\|}.$$

Puisque

$$\|A-B\| \leq \eta \leq \frac{1}{2\|B^{-1}\|},$$

on en déduit que

$$|(A-B)z| \leq \frac{|z|}{2\|B^{-1}\|}.$$

On peut alors minorer $|Az|$ *de la manière suivante*

$$|Az| = |Bz+(A-B)z| \geq |Bz| - |(A-B)z| \geq \frac{|z|}{\|B^{-1}\|} - \frac{|z|}{2\|B^{-1}\|} = \frac{|z|}{2\|B^{-1}\|}.$$

En remplaçant $z = A^{-1}u$ *pour* $u \in \mathbb{R}^k$, *on obtient*

$$|A^{-1}u| \leq 2\|B^{-1}\||u|.$$

Puisque z *étant quelconque,* u *est aussi quelconque. On en déduit que*

$$\|A^{-1}\| \leq 2\|B^{-1}\|.$$

Pour conclure, il nous suffit de remarquer que

$$\begin{aligned}\left|x^TA^{-1}-y^TB^{-1}\right| &\leq \|A^{-1}\||x-y|+|y|\|A^{-1}-B^{-1}\| \\ &\leq 2\|B^{-1}\||x-y|+2\|B^{-1}\|^2|y|\|A-B\| \\ &\leq 2\|(x,A)-(y,B)\|_\infty\|B^{-1}\|\left(1+\|B^{-1}\||y|\right) \\ &\leq 2\eta\|B^{-1}\|\left(1+\|B^{-1}\||y|\right) \\ &\leq \lambda.\end{aligned}$$

□

Ainsi la contraposée de l'implication du lemme précédent nous donne

$$\forall (x,A) \in \mathbb{R}^m \times \mathscr{S}_m^{++}(\mathbb{R}),\ \left(\left|x^T A^{-1} - y^T B^{-1}\right| > \lambda \implies \|(x,A)-(y,B)\|_\infty > \eta\right).$$

On peut donc appliquer le lemme 5.2.4 en posant

$$\begin{aligned}\lambda &= \sqrt{\varepsilon t} n^\delta,\\ y &= cov(g,f),\\ B &= Var_P g.\end{aligned}$$

Ainsi

$$\eta := \eta_n(t) = \min\left(\frac{1}{2\|(Var_P(g))^{-1}\|}, \frac{\sqrt{\varepsilon t} n^\delta}{2\|(Var_P(g))^{-1}\|\left(1+\|(Var_P(g))^{-1}\||cov(g,f)|\right)}\right).$$

En notant $\lambda = \sqrt{\varepsilon t} n^\delta$, on a

$$\begin{aligned}\mathbb{P}\left(|\Lambda_n(f)-\Lambda(f)|>\lambda\right) &= \mathbb{P}\left(\left|cov_n(f,g)(Var_n(g))^{-1} - cov_P(f,g)(Var_P(g))^{-1}\right| > \lambda\right)\\ &\le \mathbb{P}\left(\left\{|cov_n(f,g)-cov_P(f,g)|>\eta\right\} \cup \left\{\|var_n(g)-var_P(g)\|>\eta\right\}\right)\\ &\le \mathbb{P}\left(\left\{|cov_n(f,g)-cov_P(f,g)|>\eta\right\}\right) + \mathbb{P}\left(\left\{\|var_n(g)-var_P(g)\|>\eta\right\}\right).\end{aligned}$$

Il nous reste à étudier ces deux termes

$$\begin{aligned}\mathbb{P}\left(\left\{|cov_n(f,g)-cov_P(f,g)|>\eta\right\}\right) &= \mathbb{P}\left(\left|\frac{1}{\sqrt{n}}(\alpha_n(fg)-\mathbb{P}_n f\alpha_n(g))\right| > \eta\right)\\ &= \mathbb{P}\left(\left|\alpha_n(fg)-\mathbb{P}_n f\alpha_n(g)\right| > \sqrt{n}\eta\right)\\ &\le \mathbb{P}\left(\left|\alpha_n(fg)\right| > \sqrt{n}\frac{\eta}{2}\right) + \mathbb{P}\left(\left|\mathbb{P}_n f\right|\left|\alpha_n(g)\right| > \sqrt{n}\frac{\eta}{2}\right)\\ &\le \sum_{j=1}^m \mathbb{P}\left(\left|\alpha_n(fg_j)\right| > \sqrt{n}\frac{\eta}{2}\right) + \mathbb{P}\left(\left|\mathbb{P}_n f\right|\left|\alpha_n(g)\right| > \sqrt{n}\frac{\eta}{2}\right).\end{aligned}$$

Le premier terme avec la somme de se traite directement en utilisant la proposition 5.2.1 et le second terme de la manière suivante

$$\begin{aligned}\mathbb{P}\left(\left|\mathbb{P}_n f\right|\left|\alpha_n(g)\right| > \sqrt{n}\frac{\eta}{2}\right) &\le \mathbb{P}\left(\left|\mathbb{P}_n f\right| > n^{\frac{1}{4}}\sqrt{\frac{\eta}{2}}\right) + \mathbb{P}\left(\left|\alpha_n(g)\right| > n^{\frac{1}{4}}\sqrt{\frac{\eta}{2}}\right)\\ &\le \mathbb{P}\left(\left|\alpha_n(f)\right| > \sqrt{n}\left(n^{\frac{1}{4}}\sqrt{\frac{\eta}{2}} - |Pf|\right)\right) + \sum_{j=1}^m \mathbb{P}\left(\left|\alpha_n(g_j)\right| > n^{\frac{1}{4}}\sqrt{\frac{\eta}{2}}\right).\end{aligned}$$

Le second terme de l'inégalité précédente avec la somme se traite de même en utilisant la proposition 5.2.1. Concernant le premier terme, il est nécessaire de s'assurer que pour n suffisamment grand $n^{\frac{1}{4}}\sqrt{\frac{\eta}{2}} - |Pf| > 0$. En effet, cela est le cas puisque

$$\begin{aligned}n^{\frac{1}{4}}\sqrt{\frac{\eta_n\left(\frac{1}{\sqrt{n}}\right)}{2}} &\sim n^{\frac{1}{4}}\sqrt{\frac{\sqrt{\varepsilon}n^{\delta-\frac{1}{4}}}{4\|(Var_P(g))^{-1}\|\left(1+\|(Var_P(g))^{-1}\||cov(g,f)|\right)}}\\ &\sim n^{\frac{\delta}{2}+\frac{1}{8}}\frac{\varepsilon^{\frac{1}{4}}}{\sqrt{4\|(Var_P(g))^{-1}\|\left(1+\|(Var_P(g))^{-1}\||cov(g,f)|\right)}}.\end{aligned}$$

On traite de la même manière le cas de la variance

$$\begin{aligned}\mathbb{P}\left(\left\{\|var_n(g)-var_P(g)\|>\eta\right\}\right) &= \mathbb{P}\left(\|\alpha_n(gg^T)-\mathbb{P}_n g(\alpha_n(g))^T\|>\sqrt{n}\eta\right)\\ &\leq \mathbb{P}\left(\|\alpha_n(gg^T)\|>\sqrt{n}\frac{\eta}{2}\right)+\mathbb{P}\left(\|\mathbb{P}_n g(\alpha_n(g))^T\|>\sqrt{n}\frac{\eta}{2}\right).\end{aligned}$$

Puisque

$$\begin{aligned}\|\alpha_n(gg^T)\| &\leq m\max_{1\leq i\leq j\leq m}|\alpha_n(g_i g_j)|\\ \|\mathbb{P}_n g(\alpha_n(g))^T\| &\leq m|\mathbb{P}_n g||\alpha_n(g)|,\end{aligned}$$

on a

$$\mathbb{P}\left(\left\{\|var_n(g)-var_P(g)\|>\eta\right\}\right)\leq \sum_{1\leq i\leq j\leq m}\mathbb{P}\left(|\alpha_n(g_i g_j)|>\sqrt{n}\frac{\eta}{2m}\right)+\mathbb{P}\left(|\mathbb{P}_n g||\alpha_n(g)|>\sqrt{n}\frac{\eta}{2m}\right).$$

Le second terme se traite de la manière suivante

$$\begin{aligned}\mathbb{P}\left(\left|\mathbb{P}_n g||\alpha_n(g)|\right|>\sqrt{n}\frac{\eta}{2m}\right) &\leq \mathbb{P}\left(\left|\mathbb{P}_n g\right|>\sqrt{\frac{\eta}{2m}}\right)+\mathbb{P}\left(\left|\alpha_n(g)\right|>\sqrt{n}\sqrt{\frac{\eta}{2m}}\right)\\ &\leq \mathbb{P}\left(\left|\alpha_n(g)\right|>\sqrt{n}\sqrt{\frac{\eta}{2m}}\right)+\mathbb{P}\left(\left|\alpha_n(g)\right|>\sqrt{n}\sqrt{\frac{\eta}{2m}}\right)\\ &= 2\mathbb{P}\left(\left|\alpha_n(g)\right|>\sqrt{n}\sqrt{\frac{\eta}{2m}}\right)\\ &\leq 2\sum_{j=1}^{m}\mathbb{P}\left(\left|\alpha_n(g_j)\right|>\sqrt{n}\sqrt{\frac{\eta}{2m}}\right).\end{aligned}$$

D'où

$$\begin{aligned}\mathbb{P}\left(\left\{\|var_n(g)-var_P(g)\|>\eta\right\}\right)\leq &\sum_{1\leq i\leq j\leq m}\mathbb{P}\left(|\alpha_n(g_i g_j)|>\sqrt{n}\frac{\eta}{2m}\right)\\ &+2\sum_{j=1}^{m}\mathbb{P}\left(\left|\alpha_n(g_j)\right|>\sqrt{n}\sqrt{\frac{\eta}{2m}}\right).\end{aligned}$$

On en déduit alors

$$\begin{aligned}\mathbb{P}\left(|\mathbb{P}_n^I f-Pf|>mt\right)\leq &\mathbb{P}\left(|\alpha_n\left(f+\Lambda(f)\cdot g\right)|>mt(1-\varepsilon)\sqrt{n}\right)+\sum_{j=1}^{m}\mathbb{P}\left(|\alpha_n(g_j)|>\sqrt{\varepsilon t}n^{\frac{1}{2}-\delta}\right)\\ &+\mathbb{P}\left(|\Lambda_n(f)-\Lambda(f)|>\sqrt{\varepsilon t}n^{\delta}\right),\end{aligned}$$

et

$$\begin{aligned}\mathbb{P}\left(|\Lambda_n(f)-\Lambda(f)|>\sqrt{\varepsilon t}n^{\delta}\right)\leq &\sum_{j=1}^{m}\mathbb{P}\left(\left|\alpha_n(fg_j)\right|>\sqrt{n}\frac{\eta}{2}\right)+\mathbb{P}\left(\left|\alpha_n(f)\right|>\sqrt{n}\left(n^{\frac{1}{4}}\sqrt{\frac{\eta}{2}}-|Pf|\right)\right)\\ &+\sum_{j=1}^{m}\mathbb{P}\left(\left|\alpha_n(g_j)\right|>n^{\frac{1}{4}}\sqrt{\frac{\eta}{2}}\right)+\sum_{1\leq i\leq j\leq m}\mathbb{P}\left(|\alpha_n(g_i g_j)|>\sqrt{n}\frac{\eta}{2m}\right)\\ &+2\sum_{j=1}^{m}\mathbb{P}\left(\left|\alpha_n(g_j)\right|>\sqrt{n}\sqrt{\frac{\eta}{2m}}\right).\end{aligned}$$

En utilisant la proposition 5.2.1, on obtient

$$\begin{aligned}\varphi_{n,I}(mt) &= 2\exp\left(-\frac{1}{4}\frac{n(mt)^2(1-\varepsilon)^2}{P(f+\Lambda(f)\cdot g)^2+t(1-\varepsilon)\|f+\Lambda(f)\cdot g\|_\infty}\right)+2\sum_{j=1}^{m}\exp\left(-\frac{1}{4}\frac{\varepsilon t n^{1-2\delta}}{Pg_j^2+\sqrt{\varepsilon t}\frac{\|g_j\|_\infty}{n^\delta}}\right)\\ &+2\sum_{j=1}^{m}\exp\left(-\frac{1}{16}\frac{n\eta^2}{P(fg_j)^2+\frac{\eta}{2}\|fg_j\|_\infty}\right)+2\sum_{j=1}^{m}\exp\left(-\frac{1}{8}\frac{\sqrt{n}\eta}{Pg_j^2+\sqrt{\frac{\eta}{2}}\frac{\|g_j\|_\infty}{n^{\frac{1}{4}}}}\right)\\ &+2\exp\left(-\frac{1}{4}\frac{n\left(\sqrt{\frac{\eta}{2}}n^{\frac{1}{4}}-|Pf|\right)^2}{Pf^2+\left(\sqrt{\frac{\eta}{2}}n^{\frac{1}{4}}-|Pf|\right)\|f\|_\infty}\right)+4\sum_{j=1}^{m}\exp\left(-\frac{1}{8m}\frac{n\eta}{Pg_j^2+\sqrt{\frac{\eta}{2m}}\|g_j\|_\infty}\right)\\ &+2\sum_{1\le i\le j\le m}\exp\left(-\frac{1}{16m^2}\frac{n\eta^2}{P(g_ig_j)^2+\frac{\eta}{2m}\|g_ig_j\|_\infty}\right).\end{aligned}$$

En faisant le changement de variable $\widetilde{t}=\frac{t}{m}$*, on en déduit le résultat.* □

Énonçons à présent le résultat principal de cette section stipulant que pour $\lambda>0$ la borne de concentration informée $\varphi_{n,I}\left(\frac{\lambda}{\sqrt{n}}\right)$ pour n suffisamment grand est meilleure que celle non informé $\varphi_n\left(\frac{\lambda}{\sqrt{n}}\right)$:

Théorème 5.2.5. *Soit* $f\in L^2(P)$ *une fonction mesurable bornée sur le support de* X*. On suppose que :*

- g *est bornée c'est à dire que* $\|g\|=\max_{1\le i\le m}|g_i|$ *est bornée sur le support de* X*.*
- $cov(f,g)\neq 0$.

Notons $\varphi_{n,I}$ *la borne de concentration informée définie dans la proposition 5.2.2 et* φ_n *la borne de concentration non informée définie dans le lemme 5.2.1. Alors pour tout* $\varepsilon\in\left]0,1-\sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}}\right[$*, pour tout* $\delta\in\left]0,\frac{1}{4}\right[$ *et pour tout* $\lambda>0$*, il existe un rang* $N>0$ *tel que pour tout* $n\ge N$*,*

$$\varphi_{n,I}\left(\frac{\lambda}{\sqrt{n}}\right)<\varphi_n\left(\frac{\lambda}{\sqrt{n}}\right).$$

Démonstration (du théorème).

Tout d'abord énonçons un premier lemme technique montrant que l'intervalle $\left]0,1-\sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}}\right[$ *n'est pas vide :*

Lemme 5.2.6. *Si* $cov(f,g)\neq 0$ *alors*

$$P(f+\Lambda(f)\cdot g)^2<Pf^2.$$

Ainsi

$$\left]0, 1-\sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}}\right[\neq \emptyset.$$

Démonstration (du lemme).
Rappelons que $\Lambda(f) = -cov(g,f)^T(Var_P g)^{-1}$. *Alors*

$$P(f+\Lambda(f)\cdot g)^2 = Pf^2 + \Lambda(f)Pgg^T\Lambda(f)^T + 2\Lambda(f)Pfg.$$

Puisque $Pg = 0$, *on a*

$$\begin{aligned} Pgg^T &= Var_P g, \\ Pfg &= cov(g,f). \end{aligned}$$

Ainsi

$$\begin{aligned} P(f+\Lambda(f)\cdot g)^2 &= Pf^2 + cov(g,f)^T(Var_P g)^{-1}cov(g,f) - 2cov(g,f)^T(Var_P g)^{-1}cov(g,f) \\ &= Pf^2 - cov(g,f)^T(Var_P g)^{-1}cov(g,f). \end{aligned}$$

En remarquant que $Var_P g$ *est une matrice symétrique définie positive, on en déduit que*

$$P(f+\Lambda(f)\cdot g)^2 < Pf^2.$$

D'où

$$1-\sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}} > 0.$$

□

Il nous reste à démontrer que pour tout $\lambda > 0$ *il existe un rang* $N > 0$ *tel que pour tout pour tout* $n \geq N$,

$$\varphi_{n,I}\left(\frac{\lambda}{\sqrt{n}}\right) < \varphi_n\left(\frac{\lambda}{\sqrt{n}}\right).$$

Fixons dorénavant $\lambda > 0$ *et posons* $t_n = \frac{\lambda}{\sqrt{n}}$. *Remarquons alors*

$$\varphi_n(t_n) = 2\exp\left(-\frac{1}{4}\frac{\lambda^2}{Pf^2 + \frac{\lambda}{\sqrt{n}}\|f\|_\infty}\right) \xrightarrow[n\to+\infty]{} 2\exp\left(-\frac{1}{4}\frac{\lambda^2}{Pf^2}\right).$$

Il nous suffit donc de démontrer que $\varphi_{n,I}\left(\frac{\lambda}{\sqrt{n}}\right)$ *converge vers une limite* $l \in \mathbb{R}_+$ *telle que* $l < 2\exp\left(-\frac{1}{4}\frac{\lambda^2}{Pf^2}\right)$.
Notons pour $t \in \mathbb{R}_+$,

$$\varphi_{n,I}(t) = a_n(t) + b_n(t)$$

où

$$a_n(t) = 2\exp\left(-\frac{1}{4}\frac{nt^2(1-\varepsilon)^2}{P(f+\Lambda(f)\cdot g)^2 + \frac{t(1-\varepsilon)}{m}\|f+\Lambda(f)\cdot g\|_\infty}\right)$$

$$\begin{aligned}
b_n(t) = {} & 2\sum_{j=1}^m \exp\left(-\frac{1}{4\sqrt{m}}\frac{\varepsilon t n^{1-2\delta}}{\sqrt{m}Pg_j^2 + \sqrt{\varepsilon}t\frac{\|g_j\|_\infty}{n^\delta}}\right) \\
& + 2\sum_{j=1}^m \exp\left(-\frac{1}{16}\frac{n\eta^2}{P(fg_j)^2 + \frac{\eta}{2}\|fg_j\|_\infty}\right) + 2\sum_{j=1}^m \exp\left(-\frac{1}{8}\frac{\sqrt{n}\eta}{Pg_j^2 + \sqrt{\frac{\eta}{2}}\frac{\|g_j\|_\infty}{n^{\frac{1}{4}}}}\right) \\
& + 2\exp\left(-\frac{1}{4}\frac{n\left(\sqrt{\frac{\eta}{2}}n^{\frac{1}{4}} - |Pf|\right)^2}{Pf^2 + \left(\sqrt{\frac{\eta}{2}}n^{\frac{1}{4}} - |Pf|\right)\|f\|_\infty}\right) + 4\sum_{j=1}^m \exp\left(-\frac{1}{8m}\frac{n\eta}{Pg_j^2 + \sqrt{\frac{\eta}{2m}}\|g_j\|_\infty}\right) \\
& + 2\sum_{1\le i\le j\le m} \exp\left(-\frac{1}{16m^2}\frac{n\eta^2}{P(g_ig_j)^2 + \frac{\eta}{2m}\|g_ig_j\|_\infty}\right).
\end{aligned}$$

Il suffit donc de montrer que $a_n\left(\frac{\lambda}{\sqrt{n}}\right) \xrightarrow[n\to+\infty]{} l < 2\exp\left(-\frac{1}{4}\frac{\lambda^2}{Pf^2}\right)$ *et* $b_n\left(\frac{\lambda}{\sqrt{n}}\right) \xrightarrow[n\to+\infty]{} 0$. *Remarquons que*

$$\begin{aligned}
\frac{a_n\left(\frac{\lambda}{\sqrt{n}}\right)}{\varphi_n\left(\frac{\lambda}{\sqrt{n}}\right)} &= \exp\left(-\frac{1}{4}\left(\frac{nt_n^2(1-\varepsilon)^2}{P(f+\Lambda(f)\cdot g)^2 + \frac{t_n(1-\varepsilon)}{m}\|f+\Lambda(f)\cdot g\|_\infty} - \frac{nt_n^2}{Pf^2 + t_n\|f\|_\infty}\right)\right) \\
&\xrightarrow[n\to+\infty]{} \exp\left(-\frac{1}{4}\left(\frac{\lambda^2(1-\varepsilon)^2}{P(f+\Lambda(f)\cdot g)^2} - \frac{\lambda^2}{Pf^2}\right)\right).
\end{aligned}$$

Définissons pour $\varepsilon \in]0,1[$,

$$\psi^*(\varepsilon) := \frac{\lambda^2(1-\varepsilon)^2}{P(f+\Lambda(f)\cdot g)^2} - \frac{\lambda^2}{Pf^2}.$$

Et montrons que pour tout $\varepsilon \in \left]0, 1-\sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}}\right[$, $\psi^*(\varepsilon) > 0$. *En effet*

$$\begin{aligned}
\psi^*(\varepsilon) > 0 &\Leftrightarrow (1-\varepsilon)^2 > \frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2} \\
&\Leftrightarrow 1-\varepsilon > \sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}} \\
&\Leftrightarrow 1-\sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}} > \varepsilon.
\end{aligned}$$

Ainsi pour tout $\varepsilon \in \left]0, 1-\sqrt{\frac{P(f+\Lambda(f)\cdot g)^2}{Pf^2}}\right[$,

$$\frac{a_n\left(\frac{\lambda}{\sqrt{n}}\right)}{\varphi_n\left(\frac{\lambda}{\sqrt{n}}\right)} \xrightarrow[n\to+\infty]{} \exp\left(-\frac{1}{4}\left(\frac{\lambda^2(1-\varepsilon)^2}{P(f+\Lambda(f)\cdot g)^2} - \frac{\lambda^2}{Pf^2}\right)\right) < 1.$$

Concernant la suite $(b_n(t_n))_{n\geq 1}$*, rappelons que*

$$\eta_n(t_n) = \min\left(\frac{1}{2\|(Var_P(g))^{-1}\|}, \frac{\sqrt{\varepsilon\lambda}n^{\delta-\frac{1}{4}}}{2\sqrt{m}\|(Var_P(g))^{-1}\|\left(1+\|(Var_P(g))^{-1}\||cov(g,f)|\right)}\right).$$

Puisque $0<\delta<\frac{1}{4}$*, pour n suffisamment grand on a*

$$\eta_n(t_n) = \frac{\sqrt{\varepsilon\lambda}n^{\delta-\frac{1}{4}}}{2\sqrt{m}\|(Var_P(g))^{-1}\|\left(1+\|(Var_P(g))^{-1}\||cov(g,f)|\right)}.$$

Pour simplifier l'écriture, notons

$$C = \frac{\sqrt{\varepsilon\lambda}}{2\sqrt{m}\|(Var_P(g))^{-1}\|\left(1+\|(Var_P(g))^{-1}\||cov(g,f)|\right)} > 0.$$

Il nous reste à démontrer que :

- $t_n n^{1-2\delta} \xrightarrow[n\to+\infty]{} +\infty$ *:*

$$t_n n^{1-2\delta} = \lambda n^{\frac{1}{2}-2\delta} \xrightarrow[n\to+\infty]{} +\infty$$

  car $0<\delta<\frac{1}{4}$.
- $\sqrt{n}\eta(t_n) \xrightarrow[n\to+\infty]{} +\infty$ *:*

$$\sqrt{n}Cn^{\delta-\frac{1}{4}} = Cn^{\frac{1}{4}+\delta} \xrightarrow[n\to+\infty]{} +\infty.$$

- $\frac{n}{\sqrt{\eta(t_n)}n^{\frac{1}{4}}} \xrightarrow[n\to+\infty]{} +\infty$ *:*

$$\frac{n}{\sqrt{\eta(t_n)}n^{\frac{1}{4}}} = \frac{n}{\sqrt{Cn^{\delta-\frac{1}{4}}n^{\frac{1}{4}}}} = \frac{1}{\sqrt{C}}n^{1-\frac{\delta}{2}-\frac{1}{8}} \xrightarrow[n\to+\infty]{} +\infty$$

  car $\frac{\delta}{2}+\frac{1}{8}<\frac{1}{8}+\frac{1}{8}=\frac{1}{4}<1$.

D'où

$$b_n(t_n) \xrightarrow[n\to+\infty]{} 0.$$

On peut donc conclure que pour tout $\lambda>0$ *il existe un rang* $N>0$ *tel que pour tout pour tout* $n\geq N$*,*

$$\varphi_{n,I}\left(\frac{\lambda}{\sqrt{n}}\right) < \varphi_n\left(\frac{\lambda}{\sqrt{n}}\right).$$

□

6

Information auxiliaire fausse et adaptativité de la mesure informée

Ce chapitre a pour objectif de montrer l'impact négatif d'une information auxiliaire fausse, de détecter les informations auxiliaires fausses et de sélectionner la meilleure information auxiliaire I^* engendrée par des informations auxiliaires vraies et significatives en un certain sens.

6.1 Impact et détection d'une information auxiliaire forte fausse

Dans cette section, nous souhaitons étudier l'impact d'une information auxiliaire forte fausse dans l'estimation d'un paramètre d'intérêt. On observe un échantillon $X_1, \cdots, X_n$ *i.i.d.* de taille $n \in \mathbb{N}^*$ et de loi P. On suppose qu'on dispose d'une information auxiliaire fausse notée I^F. Plus précisément l'information auxiliaire fausse est donnée par une approximation $\beta \in \mathbb{R}^m$ de Pg avec $m \geq 1$ et $g : \mathscr{X} \to \mathbb{R}^m$ une fonction intégrable. La mesure empirique informée par I^F est donnée pour toute fonction f mesurable de carré intégrable par

$$\mathbb{P}_n^{I^F} f = \mathbb{P}_n f + \Lambda_n(f) \cdot \left(\mathbb{P}_n g - \beta\right)$$

où

$$\Lambda_n(f) = -cov_n(f,g)(Var_n g)^{-1}.$$

Notons $\Lambda(f) = -cov_P(f,g)(Var_P g)^{-1}$ la limite presque sûre de $\Lambda_n(f)$. La proposition suivante montre l'impact négatif d'une information auxiliaire fausse sur l'estimation statistique.

Proposition 6.1.1. *Soit $f \in L^2(P)$. On suppose que $\langle \Lambda(f)^T, Pg - \beta \rangle \neq 0$. Alors*

$$\mathbb{P}_n^{I^F} f \xrightarrow[n\to\infty]{p.s.} Pf + \langle \Lambda(f)^T, Pg - \beta \rangle,$$

$$\forall \alpha \in \mathbb{R}_+^*, \ \ n^\alpha \left| \mathbb{P}_n^{I^F} f - Pf \right| \xrightarrow[n\to+\infty]{p.s.} +\infty.$$

De plus pour tout $x \in \left]0, \left|\langle \Lambda(f)^T, Pg - \beta \rangle\right|\right[$,

$$\mathbb{P}\left(\left|\mathbb{P}_n^{I^F} f - Pf\right| > x\right) \xrightarrow[n\to+\infty]{} 1.$$

Démonstration.
Soit $f \in L^2(P)$. La première assertion est obtenue par simple application de la loi forte des grands nombres

$$\mathbb{P}_n^{I^F} f \xrightarrow[n\to+\infty]{p.s.} Pf + \langle \Lambda(f)^T, Pg-\beta\rangle.$$

Montrons la seconde assertion

$$\forall \alpha \in \mathbb{R}_+^*, \ \ n^\alpha \left|\mathbb{P}_n^{I^F} f - Pf\right| \xrightarrow[n\to+\infty]{p.s.} +\infty.$$

Puisque

$$\left|\mathbb{P}_n^{I^F} f - Pf\right| \xrightarrow[n\to\infty]{p.s.} \left|\langle \Lambda(f)^T, Pg-\beta\rangle\right| \neq 0,$$

on en déduit que

$$\forall \alpha \in \mathbb{R}_+^*, \ \ n^\alpha \left|\mathbb{P}_n^{I^F} f - Pf\right| \xrightarrow[n\to+\infty]{p.s.} +\infty.$$

Enfin concernant la troisième assertion. Pour $x \in \left]0, \left|\langle \Lambda(f)^T, Pg-\beta\rangle\right|\right[$, on a

$$\mathbb{P}\left(\left|\mathbb{P}_n^{I^F} f - Pf\right| > x\right) \xrightarrow[n\to+\infty]{} \mathbb{P}\left(\left|\langle \Lambda(f)^T, Pg-\beta\rangle\right| > x\right) = 1.$$

□

Remarque 6.1.2. *Remarquons que pour $m = 1$, la condition $\langle \Lambda(f)^T, Pg-\beta\rangle \neq 0$ est équivalente à*

$$cov_P(f,g) \neq 0 \text{ et } \beta \neq Pg.$$

De plus si β est très proche de Pg, la mesure empirique informée par β est relativement proche de la mesure empirique informée par Pg. En effet pour $f \in L^2(P)$,

$$\mathbb{P}_n^I f - \mathbb{P}_n^{I_F} = cov_n(f,g)\left(Var_n g\right)^{-1}(Pg-\beta).$$

Illustrons cela par quelques simulations.

Exemple 6.1.1. On génère $n = 100$ variables aléatoires *i.i.d.* de loi $P = \mathcal{N}(0,1)$. Notons l'information auxiliaire I donnée par Pg avec $g(x) = (x, x^2)$, $x \in \mathbb{R}$ et on suppose qu'on dispose d'une information auxiliaire fausse I^F donnée par $(\theta, 1)$ où $\theta \in \left\{\frac{1}{2}, \frac{2}{100}\right\}$. On trace sur un même graphique $\mathbb{F}_{n,I}$ (en verte), $\mathbb{F}_{n,I^F}$ (en magenta), $\mathbb{F}_n$ (en rouge) et F la fonction de répartition de la loi P (en bleue) pour 20

réalisations (Figure 6.1). On remarque que lorsque $\theta = \frac{1}{2}$, le faisceau en magenta s'écarte fortement de la fonction de répartition F. Cependant lorsque θ se rapproche de 0 (ici $\theta = \frac{1}{50}$), le faisceau en magenta est relativement proche de F. De ce fait, injecter une information auxiliaire fausse et proche de la vérité peut être bénéfique pour les petits échantillons.

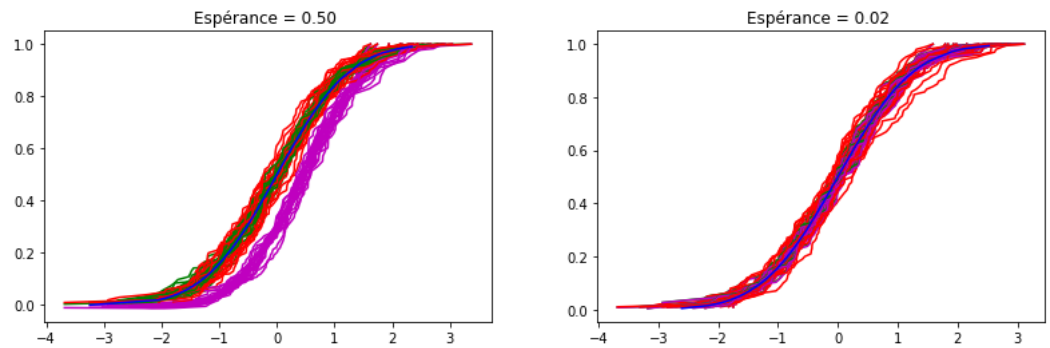

FIGURE 6.1 – Comparaison entre $\mathbb{F}_{n,I}$ (verte), $\mathbb{F}_n$ (magenta), $\mathbb{F}_n$ (rouge) et F (bleue).

On souhaite à présent détecter une information auxiliaire fausse. Nous avons à notre disposition une information auxiliaire donnée par $\beta \in \mathbb{R}^m$ qui serait la valeur de Pg avec $g : \mathscr{X} \to \mathbb{R}^m$ une fonction intégrable et $m \in \mathbb{N}^*$. Ainsi il se peut qu'il existe $k \in [\![1, m]\!]$ tel que $\beta_k \neq Pg_k$. Pour éviter cela, nous allons effectuer m tests statistiques pour les hypothèses suivantes

$$H_0^k : \ Pg_k = \beta_k \text{ contre } H_1^k : \ Pg_k \neq \beta_k$$

avec $k \in [\![1, m]\!]$. Nous avons donc affaire un problème de tests multiples. Pour chaque $k \in [\![1, m]\!]$, on pose

$$Z_{k,n} = \frac{\sqrt{n}|\mathbb{P}_n g_k - \beta_k|}{\sqrt{Var_n g_k}},$$
$$T_{k,n} = 1_{Z_{k,n} > z_{1-\alpha}}$$

où $z_{1-\alpha}$ est le quantile d'ordre $\alpha \in [0,1]$ de $|Y|$ où Y suit une loi normale centrée réduite. On obtient alors m tests statistiques asymptotiques de niveau $\alpha \in [0,1]$. Ainsi pour chaque test statistique, on peut calculer sa p-valeur notée p_k pour $k \in [\![1, m]\!]$. On pose $p = (p_1, \cdots, p_m)$ et

$$R(p) = \left\{k \in [\![1, m]\!], \ p_k \leq s(p)\right\}$$

où $s(p)$ est un seuil pouvant dépendre de p. $R(p)$ représente l'ensemble des hypothèses nulles à rejeter. Pour la procédure de Bonferroni, la fonction seuil est constante et égale à $s(\cdot) = \frac{\alpha}{m}$. D'autres procédures adaptatives existent comme par exemple la procédure de Holm. On pourra consulter à ce sujet [10],[11], [35] et [26].

6.2 Adaptativité de la mesure informée par une information auxiliaire forte pour l'estimation de Pf

Soit $X_1, \cdots, X_n$ un échantillon de taille $n \in \mathbb{N}^*$ *i.i.d.* de loi P inconnue. On dispose d'une information auxiliaire I portant sur Pg où $g = (g_1, \cdots, g_r)$ est une fonction donnée avec $r \geq 1$. Cette information auxiliaire I est engendrée par r informations auxiliaires $I_1, \cdots, I_r$ portant respectivement sur $Pg_1, \cdots, Pg_r$. On suppose que ces informations auxiliaires $I_1, \cdots, I_r$ sont vraies. On souhaite estimer au mieux un paramètre du type

$$Pf = \int_{\mathscr{X}} f dP$$

pour une fonction $f : \mathscr{X} \to \mathbb{R}$ intégrable donnée. L'objectif est de sélectionner l'information auxiliaire I^* engendrée par les informations auxiliaires informatives c'est à dire celles qui améliorent l'estimation de Pf.

Sans perte de généralité, on peut supposer que $Pg = 0$ et on supposera aussi que $\Sigma = Var_P g$ est définie positive. On note Γ l'ensemble des informations engendrées par une partie des r informations auxiliaires

$$\Gamma = \left\{ I_J = \{I_k\}_{k \in J} \ , \ J \subset [\![1, r]\!] \right\}.$$

Remarquons que $\varnothing \subset \{1, \cdots, K\}$, $\{I_k\}_{k \in \varnothing} = \varnothing$ et $\mathbb{P}_n^{\varnothing} = \mathbb{P}_n$. On notera $g_J = (g_k)_{k \in J}$ pour tout $J \subset [\![1, r]\!]$.

Nous avons alors à notre disposition les estimateurs $(\mathbb{P}_n^I f)_{I \in \Gamma}$ et on souhaite choisir celui qui approche le mieux Pf. Pour tout $I_J \in \Gamma$ avec $J \subset [\![1, r]\!]$ telles que $Var_P g_J > 0$, on a par le théorème 3.1.2,

$$\sqrt{n}(\mathbb{P}_n^{I_J} f - Pf) \Rightarrow G_{I_J}(f)$$

où G_{I_J} est un P-pont brownien informé par I de matrice de covariance

$$Var\ G_{I_J}(f) = Var_P f - Cov_P(g_J, f)^T (Var_P g_J)^{-1} Cov_P(g_J, f).$$

De plus par la proposition 3.1.4, on a

$$Var\ G_{I_J}(f) \leq Var\ G(f) = Var_P f.$$

Definition 6.2.1. L'information auxiliaire $I_J \in \Gamma$ avec $J \subset [\![1, r]\!]$ est dite **informative** pour f si $Var_P g_J > 0$ et

$$Var\ G_{I_J}(f) < Var\ G(f).$$

On note Γ^* l'ensemble des informations auxiliaires informatives.

Observons que pour tout $I_J \in \Gamma^*$ avec $J \subset [\![1,r]\!]$,

$$\begin{aligned} Var\ G_{I_J}(f) < Var\ G(f) &\Leftrightarrow Cov_P(g_J,f)^T Var_P g_J^{-1} Cov_P(g_J,f) > 0 \\ &\Leftrightarrow cov_P(g_J,f) \neq 0 \end{aligned}$$

car $\Sigma_J := Var_P g_J > 0$. Puisque $Pg_J = 0$ et $cov_P(f,g_J) = Pfg_J - PfPg_J$ alors $cov_P(f,g_J) = \langle f,g\rangle_{L^2(P)}$ où $\langle f,g_J\rangle_{L^2(P)} := \left(\langle f,g_k\rangle_{L^2(P)}\right)_{k\in J}$. Ainsi une information auxiliaire $I_J \in \Gamma$ avec $J \subset [\![1,r]\!]$ est informative pour f si $Var_P g_J > 0$ et f **est orthogonale à** g dans $L^2(P)$.

Dans un premier temps, nous voulons tester si l'information auxiliaire totale $I_{[\![1,r]\!]} = \{I_1,\cdots,I_r\}$ est informative pour f. Cela nous amène à poser le problème de test suivant

$$\mathscr{H}_0 : \langle f,g\rangle_{L^2(P)} = 0 \ \text{ contre } \ \mathscr{H}_1 : \langle f,g\rangle_{L^2(P)} \neq 0.$$

La proposition suivante établit la statistique de test permettant d'aboutir à une règle de décision.

Proposition 6.2.1. *On note pour tout $n \in \mathbb{N}^*$, $\Sigma_{n,f,g} = (cov_{P_n}(fg_i, fg_j))_{i,j\in[\![1,r]\!]}$ et $\Sigma_{f,g} = (cov_P(fg_i, fg_j))_{i,j\in[\![1,r]\!]}$. On suppose que $\Sigma_{f,g}$ est inversible. On pose pour tout $n \in \mathbb{N}^*$*

$$W_{n,f,g} = n(\mathbb{P}_n fg)^T \Sigma_{n,f,g}^{-1} \mathbb{P}_n fg.$$

Alors

$$\begin{aligned} &\text{Sous } \mathscr{H}_0,\ W_{n,f,g} \Rightarrow \chi_r^2, \\ &\text{Sous } \mathscr{H}_1,\ W_{n,f,g} \xrightarrow[n\to+\infty]{p.s.} +\infty. \end{aligned}$$

Remarque 6.2.2.

- *En posant pour tout $n \in \mathbb{N}^*$ et $\alpha \in]0,1[$, $T_n = 1_{W_{n,f,g} > q_{1-\alpha}}$ où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi du χ_r^2 ayant r degrés de liberté, on obtient donc un test asymptotique de niveau α.*
- *Dès que $Var_P g_J > 0$ et Σ_{f,g_J} est inversible pour $J \subset [\![1,r]\!]$, on peut remplacer g par g_J pour résoudre le problème de test suivant*

$$\mathscr{H}_0 : \langle f,g_J\rangle_{L^2(P)} = 0 \ \text{ contre } \ \mathscr{H}_1 : \langle f,g_J\rangle_{L^2(P)} \neq 0.$$

Démonstration.
Puisque $\Sigma_{f,g}$ est inversible, alors p.s. à partir d'un certain rang $\Sigma_{n,f,g}$ est aussi inversible. Puisque ces matrices sont symétriques et définies positives, on peut définir la racine carrée de la matrice inverse de $\Sigma_{n,f,g}$ et de $\Sigma_{f,g}$. Posons pour tout $n \in \mathbb{N}^*$,

$$Z_n = \Sigma_{n,f,g}^{-1/2} \sqrt{n} \mathbb{P}_n fg.$$

Remarquons que $W_{n,f,g} = \|Z_n\|^2$ *et sous* $\mathscr{H}_0$,

$$Z_n \Rightarrow \mathscr{N}_r(0, Id).$$

On en déduit alors que sous $\mathscr{H}_0$, $W_{n,f,g} \Rightarrow \chi^2_r$. *Puisque* $\mathbb{P}_n fg \xrightarrow[n\to+\infty]{p.s.} Pfg$ *alors on en déduit aussi que sous* $\mathscr{H}_1$, $W_{n,f,g} \xrightarrow[n\to+\infty]{p.s.} +\infty$. □

Pour des raisons computationnelles, on souhaite éliminer les coordonnées du vecteur g qui sont orthogonales à f dans $L^2(P)$. Pour des raisons algorithmiques, on se donne un seuil de tolérance $\varepsilon > 0$ et on s'intéresse aux problèmes de tests suivants donnés pour tout $i \in [\![1, r]\!]$,

$$\mathscr{H}_{0,i} : |\langle f, g_i\rangle_{L^2(P)}| < \varepsilon \text{ contre } \mathscr{H}_{1,i} : |\langle f, g_i\rangle_{L^2(P)}| > \varepsilon.$$

Posons pour chaque $i \in [\![1, r]\!]$ et pour tout $n \in \mathbb{N}^*$,

$$V_{i,n} = \frac{\sqrt{n}((\mathbb{P}_n fg_i)^2 - \epsilon^2)}{2|\mathbb{P}_n fg_i|\sqrt{Var_n fg_i}},$$
$$L_{i,n} = 1_{V_{i,n} > z_{1-\alpha}}$$

où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi normale centrée réduite $\mathscr{N}(0,1)$. Définissons aussi l'ensemble $\mathscr{K}_n = \{i \in [\![1, r]\!],\ L_{i,n} = 1\}$.

Definition 6.2.2. L'estimateur adaptatif informé de Pf notée $\mathbb{P}_n^{I_n} f$ est défini en suivant pour tout $n \in \mathbb{N}^*$ la procédure suivante :

- Si $\mathscr{K}_n \neq \emptyset$ alors on retient les coordonnées $\tilde{g}_n = (g_i)_{i\in\mathscr{K}_n}$ et on injecte l'information auxiliaire $I_{\mathscr{K}_n}$. Dans ce cas, $\mathbb{P}_n^{I_n} f := \mathbb{P}_n^{I_{\mathscr{K}_n}} f$.
- Sinon $\mathbb{P}_n^{I_n} f := \mathbb{P}_n f$.

La proposition suivante nous donne un théorème limite pour l'estimateur adaptatif informé $\mathbb{P}_n^{I_n} f$.

Proposition 6.2.3. *Soit* $\varepsilon \notin \{Pfg_1, \cdots, Pfg_r\}$ *et posons* $\mathscr{K}_\varepsilon = \{i \in [\![1, r]\!],\ |\langle f, g_i\rangle_{L^2(P)}| > \varepsilon\}$. *Alors*

$$\sqrt{n}(\mathbb{P}_n^{I_n} f - Pf) \Rightarrow G_{I_{\mathscr{K}_\varepsilon}} f.$$

Remarque 6.2.4. *D'un point de vue méthodologique, nous pouvons choisir* ε *en utilisant les estimations* $\mathbb{P}_n fg_i$ *pour tout* $i \in [\![1, r]\!]$ *afin que* $\varepsilon \notin \{Pfg_1, \cdots, Pfg_r\}$.

Démonstration.
Tout d'abord remarquons que pour tout $i \in [\![1,r]\!]$, on a

$$\text{Sous } \mathscr{H}_{0,i},\ L_{i,n} \xrightarrow{p.s.} 0,$$
$$\text{Sous } \mathscr{H}_{1,i},\ L_{i,n} \xrightarrow{p.s.} 1$$

car $\varepsilon \notin \{Pfg_1, \cdots, Pfg_r\}$. De plus puisque pour tout $i \in [\![1,r]\!]$, $L_{i,n}$ est une indicatrice, alors p.s. à partir d'un certain rang cette indicatrice est constante. Ainsi p.s. à partir d'un certain rang on a $\mathscr{K}_n = \mathscr{K}_\varepsilon$. On en déduit que

$$\sqrt{n}(\mathbb{P}_n^{I_n} f - Pf) \Rightarrow G_{I_{\mathscr{K}_\varepsilon}} f.$$

□

Exemple 6.2.1. Nous générons à l'aide d'une simulation $n = 150$ données *i.i.d.* de loi $P := \mathscr{N}(0,1)$. L'observateur dispose uniquement de l'information auxiliaire forte donnée par Pg où

$$g(x) := (g_1(x), g_2(x), g_3(x)) := (x^3, x^2, x^4)$$

avec $x \in \mathbb{R}$. Pour tout $k \in [\![1,3]\!]$, on note I_k l'information auxiliaire apportée par Pg_k. L'observateur souhaite estimer Pf avec $f(x) := x$, $x \in \mathbb{R}$. Il applique le test statistique de la définition 6.2.2 en calculant $\mathscr{K}_m$ pour les $m \in \{10, 20, 30, \cdots, 150\}$ premières données. Nous prenons $\alpha = \frac{5}{100}$ et $\varepsilon = 10^{-3}$. Notons que plus ε est grand, plus le risque est important de rejeter des informations auxiliaires réellement informatives. On observe qu'aucune des informations auxiliaires I_1, I_2, I_3 est informative pour m petit (généralement $m \leq 40$) et à partir d'un certain rang ($m = 50$ en général) I_1 est la seule information auxiliaire informative. Cela est conforme au fait que $cov_P(f, g_2) = cov_P(f, g_3) = 0$ et $cov_P(f, g_1) = 3$.

6.3 Adaptativité de la mesure informée au temps n par une information auxiliaire générale en apprentissage

On observe un échantillon d'apprentissage $D_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ *i.i.d.* de taille $n \in \mathbb{N}^*$ et de loi P inconnue. Plaçons nous dans le cadre où nous disposons de K informations auxiliaires $I_1, ..., I_K$ pour lesquelles nous ne connaissons pas forcément leur véracité. On note Γ l'ensemble des informations engendrées par une partie des K informations auxiliaires

$$\Gamma = \left\{ I_J = \{I_k\}_{k \in J}\ ,\ J \subset \{1, \cdots, K\} \right\}.$$

Remarquons que $\varnothing \subset \{1, ..., K\}$, $\cup_{k \in \varnothing} I_k = \varnothing$ et $\mathbb{P}_n^{\varnothing} = \mathbb{P}_n$.

Dans cette sous-section, on souhaite classer les informations combinées pertinentes pour une fonction statistique d'intérêt et rejeter les informations qui font moins bien que la mesure empirique au temps n. Les fausses informations se trouveront dans cette dernière catégorie. Ce classement se

fera par **validation croisée**. Plus précisément, nous supposons que nous avons à notre disposition une collection de règles d'apprentissage plus ou moins informées $(\hat{f}_I)_{I\in\Gamma}$.

Exemple 6.3.1.
- Dans le cas de la **régression** $\beta_n^I = \left(\mathbb{P}_n^I xx^T\right)^{-1}\mathbb{P}_n^I yx$ basé sur l'échantillon D. On pose $\hat{f}_I(x;D) = \langle x, \beta_n^I\rangle$ (cf section 7.3).
- Dans le cas de la **classification**, on peut prendre par exemple l'estimateur des k-plus proches voisins avec informations auxiliaires (cf section 7.4.5).

Un sous ensemble propre $E \subset \{1,\cdots,n\}$ est appelé **découpage** car il partitionne l'échantillon en deux sous échantillons indépendants appelés respectivement d'entraînement et de validation $D_n^E = (X_i)_{i\in E}$ et $D_n^{E^c} = (X_i)_{i\in E^c}$. Ainsi $D_n = D_n^E \cup D_n^{E^c}$. On appelle risque empirique sur le sous échantillon D_n^E

$$\mathscr{R}_n^E : f \mapsto \frac{1}{card(E)}\sum_{i\in E} c\left(f(X_i), Y_i\right)$$

où c est une fonction de coût adaptée au problème. L'estimateur par validation simple du risque de $\hat{f}_I$ pour l'échantillon D_n et le découpage E

$$\begin{aligned}\mathscr{R}_n^{val}(\hat{f}_I; D_n; E) &= \mathscr{R}_n^{E^c}(\hat{f}_I(\cdot, D_n^E)) \\ &= \frac{1}{card(E^c)}\sum_{i\in E^c} c\left(\hat{f}_I(X_i; D_n^E), Y_i\right).\end{aligned}$$

L'estimateur par validation croisée pour l'échantillon D_n et la suite de découpages $(E_j)_{1\le j\le V}$ est défini par

$$\mathscr{R}_n^{vc}(\hat{f}_I; D_n; (E_j)_{1\le j\le V}) = \frac{1}{V}\sum_{j=1}^{V}\mathscr{R}_n^{val}(\hat{f}_I; D_n; E_j).$$

Posons pour tout $I \in \Gamma$

$$\mu_{I,n} = \mathscr{R}_n^{vc}(\hat{f}_I; D_n; (E_j)_{1\le j\le V}).$$

On peut donc ordonner ces $\mu_{I,n}$ ce qui donne un classement de la pertinence des informations auxiliaires au temps n pour une fonction d'intérêt. L'information la plus pertinente au temps n est

$$I_n^* = arg\min_{I\in\Gamma}\mu_{I,n}.$$

On en déduit la définition suivante.

Definition 6.3.1. La règle d'apprentissage optimale au temps $n \in \mathbb{N}^*$ (au sens de la validation croisée) parmi une collection de règles d'apprentissage $(\hat{f}_I)_{I\in\Gamma}$ est donnée par $\hat{f}_{I_n^*}$.

Remarque 6.3.1. *Notons que pour $I = \varnothing$, $\hat{f}_I$ est l'estimateur empirique. Posons*

$$\mathcal{O}_n = \left\{ I \in \Gamma,\ \mu_{I,n} \geq \mu_{\varnothing,n} \right\}.$$

Cet ensemble représente les ***mauvaises informations*** *au temps n.*

7

Applications

L'objectif de ce chapitre est de montrer que la notion d'information auxiliaire peut être utilisée dans des contextes divers et variés. Ainsi nous proposons d'incorporer une information auxiliaire à disposition afin d'améliorer les méthodes génériques en vigueur en statistique et en simulation stochastique (Méthode de Monte-Carlo, estimation paramétrique etc). Ce chapitre est donc un recueil d'applications plus ou moins détaillées. En particulier la dernière section de chapitre donne des pistes méthodologiques d'utilisation d'une information auxiliaire en apprentissage statistique.

Tout au long de ce chapitre $\mathbb{P}_n^I$ représente la mesure empirique informée par une information auxiliaire générale I. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires *i.i.d.* de loi P. Alors on écrira la mesure empirique informée de la manière suivante

$$\mathbb{P}_n^I = \sum_{i=1}^{n} q_i \delta_{X_i}.$$

et on notera le processus empirique informé $\alpha_n^I = \sqrt{n}\left(\mathbb{P}_n^I - P\right)$.

Remarque 7.0.1. *Néanmoins notons que tous les résultats de ce chapitre sont valables pour une information auxiliaire générale (ajout de données auxiliaires, mélange de plusieurs types d'informations auxiliaires etc) du fait que nous avons un résultat de type Donsker pour chaque processus empirique informé et que la mesure empirique informée est presque sûrement à partir d'un certain rang une mesure de probabilité discrète à support fini.*

Sous certaines conditions, le processus empirique informé converge en loi (théorème 3.1.2, 3.2.3, 4.1.6 par exemple) et on notera G_I le processus limite appelé le P-pont brownien informé par l'information auxiliaire I. Par exemple :

- Dans le cas où l'information auxiliaire est donnée par une fonctionnelle $g(P)$ (ou par des estimations issues de sources indépendantes), la variance asymptotique s'exprime pour une

fonction f de carré intégrable par

$$Var(G_I(f)) = Var_P f - cov(f, \varphi_g(P))\Sigma^{-1} cov(f, \varphi_g(P))^T$$

où Σ diffère en fonction de l'information auxiliaire I.

- Dans le cas d'ajout de données auxiliaires $G_I = \lambda G$ où $0 \leq \lambda \leq 1$ et G est un P-pont brownien.

7.1 Méthode de Monte-Carlo, variables auxiliaires et données manquantes massives

7.1.1 Amélioration de la méthode de Monte-Carlo à l'aide d'une information auxiliaire

Les méthodes de Monte-Carlo sont un ensemble de procédures permettant d'estimer des quantités d'intérêt en utilisant des procédés aléatoires. Plus précisément, supposons que nous souhaitons estimer un paramètre pouvant s'écrire sous la forme d'une espérance

$$Pf = \int_{\mathscr{X}} f(x)dP$$

où f est une fonction intégrable par rapport à la loi P. La solution à ce problème consiste à générer un échantillon *i.i.d.* $\{X_1, \cdots, X_n\}$ de taille $n \in \mathbb{N}^*$ et de loi P puis à estimer l'intégrale par la moyenne empirique

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Nous proposons dans cette sous-section de rechercher une information auxiliaire pertinente afin d'améliorer cette estimation. Plus précisément nous cherchons une fonction g de carré intégrable (ou un ensemble de fonctions) **calculable** au sens où nous connaissons de manière exacte Pg. De ce fait, les fonctions calculables vont dépendre de la loi P choisie dans notre contexte. Par la suite, nous injectons cette information auxiliaire forte afin d'obtenir la mesure empirique informée. La proposition 3.1.4 nous indique qu'il y a diminution stricte de la variance du processus limite dès que $cov_P(g, f) \neq 0$. Ainsi une information auxiliaire pertinente consiste à trouver une fonction g (ou un ensemble de fonctions) corrélée à f sous P. Cela consiste donc à s'intéresser aux fonctions g calculables qui *maximisent* la quantité $| cov_P(g, f) |$. D'un point de vue méthodologique, nous pouvons rechercher les fonctions calculables pour P puis tester si l'information auxiliaire est pertinente. La section 6.2 propose des tests statistiques permettant de vérifier si l'information auxiliaire est pertinente avec ou sans seuil de tolérance.

Supposons dorénavant que nous disposons d'un vecteur de fonctions g pertinent. En utilisant cette information auxiliaire, nous obtenons la mesure empirique informée $\mathbb{P}_n^I$ donnée par la définition 2.3.1. Par le théorème 3.1.1,

$$\mathbb{P}_n^I f \xrightarrow[n\to+\infty]{p.s.} Pf.$$

De plus, par le théorème 3.1.2, nous avons

$$\sqrt{n}\left(\mathbb{P}_n^I f - Pf\right) \Rightarrow \mathscr{N}\left(0, \sigma_f\right)$$

où σ_f est donnée par

$$\sigma_f = Var_P f - cov_P(g,f)^T \Sigma^{-1} cov_P(g,f).$$

Remarquons que par la proposition 3.1.4, nous avons que $\sigma_f < Var_P f$. Ainsi nous obtenons une réduction de la variance en injectant une information auxiliaire et donc **une amélioration significative** de la méthode de Monte-Carlo standard. De plus notons que toute méthode utilisant la méthode de Monte-Carlo standard est améliorée du fait que l'estimation est plus précise. On peut citer par exemple l'échantillonnage préférentiel.

Exemple 7.1.1. Nous souhaitons estimer la valeur de l'intégrale suivante

$$\int_1^3 \frac{\sin t}{t} dt$$

Posons $w(t) = 2\frac{\sin t}{t}$ avec $t \in \mathbb{R}^*$. De plus, nous remarquons que

$$\int_1^3 \frac{\sin t}{t} dt = \mathbb{E}(w(X))$$

où X suit une loi uniforme sur l'intervalle $[1,3]$. Nous injectons l'information auxiliaire forte donnée par $\mathbb{E}(X^2) = \frac{13}{3}$. En utilisant une procédure de test, nous remarquons que l'information auxiliaire est pertinente. Par la suite nous générons $n = 500$ données et nous calculons $\mathbb{P}_k w$ et $\mathbb{P}_k^I w$ pour tout $k \in [\![2,500]\!]$. Dans la Figure 7.1, nous observons que $\mathbb{P}_k^I w$ est beaucoup plus proche de $\int_1^3 \frac{\sin t}{t} dt$ que ne l'est $\mathbb{P}_k w$.

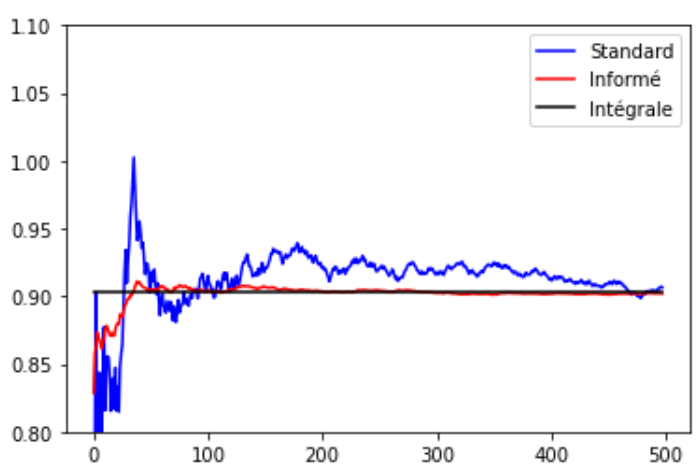

FIGURE 7.1 – Pour tout $k \in [\![2,500]\!]$, nous comparons $\mathbb{P}_k^I f$ (en rouge) et $\mathbb{P}_k f$ (en bleu).

Remarque 7.1.1. *Nous remarquons que l'utilisation de la mesure empirique informée dans ce contexte est similaire à la méthode de la variable de contrôle permettant de réduire la variance de la méthode de Monte-Carlo. On pourra consulter à ce sujet le livre [29].*

7.1.2 Utilisation de variables auxiliaires

En statistique, il n'est pas rare d'avoir à notre disposition des variables auxiliaires. Dans cette sous-section, on souhaite utiliser des variables auxiliaires afin d'améliorer l'estimation d'un paramètre. Nous montrons que l'injection de variables auxiliaires est un cas particulier d'information auxiliaire forte.

Tout d'abord posons le cadre mathématique. On observe un échantillon $X_1, \cdots, X_n$ de taille $n \in \mathbb{N}^*$ *i.i.d.* de loi P_1 à valeurs dans $\mathscr{X}$. On dispose aussi de variables auxiliaires $A_1, \cdots, A_n$ *i.i.d.* de loi P_2 à valeurs dans $\mathscr{A} \subset \mathbb{R}^k$ pour un certain $k \in \mathbb{N}^*$. De plus, on suppose que les variables auxiliaires sont indépendantes aux données. On note pour tout $i \in [\![1, n]\!]$, $Z_i = (X_i, A_i)$ et $\mathbb{P}_n$ la mesure empirique associée à l'échantillon $Z_1, \cdots, Z_n$. Observons que l'échantillon $Z_1, \cdots, Z_n$ est i.i.d et on note Q la loi du couple. On suppose de plus qu'on dispose d'une information auxiliaire de la forme $g(Q)$. Cela est assez courant en théorie des sondages de disposer d'une variable auxiliaire sur l'ensemble de la population (recensement etc). Dans ce cas, le statisticien dispose d'une information du type $P_2 g$ et il pose pour tout $(x, a) \in \mathscr{X} \times \mathscr{A}$, $\tilde{g}(x, a) = g(a)$, on a $P_2 g = Q\tilde{g}$.

La mesure empirique informée est définie pour toute fonction $f : \mathscr{X} \times \mathscr{A} \to \mathbb{R}$ de carré intégrable par

$$\mathbb{P}_n^I(f) = \mathbb{P}_n f + \Lambda_n(f) \cdot (g[\mathbb{P}_n] - g[Q])$$

où

$$\Lambda_n(f) = -\Sigma_{1,n}(f)\Sigma_{2,n}^{-1} = -cov_n(\varphi_g(\mathbb{P}_n), f)^T Var_n(\varphi_g[\mathbb{P}_n])^{-1}.$$

Remarquons qu'on pourrait s'intéresser seulement aux fonctions f définies sur $\mathscr{X}$. Dans ce cas, il faudrait seulement poser $\tilde{f}(x, a) = f(x)$ pour tout $(x, a) \in \mathscr{X} \times \mathscr{A}$. Pour une classe de fonctions $\mathscr{F}$ de Donsker, on a par le théorème 3.2.3

$$\alpha_n^I \Rightarrow G_I = G(\cdot) - cov_Q(\varphi_g(Q), \cdot)^T Var_Q(\varphi_g(Q))^{-1} G(\varphi_g(Q)) \text{ dans } l^\infty(\mathscr{F})$$

où G est Q-pont brownien. Généralement, g dépend seulement des variables auxiliaires et donc de P_2 et f dépend seulement de X. Cela est le cas dans le cadre d'un sondage. Ainsi dans ce cadre la variance du processus limite est donnée pour toute fonction $f : \mathscr{X} \times \mathscr{A} \to \mathbb{R}$ de carré intégrable par

$$\begin{aligned} Var\ G_I(f) &= Var_{P_1}(f) - cov_Q(\varphi_g(P_2), f)^T Var_{P_2}(\varphi_g(P_2))^{-1} cov_Q(\varphi_g(P_2), f) \\ &= Var_{P_1}(f) - cov(\varphi_g(P_2)(A), f(X))^T Var_{P_2}(\varphi_g(P_2))^{-1} cov(\varphi_g(P_2)(A), f(X)). \end{aligned}$$

On observe alors que plus les variables auxiliaires sont corrélées à X, plus la variance diminue.

Exemple 7.1.2. Illustrons cela par un exemple sur des données simulées. Soit $\sigma \in \mathbb{R}_+$. On génère 500 réalisations de (Z, Y) avec $Y = \sigma Z + \varepsilon$ où Z et ε sont deux variables aléatoires indépendantes de loi $P_1 = \mathcal{N}(0,1)$. On peut remarquer que Y et Z sont deux variables corrélées et Y suit la loi $P_2 = \mathcal{N}\left(0, 1+\sigma^2\right)$. Dans cet exemple, Y joue le rôle de variable auxiliaire pour laquelle on connaît son moment d'ordre 2 à savoir $1+\sigma^2$. De plus, on rappelle que l'observateur de l'expérience statistique observe seulement les données $(Z_1, Y_1), \cdots, (Z_{500}, Y_{500})$ et connaît le moment d'ordre 2 de Y. On souhaite estimer le moment d'ordre 2 de Z égal à 1 en utilisant l'information auxiliaire à notre disposition. On pose pour tout $(y,z) \in \mathbb{R}^2$, $f(z,y) = z^2$ et $g(z,y) = y^2$. La variance du processus limite est alors

$$\begin{aligned} Var\ G_I(f) &= Var Z^2 - \frac{\left(cov\left(Z^2, Y^2\right)\right)^2}{Var\left(Y^2\right)} \\ &= 2 - \frac{4\sigma^4}{2\sigma^4 + 4\sigma^2 + 2} \\ &= \frac{4\sigma^2 + 2}{\sigma^4 + 2\sigma^2 + 1}. \end{aligned}$$

Ainsi le paramètre σ joue un rôle important dans l'estimation de $\mathbb{E}\left(Z^2\right)$ car plus σ est grand, plus la variance diminue. Observons aussi que puisque $cov\left(Z^2, Y^2\right) = 2\sigma^2$ alors la variable auxiliaire Y est d'autant plus corrélée à Z que σ est grand. Les graphiques suivants (Figure 7.2) illustrent que $\mathbb{P}_n^I f$ est plus proche de Pf que ne l'est $\mathbb{P}_n f$ lorsque σ devient de plus en plus grand. De plus, on remarque que lorsque $\sigma \geq 5$ l'estimation $\mathbb{P}_n^I f$ est extrêmement proche de $P_1 f = 1$.

7.1.3 Données manquantes massives et information auxiliaire

Lors d'un recueil de données pour une expérience statistique, il est courant d'avoir des données manquantes. Dans cette sous-section, on suppose qu'on a affaire à des données manquantes massives. De plus on s'intéresse à un certain type de données manquantes à savoir des réalisations d'un couple de variables aléatoires (X, Y) à valeurs dans $\mathcal{X} \times \mathcal{Y}$ pour lesquelles parfois on n'observe pas la variable Y. Formellement, on observe $n \in \mathbb{N}^*$ réalisations du couple (X, Y) notées $(X_1, Y_1), \cdots, (X_n, Y_n)$ et $m \in \mathbb{N}^*$ réalisations de la variable X notées $X_{n+1}, \cdots, X_{n+m}$. Ainsi les variables $Y_{n+1}, \cdots, Y_{n+m}$ ne sont pas observées. Plusieurs méthodes existent pour traiter les données manquantes. Certaines méthodes suppriment des données, d'autres utilisent un modèle de régression ou des méthodes de classification pour attribuer une valeur à $Y_{n+1}, \cdots, Y_{n+m}$. On pourra consulter à ce sujet [15] [17].

Nous proposons d'utiliser les variables $X_{n+1}, \cdots, X_{n+m}$ comme une information auxiliaire à notre disposition. En effet, ces dernières nous donnent une information auxiliaire sur la loi de X. Puisqu'on se place dans un cadre de données manquantes massives, on supposera que $m \geq n$. Notons $Q = \mathbb{P}^{(X,Y)}$ et $P = \mathbb{P}^X$. Soit $\mathcal{F}$ une classe de fonctions Q-Donsker. Puisque les données auxiliaires $X_{n+1}, \cdots, X_{n+m}$ ne sont pas des couples, on ne peut pas ajouter ces données comme ce qui a été fait dans la sous-section 4.1.2. L'idée est alors d'injecter une estimation de Pg pour une

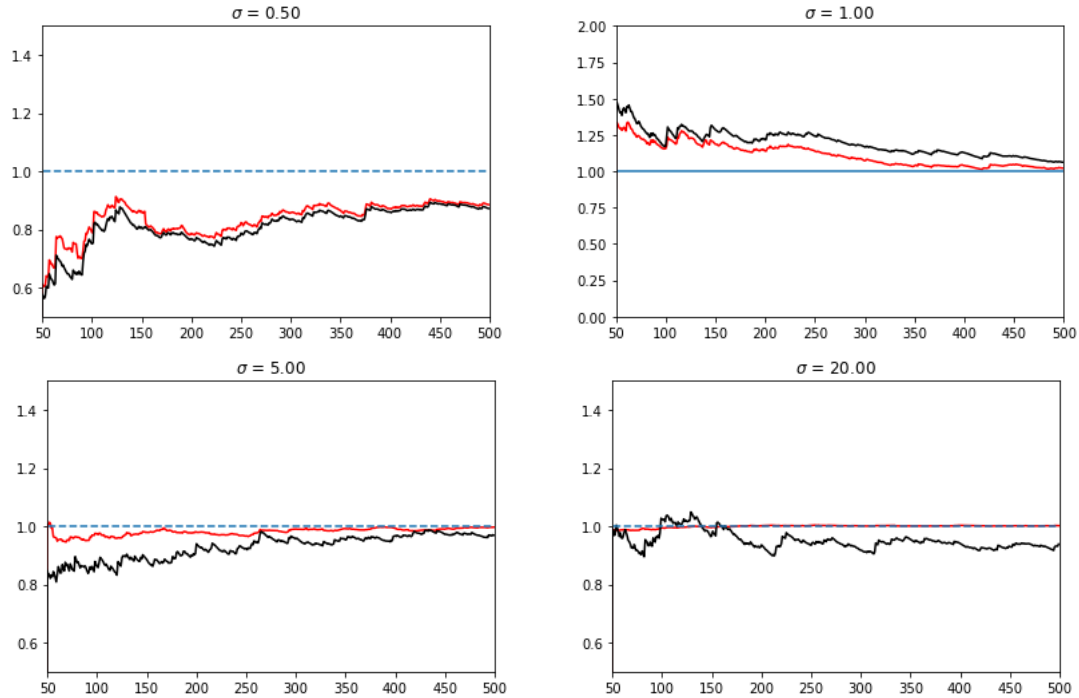

FIGURE 7.2 – Pour $n \in [\![50, 500]\!]$, on compare $\mathbb{P}_n^I f$ (en rouge) et $\mathbb{P}_n f$ (en noir) pour différentes valeurs de σ. On observe que l'estimation $\mathbb{P}_n^I$ est de plus en plus précise au fur à mesure que la quantité $cov(Z^2, Y^2) = 2\sigma^2$ est grande.

fonction auxiliaire choisie g en utilisant les données auxiliaires $X_{n+1}, \cdots, X_{n+m}$. Posons

$$\mathbb{Q}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)},$$
$$\mathbb{P}_m = \frac{1}{m} \sum_{i=n+1}^{n+m} \delta_{X_i}.$$

Remarquons que $Q(g \circ \pi) = Pg$ où π est la projection canonique par rapport à la première variable. Notons $\widetilde{g} = g \circ \pi$ définie sur $\mathscr{X} \times \mathscr{Y}$. Puisque $Q\widetilde{g} = Pg$, l'injection de l'estimation $\mathbb{P}_m g$ est un cas particulier de la sous-section 4.1.1. Ainsi la mesure empirique informée est donnée pour tout $f \in \mathscr{F}$ par

$$\mathbb{Q}_n^I f = \mathbb{Q}_n f - \frac{1}{1 + \frac{n}{m}} cov_{\mathbb{Q}_n}(f, \widetilde{g}) \left(Var_{\mathbb{P}_m} g\right)^{-1} \left(\mathbb{Q}_n \widetilde{g} - \mathbb{P}_m g\right).$$

On note $c = \lim_{n \to +\infty} \frac{n}{m}$ et $\alpha_n^I = \sqrt{n}\left(\mathbb{Q}_n^I - Q\right)$ le processus empirique associé à la mesure empirique informée. Par le théorème 4.1.6, on a

$$\alpha_n^I \Rightarrow G_I \text{ dans } l^\infty(\mathscr{F})$$

où pour tout $f \in \mathscr{F}$, $G_I(f) = G_1(f) - \frac{1}{1+c} cov_Q(f, \widetilde{g}) \left(Var_P g\right)^{-1} \left(G_1(\widetilde{g}) - \sqrt{c}\, G_2(g)\right)$ avec G_1 et G_2 respectivement un Q-pont bronwien et un P-pont brownien indépendants. Par le lemme 4.1.8, la variance est

$$Var\ G_I(f) = Var_Q f - \frac{1}{1+c} cov_Q(f, \widetilde{g}) \left(Var_P g\right)^{-1} cov_Q(\widetilde{g}, f) \tag{7.1}$$

avec $f \in \mathscr{F}$.

Exemple 7.1.3. Illustrons cela par une simulation sur des données simulées. Soit $\sigma \in \mathbb{R}_+^{\setminus\{1\}}$. Supposons que nous observons $n = 500$ réalisations du couple (X, Y) avec $Y = \sigma X + \varepsilon$ et X, ε sont deux variables indépendantes de loi $\mathscr{N}(0,1)$. Précisons que l'observateur ne connaît pas la relation qui lie Y à X. De plus, nous avons aussi $m = 99500$ données partielles pour lesquelles nous observons seulement des réalisations de X. Nous souhaitons estimer $\mathbb{E}(X-Y)^2$ et nous posons $f(x,y) = (x-y)^2$ avec $x, y \in \mathbb{R}$. Notons que $\mathbb{E} f(X,Y) = (\sigma - 1)^2 + 1$. Nous choisissons la fonction $\widetilde{g}(x,y) := g(x) = x^2$ avec $x, y \in \mathbb{R}$ car

$$cov_P(\widetilde{g}, f) = 2(\sigma - 1)^2 > 0.$$

Ainsi nous injectons l'estimation $\mathbb{P}_m g$ où

$$\mathbb{P}_m = \frac{1}{m} \sum_{i=n+1}^{n+m} \delta_{X_i}.$$

Dans la figure 7.3, on prend $\sigma = 5$ et nous comparons $\mathbb{Q}_k^I f$ en rouge et $\mathbb{Q}_k f$ en bleu pour tout $k \in [\![2, 500]\!]$. Nous remarquons que l'estimation de $\mathbb{E} f(X,Y)$ est beaucoup plus précise en utilisant les données partielles que si l'on se restreignait aux données complètes. En effet, dès que $n \geq 50$ la courbe en rouge est quasiment *collée* à la droite noire $y = 17$ (sur $[50, +\infty[$). Ainsi l'utilisation des données partielles vues comme une information auxiliaire est bénéfique pour l'estimation d'une quantité d'intérêt.

On souhaite à présent trouver $g \in L^2(P)$ à valeurs réelles qui minimise la variance donnée par l'égalité (7.1). La proposition suivante nous donne la fonction auxiliaire optimale.

Proposition 7.1.2. *Soit $f \in \mathscr{F}$. On pose pour tout $x \in \mathscr{X}$,*

$$g_f(x) := g(x) := \mathbb{E}\left(f(X,Y) | X = x\right).$$

Alors on a

$$\sup_{\varphi \in L^2(P)} \frac{\left(cov_Q\left(\varphi \circ \pi, f\right)\right)^2}{Var_P \varphi} = Var_P g_f$$

et le sup est atteint en g. De plus $Var_P g_f \leq Var_Q f$.

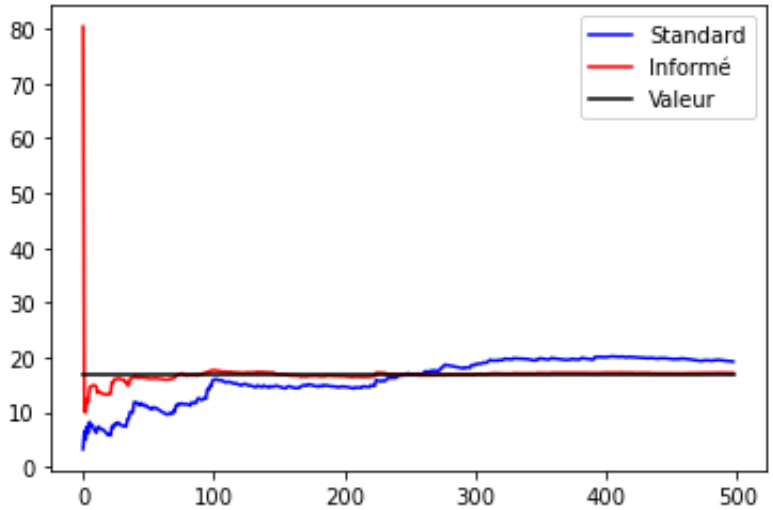

FIGURE 7.3 – Pour $\sigma = 5$ et pour tout $k \in [\![2,500]\!]$, nous comparons $\mathbb{Q}_k^I f$ (en rouge) et $\mathbb{Q}_k f$ (en bleu) .

Démonstration.
Tout d'abord observons que

$$cov_Q(g \circ \pi, f) = \mathbb{E}\big(g(X)f(X,Y)\big) - \mathbb{E}\big(g(X)\big)\mathbb{E}\big(f(X,Y)\big).$$

Puisque

$$\begin{aligned} \mathbb{E}(g(X)) &= \mathbb{E}\big(f(X,Y)\big), \\ \mathbb{E}\big(g(X)f(X,Y)\big) &= \mathbb{E}\big(g(X)^2\big) \end{aligned}$$

on en déduit que $cov_Q(g \circ \pi, f) = Var_P g$. *Ainsi*

$$\frac{\big(cov_Q\big(g \circ \pi, f\big)\big)^2}{Var_P g} = Var_P g.$$

Fixons $\varphi \in L^2(P)$. *Alors*

$$\begin{aligned} cov_Q\big(\varphi \circ \pi, f\big) &= \mathbb{E}\big(\varphi(X)f(X,Y)\big) - \mathbb{E}\big(\varphi(X)\big)\mathbb{E}\big(f(X,Y)\big) \\ &= \mathbb{E}\big(\varphi(X)g(X)\big) - \mathbb{E}\big(\varphi(X)\big)\mathbb{E}\big(g(X)\big) \\ &= cov_P\big(\varphi, g\big) \end{aligned}$$

car $\mathbb{E}\big(f(X,Y)\big) = \mathbb{E}\big(g(X)\big)$ *et* $\mathbb{E}\big(\varphi(X)f(X,Y)\big) = \mathbb{E}\big(\varphi(X)g(X)\big)$ *par définition de l'espérance conditionnelle. Ainsi par l'inégalité de Cauchy-Schwarz*

$$cov_P\big(\varphi, g\big)^2 \leq Var_P\varphi Var_P g,$$

on en déduit que

$$\frac{\left(cov_Q\left(\varphi\circ\pi, f\right)\right)^2}{Var_P\varphi} \leq Var_P g.$$

Enfin montrons que $Var_P g \leq Var_Q f$. *Puisque* $\mathbb{E}(g(X)) = \mathbb{E}(f(X,Y))$ *et*

$$\mathbb{E}\left(\left(f(X,Y)-g(X)\right)\left(g(X)-\mathbb{E}(g(X))\right)\right) = 0,$$

on a

$$Var_Q f = \mathbb{E}\left(\left(f(X,Y)-g(X)\right)^2\right) + Var_P g \geq Var_P g.$$

□

Cependant cette fonction g_f n'est pas accessible du fait que l'espérance conditionnelle ne peut pas être déterminée de manière exacte. Ainsi on devra avoir recours à une estimation de g_f qu'on notera g_n avec $n \in \mathbb{N}^*$. L'information auxiliaire apportée par g_n pour tout $n \in \mathbb{N}^*$ sera notée I_n. On pose pour tout $(x,y) \in \mathscr{X} \times \mathscr{Y}$, $\widetilde{g}_n(x,y) = g_n(x)$ et $\widetilde{g}_f(x,y) = g_f(x)$.

Proposition 7.1.3. *Soit* $f \in \mathscr{F}$. *Supposons que*

1. $\|\widetilde{g}_n - \widetilde{g}_f\|_{L^2(\mathbb{Q}_n)} \xrightarrow[n\to+\infty]{p.s.} 0$,
2. $\sqrt{n}\left(\mathbb{Q}_n(\widetilde{g}_n - \widetilde{g}_f) - \mathbb{P}_m(g_n - g_f)\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0$.

Alors

$$\alpha_n^{I_n}(f) = \sqrt{n}\left(\mathbb{Q}_n^{I_n} f - Qf\right) \Rightarrow G_I(f).$$

Démonstration.
Posons $g := g_f$ *et* $\widetilde{g} := \widetilde{g}_f$. *La mesure empirique informée est*

$$\mathbb{Q}_n^{I_n} f = \mathbb{Q}_n f - \frac{1}{1+\frac{n}{m}} cov_{\mathbb{Q}_n}(f,\widetilde{g}_n)\left(Var_{\mathbb{Q}_n} g_n\right)^{-1}\left(\mathbb{Q}_n\widetilde{g}_n - \mathbb{P}_m g_n\right).$$

Tout d'abord, observons que

$$\begin{aligned} |\, cov_{\mathbb{Q}_n}(f,\widetilde{g}_n) - cov_Q(f,\widetilde{g})\,| &\leq |\, cov_{\mathbb{Q}_n}(f,\widetilde{g}_n) - cov_{\mathbb{Q}_n}(f,\widetilde{g})\,| + |\, cov_{\mathbb{Q}_n}(f,\widetilde{g}) - cov_Q(f,\widetilde{g})\,| \\ &\leq \|f\|_{L^2(\mathbb{Q}_n)}\|\widetilde{g}_n - \widetilde{g}\|_{L^2(\mathbb{Q}_n)} + |\, cov_{\mathbb{Q}_n}(f,\widetilde{g}) - cov_Q(f,\widetilde{g})\,|\,. \end{aligned}$$

En utilisant le lemme 3.2.1 et puisque $\|\widetilde{g}_n - g\|_{L^2(\mathbb{Q}_n)} \xrightarrow[n\to+\infty]{p.s.} 0$, *on en déduit que*

$$cov_{\mathbb{Q}_n}(f,\widetilde{g}_n) \xrightarrow[n\to+\infty]{p.s.} cov_Q(f,\widetilde{g}).$$

On traite de manière similaire la variance

$$\begin{aligned}|\, Var_{\mathbb{Q}_n}\widetilde{g}_n - Var_Q\widetilde{g}\,| &\leq |\, Var_{\mathbb{Q}_n}\widetilde{g}_n - Var_{\mathbb{Q}_n}\widetilde{g}\,| + |\, Var_{\mathbb{Q}_n}\widetilde{g} - Var_Q\widetilde{g}\,| \\ &\leq |\, \mathbb{Q}_n(\widetilde{g}_n^2 - \widetilde{g}^2)\,| + |\, \left(\mathbb{Q}_n\widetilde{g}_n\right)^2 - \left(\mathbb{Q}_n\widetilde{g}\right)^2\,| + |\, Var_{\mathbb{Q}_n}\widetilde{g} - Var_Q\widetilde{g}\,|\end{aligned}$$

Par le lemme 3.2.1, on en déduit que $|\, Var_{\mathbb{Q}_n}\widetilde{g} - Var_Q\widetilde{g}\,| \xrightarrow[n\to+\infty]{p.s.} 0$. *Concernant les deux premiers termes, remarquons que*

$$\begin{aligned}|\, \mathbb{Q}_n(\widetilde{g}_n^2 - \widetilde{g}^2)\,| + |\, \left(\mathbb{Q}_n\widetilde{g}_n\right)^2 - \left(\mathbb{Q}_n\widetilde{g}\right)^2\,| &\leq \|\widetilde{g}_n - \widetilde{g}\|_{L^2(\mathbb{Q}_n)}\|\widetilde{g}_n + \widetilde{g}\|_{L^2(\mathbb{Q}_n)} + |\, \mathbb{Q}_n(\widetilde{g}_n - \widetilde{g})\,|\,|\, \mathbb{Q}_n(\widetilde{g}_n + \widetilde{g})\,| \\ &\leq 2\|\widetilde{g}_n - \widetilde{g}\|_{L^2(\mathbb{Q}_n)}\|\widetilde{g}_n + \widetilde{g}\|_{L^2(\mathbb{Q}_n)}\end{aligned}$$

Ainsi puisque $\|\widetilde{g}_n - g\|_{L^2(\mathbb{Q}_n)} \xrightarrow[n\to+\infty]{p.s.} 0$, *on en déduit que* $Var_{\mathbb{Q}_n}\widetilde{g}_n \xrightarrow[n\to+\infty]{p.s.} Var_Q\widetilde{g}$. *Enfin observons que*

$$\begin{aligned}\sqrt{n}\left(\mathbb{Q}_n\widetilde{g}_n - \mathbb{P}_m g_n\right) &= \sqrt{n}\left(\mathbb{Q}_n\widetilde{g}_n - Q\widetilde{g}\right) - \sqrt{n}\left(\mathbb{P}_n g_n - Pg\right) \\ &= \sqrt{n}\left(\mathbb{Q}_n\widetilde{g} - Q\widetilde{g}\right) + \sqrt{n}\mathbb{Q}_n(\widetilde{g}_n - \widetilde{g}) - \sqrt{n}\left(\mathbb{P}_m g - Pg\right) - \sqrt{n}\mathbb{P}_m(g_n - g) \\ &= \sqrt{n}\left(\mathbb{Q}_n\widetilde{g} - \mathbb{P}_m g\right) + \sqrt{n}\left(\mathbb{Q}_n(\widetilde{g}_n - \widetilde{g}) - \mathbb{P}_m(g_n - g)\right)\end{aligned}$$

car $Q\widetilde{g} = Pg$. *Puisque* $\sqrt{n}\left(\mathbb{Q}_n(\widetilde{g}_n - \widetilde{g}_f) - \mathbb{P}_m(g_n - g_f)\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0$, *on a*

$$\sqrt{n}\left(\mathbb{Q}_n\widetilde{g}_n - \mathbb{P}_m g_n\right) = \sqrt{n}\left(\mathbb{Q}_n\widetilde{g} - \mathbb{P}_m g\right) + o_p(1).$$

Ainsi par ce qui précède, on conclut en remarquant que

$$\alpha_n^{I_n}(f) = \alpha_n^I(f) + o_p(1).$$

□

7.2 Applications en statistique paramétrique et non paramétrique

7.2.1 Estimateur du maximum de vraisemblance informé et matrice d'information de Fisher informée

Soit $\mathscr{P} = \{P_\theta,\, \theta \in \Theta \subset \mathbb{R}^k\}$ avec $k \in \mathbb{N}^*$. On observe un échantillon $X_1, \cdots, X_n$ *i.i.d.* à valeurs dans $\mathscr{X}$ de taille $n \in \mathbb{N}^*$ et de loi P_{θ_0} avec $\theta_0 \in \Theta$. Puisque

$$\theta_0 = arg\min_{\theta\in\Theta} KL(P_{\theta_0}, P_\theta) = arg\max_{\theta\in\Theta} E_{\theta_0}\log p_\theta\ ,$$

on définit le **maximum de vraisemblance** (EMV) de la manière suivante

$$\hat{\theta}_n = arg\max_{\theta\in\Theta}\ \mathbb{P}_n\log p_\theta.$$

Ainsi on peut définir le maximum de vraisemblance informé de manière analogue.

Definition 7.2.1. Le **maximum de vraisemblance informé** $\hat{\theta}_n^I$ est défini par

$$\hat{\theta}_n^I := arg \max_{\theta \in \Theta} \mathbb{P}_n^I \log p_\theta.$$

Plaçons nous dans le cadre d'un modèle régulier au sens de la définition donnée dans le livre de *Benoît Cadre* [13] :

Definition 7.2.2. Un modèle statitique paramétrique $\mathscr{P}$ avec pour espace de paramètre $\Theta \subset \mathbb{R}^k$ pour $k \geq 0$ est **régulier** si pour chaque $\theta \in \Theta$:

1. Son information de Fisher I_θ en θ existe et est inversible,
2. $\mathbb{E}_\theta \nabla \log p_\theta = 0$ et $I_\theta = -\mathbb{E}_\theta \nabla^2 \log p_\theta$.

Le théorème suivant porte sur la normalité asymptotique de l'estimateur du maximum de vraisemblance informé.

Théorème 7.2.1. *Supposons que le modèle $\mathscr{P}$ est régulier et que pour chaque $\theta \in \Theta$, il existe un voisinage $\mathscr{V} \subset \Theta$ de θ pour lequel $\sup_{\alpha \in \mathscr{V}} \|\nabla^2 \log p_\theta\|^2 \in L^1(P_\theta)$. Si l'EMV informé $\hat{\theta}_n^I$ est consistant alors*

$$\sqrt{n}(\hat{\theta}_n^I - \theta) \Rightarrow \mathscr{N}\left(0, \tilde{I}_\theta^{-1}\right) \quad sous\ P_\theta$$

où

$$\tilde{I}_\theta^{-1} = Var_P\left(I_\theta^{-1} G_I(\nabla \log p_\theta)\right) = I_\theta^{-1} Var_P\left(G_I(\nabla \log p_\theta)\right) I_\theta^{-1}.$$

De plus $\tilde{I}_\theta^{-1} \leq I_\theta^{-1}$ pour tout $\theta \in \Theta$.

Démonstration.
La preuve de ce théorème s'inspire de la démonstration faite par Benoît Cadre et Celine Vial sur la normalité asymptotique du maximum de vraisemblance [13].

Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires i.i.d. de loi P_θ. On pose pour tout $\alpha \in \Theta$

$$U_n(\alpha) = \mathbb{P}_n^I \log p_\alpha(X_i)).$$

Par régularité de $\theta \mapsto \log(p_\theta)$, U_n est de classe C^2. Ainsi $\nabla U_n(\hat{\theta}_n^I) = 0$ car $\hat{\theta}_n^I$ maximise U_n. Un developpement de Taylor avec reste intégral nous donne

$$0 = U_n(\theta) + \left(\int_0^1 \nabla^2 U_n(\theta + t(\hat{\theta}_n^I - \theta))\, dt\right)(\hat{\theta}_n^I - \theta).$$

En posant $\overline{U}_n(\theta) = \int_0^1 \nabla^2 U_n(\theta + t(\hat{\theta}_n^I - \theta))\, dt$, *on a*

$$\overline{U}_n \sqrt{n}(\hat{\theta}_n^I - \theta) = -\sqrt{n} U_n(\theta).$$

Puisque $Var_\theta(\nabla \log p_\theta) = I_\theta$ *et* $\mathbb{E}(\nabla \log p_\theta) = 0$ *par régularité du modèle, on a par un résultat de type Donsker pour la mesure empirique informée (par exemple les théorèmes 3.1.2 et 3.2.3),*

$$\sqrt{n} U_n(\theta) = \sqrt{n} \mathbb{P}_n^I \nabla \log p_\theta \Rightarrow G_I(\nabla \log p_\theta)$$

sous P_θ *où* G_I *est le* P*-pont brownien informé.*

Montrons que

$$\overline{U}_n \xrightarrow{\mathbb{P}} -I_\theta.$$

sous P_θ. *On pose pour tout* $(x, r) \in \mathscr{X} \times \mathbb{R}_+$

$$\sigma(x, r) = \sup_{\|\alpha - \theta\| \leq r} \|\nabla^2 \log p_\alpha(x) - \nabla^2 \log p_\theta(x)\|.$$

Puisque $\theta \mapsto \log(p_\theta)(x)$ *est de classe* C^2 *pour tout* $x \in \mathscr{X}$ *et* $\sigma(\cdot, r) \in L^1(P_\theta)$ *pour* r *suffisamment petit, on a par théorème de convergence dominée*

$$\forall \epsilon > 0,\ \exists r > 0,\ \mathbb{E}(\sigma(X, r)) < \frac{\epsilon}{2}$$

car $\sigma(x, r) \xrightarrow{r \to 0} 0$. *Soit* $\epsilon > 0$ *fixé. Remarquons que*

$$\overline{U}_n(\theta) = \mathbb{P}_n^I \left(\int_0^1 \nabla^2 \log p_{\theta + t(\hat{\theta}_n^I - \theta)}(\cdot)\, dt \right).$$

Notons $|\cdot|$ *une norme quelconque. Ainsi*

$$\begin{aligned}
&\mathbb{P}(|\overline{U}_n(\theta) + I_\theta| \geq \epsilon) \\
&\leq \mathbb{P}\left(\left| \mathbb{P}_n^I \left(\int_0^1 \left(\nabla^2 \log p_{\theta + t(\hat{\theta}_n^I - \theta)}(X_i) - \nabla^2 \log p_\theta(X_i) \right) dt \right) \right| \geq \frac{\epsilon}{2} \right) \\
&+ \mathbb{P}\left(\left| \mathbb{P}_n^I \nabla^2 \log p_\theta(X_i) + I_\theta \right| \geq \frac{\epsilon}{2} \right) \\
&\leq \mathbb{P}\left(\left| \mathbb{P}_n^I \left(\int_0^1 \left(\nabla^2 \log p_{\theta + t(\hat{\theta}_n^I - \theta)}(X_i) - \nabla^2 \log p_\theta(X_i) \right) dt \right) \right| \geq \frac{\epsilon}{2}\ , \|\hat{\theta}_n^I - \theta\| < r \right) \\
&+ \mathbb{P}(\|\hat{\theta}_n^I - \theta\| \geq r) + \mathbb{P}\left(\left| \mathbb{P}_n^I \nabla^2 \log p_\theta(X_i) + I_\theta \right| \geq \frac{\epsilon}{2} \right) \\
&\leq \mathbb{P}\left(\mathbb{P}_n^I \sigma(X_i, r) \geq \frac{\epsilon}{2} \right) + \mathbb{P}(\|\hat{\theta}_n^I - \theta\| \geq r) + \mathbb{P}\left(\left| \mathbb{P}_n^I \nabla^2 \log p_\theta(X_i) + I_\theta \right| \geq \frac{\epsilon}{2} \right).
\end{aligned}$$

Par un résultat du type Glivenko-Cantelli (théorème 3.1.1, 3.2.2 et 4.1.5) pour la mesure empirique informée, on a

$$\mathbb{P}_n^I \sigma(X_i, r) \xrightarrow{p.s.} \mathbb{E}(\sigma(X, r)) < \frac{\epsilon}{2}.$$

Donc le premier terme tend vers 0. *Le second terme tend vers* 0 *par consistance de* $\hat{\theta}_n^I$. *Pour le troisième terme, on utilise à nouveau le résultat du type Glivenko-Cantelli (théorème 3.1.1, 3.2.2 et 4.1.5) pour la mesureempirique informée*

$$\mathbb{P}_n^I \nabla^2 \log p_\theta(X_i) \xrightarrow{p.s.} -I_\theta.$$

Ainsi on en déduit que

$$\overline{U}_n \xrightarrow{\mathbb{P}} -I_\theta.$$

Remarquons que si on suppose qu'on a la consistance forte alors la convergence a lieu p.s.. On conclut par le lemme de Slutsky

$$\sqrt{n}(\hat{\theta}_n^I - \theta) \Rightarrow I_\theta^{-1} \mathcal{N}\left(0, Var_P\left(G_I(\nabla \log p_\theta)\right)\right) = \mathcal{N}\left(0, \tilde{I}_\theta^{-1}\right)$$

sous P_θ *et en remarquant que* $\tilde{I}_\theta^{-1} = I_\theta^{-1} Var_P\left(G_I(\nabla \log p_\theta)\right) I_\theta^{-1}$. *Enfin concernant la dernière assertion, il suffit de remarquer que pour deux matrices symétriques* $A, B \in \mathcal{M}_p(\mathbb{R})$ *avec* $p \in \mathbb{N}^*$ *telles que* $A \leq B$, *on a que pour tout* $C \in \mathcal{M}_p(\mathbb{R})$

$$C^T AC \leq C^T BC.$$

En effet soit $x \in \mathbb{R}^p$

$$x^T C^T ACx = (Cx)^T A(Cx) \leq (Cx)^T B(Cx) = x^T C^T BCx.$$

□

Remarque 7.2.2. *Nous pouvons aussi énoncer un théorème de consistance pour l'estimateur* $\hat{\theta}_n^I$ *en adaptant la démonstration du théorème* 4.4.1 *du livre de Benoît Cadre et Celine Vial [13]. Plus précisément en reprenant les notations de la démonstration dans* [13], *il suffit de remplacer* U_n *par*

$$U_n^I(\theta) = -\mathbb{P}_n^I \log p_\theta \; \theta \in \Theta$$

et les résultats du type Glivenko-Cantelli (théorème 3.1.1, 3.2.2 et 4.1.5) pour la mesure empirique informée permettent de conserver la même démonstration.

On appelle $\tilde{I}_\theta^{-1}$ la **matrice d'information de Fisher informée inverse**. Enfin énonçons un résultat montrant que le modèle paramétrique muni d'une information auxiliaire est plus *informé* que le modèle paramétrique.

Théorème 7.2.3. *Dans le cas où* $\tilde{I}_{\theta_0}^{-1}$ *est inversible, la matrice d'information de Fisher informée vérifie l'inégalité suivante*

$$I_\theta \leq \tilde{I}_{\theta_0}.$$

Démonstration (du théorème).
Afin de montrer le résultat voulu, démontrons le lemme suivant :

Lemme 7.2.4. *Soient $A, B \in \mathcal{M}_p(\mathbb{R})$ deux matrices réelles symétriques définies positives. Si $A \leq B$ alors $B^{-1} \leq A^{-1}$.*

Démonstration (du lemme).
Démontrons ce lemme. Tout d'abord rappelons que pour $C \in \mathcal{M}_p(\mathbb{R})$, on a

$$i)\ \ A \leq B \Rightarrow C^T AC \leq C^T BC.$$

De plus remarquons aussi que

$$ii)\ \ Id \leq B \Rightarrow B^{-1} \leq Id.$$

En effet, puisque B est symétrique, on peut la diagonaliser dans une base orthonormée de vecteurs propres. Ainsi il existe une matrice P orthogonale et une matrice D diagonale telles que

$$B = PDP^T.$$

Puisque $D > 0$ car B est symétrique définie positive, on peut définir $B^r := PD^rP^T$ pour $r \in \mathbb{R}$. Ainsi

$$B^{-1} = B^{-\frac{1}{2}} IdB^{-\frac{1}{2}} \leq B^{-\frac{1}{2}} BB^{-\frac{1}{2}} = Id.$$

Il nous suffit d'appliquer i) et ii) pour démontrer le lemme

$$\begin{aligned} A \leq B &\Rightarrow A^{-\frac{1}{2}} AA^{-\frac{1}{2}} \leq A^{-\frac{1}{2}} BA^{-\frac{1}{2}} \\ &\Rightarrow Id \leq A^{-\frac{1}{2}} BA^{-\frac{1}{2}} \\ &\Rightarrow A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}} \leq Id. \end{aligned}$$

D'où

$$B^{-1} = A^{-\frac{1}{2}} A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}} A^{-\frac{1}{2}} \leq A^{-\frac{1}{2}} IdA^{-\frac{1}{2}} = A^{-1}.$$

□

Revenons à la démonstration du théorème. Puisque I_{θ_0} et $\tilde{I}_{\theta_0}$ sont des matrices symétriques définies positives (car inversibles) et $\tilde{I}_{\theta_0}^{-1} \leq I_{\theta_0}^{-1}$, on en déduit par le lemme précédent 7.2.4 que

$$I_\theta \leq \tilde{I}_{\theta_0}.$$

□

Donnons quelques exemples :

Exemple 7.2.1. 1. On observe $X_1, \cdots, X_n$ *i.i.d.* de loi $\mathrm{Ber}(\theta)$. On suppose qu'on dispose de n données auxiliaires $Y_1, \cdots, Y_n$ issues d'expériences similaires indépendantes (*information auxiliaire faible*). Ainsi le P-pont brownien informé est $G_I = \frac{1}{\sqrt{2}} G$ avec G un P-pont brownien et

$$\hat{\theta}_n^I = \frac{1}{2n}(\sum_{i=1}^n X_i + \sum_{i=1}^n Y_i),$$
$$\tilde{I}_\theta = \frac{2}{\theta(1-\theta)} > I_\theta = \frac{1}{\theta(1-\theta)}.$$

On remarque que l'information apportée par une observation dans le cadre du modèle paramétrique muni d'une information auxiliaire est deux fois plus importante que celle dans le cadre du modèle paramétrique sans information auxiliaire.

2. De manière plus générale, on observe l'échantillon $X_1, \cdots, X_n$ *i.i.d.* de loi P_{θ_0} (modèle statistique paramétrique général) et on dispose de $r \times n$ données auxiliaires $Y_1, \cdots, Y_{rn}$ (avec $r \in \mathbb{R}_+^*$) issues d'expériences similaires indépendantes (*information auxiliaire faible*). Alors le P-pont brownien informé est $G_I = \frac{1}{\sqrt{1+r}} G$ avec G un P-pont brownien et

$$\tilde{I}_\theta = (1+r) I_\theta > I_\theta.$$

7.2.2 Méthode des moments avec informations auxiliaires

Soient $X_1, \cdots, X_n$ un échantillon *i.i.d.* de taille $n \in \mathbb{N}^*$ de loi P_{θ_0} où $\theta_0 \in \Theta \subset \mathbb{R}^k$ avec $k \in \mathbb{N}^*$.

Definition 7.2.3. Soit $\theta_0 \in \Theta \subset \mathbb{R}^k$ et soit $f = (f_1, \cdots, f_m)$ un vecteur de fonctions intégrables avec $m \geq 1$. Soit $\gamma : \Theta \to \mathbb{R}^m$ définie par $\gamma(\theta) = P_\theta f$. On appelle estimateur des moments avec informations auxiliaires (MMI) une solution $\hat{\theta}_n^I$ du système d'équations suivant

$$\mathbb{P}_n^I f = \gamma(\theta_0).$$

Le résultat principal de cette section montre que la méthode des moments avec informations auxiliaires (MMI) est meilleure que celle sans information auxiliaire :

Théorème 7.2.5. *Supposons que γ est C^1 au voisinage de θ_0 et telle que $D\gamma(\theta_0)$ est inversible. On suppose de plus que $P_{\theta_0}\|f\|^2 < +\infty$. Alors l'estimateur MMI existe p.s. à partir d'un certain rang et vérifie*

$$\sqrt{n}(\hat{\theta}_n^I - \theta_0) \Rightarrow \mathcal{N}\left(0, D\gamma(\theta_0)^{-1} Var_{P_{\theta_0}} G_I(f) \, (D\gamma(\theta_0)^{-1})^T\right).$$

Démonstration.
Rappelons que $\gamma(\theta) = P_\theta f$. Par le résultat de Donsker pour la mesure empirique informé, on a

$$\sqrt{n}(\mathbb{P}_n^I f - \gamma(\theta_0)) \Rightarrow \mathcal{N}\left(0, Var_{P_{\theta_0}}\ G_I(f)\right).$$

Les hypothèses faites nous permettent d'appliquer le théorème d'inversion locale : Il existe un ouvert U de θ_0 et un ouvert V de $\gamma(\theta_0)$ tels que $\gamma : U \to V$ est un C^1 difféomorphisme. Puisque p.s. à partir d'un certain rang $\mathbb{P}_n^I f \in V$ (par un résultat de type Glivenko-Cantelli pour la mesure empirique informé), l'estimateur MMI $\hat{\theta}_n^I = \gamma^{-1}(\mathbb{P}_n^I f)$ existe p.s. à partir d'un certain rang. Ainsi il suffit d'appliquer la méthode Delta à γ^{-1} pour obtenir le résultat souhaité. □

Exemple 7.2.2. On note I une information auxiliaire portant sur P_{θ_0}.

1. Supposons que $P_{\theta_0} = \mathcal{N}\left(\mu_0, \sigma_0^2\right)$. Alors l'estimateur des moments avec informations auxiliaires est

$$\hat{\theta}_n^I = \left(\mathbb{P}_n^I x,\ Var_{\mathbb{P}_n^I} x\right).$$

2. Supposons que $P_{\theta_0} = \Gamma(\theta_1, \theta_2)$. Dans ce cas la densité est

$$f_{\theta_1,\theta_2}(x) = \frac{x^{\theta_1 - 1} e^{-\frac{x}{\theta_2}}}{\theta_2^{\theta_1}\Gamma(\theta_1)} 1_{\mathbb{R}_+}(x), \quad x \in \mathbb{R}.$$

Alors l'estimateur des moments avec informations auxiliaires est

$$\hat{\theta}_n^I = \left(\frac{\left(\mathbb{P}_n^I x\right)^2}{\mathbb{P}_n^I x^2 - \left(\mathbb{P}_n^I x\right)^2}, \frac{\mathbb{P}_n^I x^2 - \left(\mathbb{P}_n^I x\right)^2}{\mathbb{P}_n^I x}\right).$$

7.2.3 Estimateur à noyau informé

Soit $X_1, \cdots, X_n$ un échantillon *i.i.d.* de taille $n \in \mathbb{N}^*$ et de loi P sur $\mathbb{R}^d$ possédant une densité f_0 par rapport à la mesure de Lebesgue λ. On suppose qu'on dispose d'une information auxiliaire I portant sur P. On note $\mathbb{P}_n^I = \sum_{i=1}^n q_i \delta_{X_i}$ la mesure empirique informée.

Definition 7.2.4. Soient $K : \mathbb{R}^d \to \mathbb{R}$ une fonction paire telle que $\int_{\mathbb{R}^d} k(y)dy1$ et h_n une fenêtre pouvant dépendre de n. **La mesure empirique informée lisse** est définie par

$$\mathbb{P}_n^I * K_{h_n}(dx) = \frac{1}{h_n^d} \sum_{i=1}^n q_i K\left(\frac{x - X_i}{h_n}\right) \lambda(dx).$$

L'estimateur à noyau informé est alors défini par

$$\forall x \in \mathbb{R}^d,\ \hat{f}_n^I(x) = \frac{1}{h_n^d} \sum_{i=1}^n q_i K\left(\frac{x - X_i}{h_n}\right).$$

Remarque 7.2.6.

- *Pour toute fonction intégrable f, on note*

$$\mathbb{P}_n^I * K_{h_n}(f) = \int_{\mathbb{R}^d} f(x)\ d\mathbb{P}_n^I * K_{h_n}(x).$$

- *L'estimateur à noyau non informé est noté $\hat{f}_n$ et est défini en prenant les poids $q_i = \frac{1}{n}$ pour tout $i \in [\![1, n]\!]$.*

Afin de quantifier l'apport de l'information auxiliaire, on s'intéresse à la convergence du processus

$$\sqrt{n}\left(\mathbb{P}_n^I * K_{h_n} - P\right)$$

indexé par une classe de fonctions $\mathscr{F}$. Dans cette sous-section, nous souhaitons montrer que la mesure empirique lisse fait aussi mieux que la mesure empirique informée $\mathbb{P}_n^I$ au sens où le processus $\sqrt{n}\left(\mathbb{P}_n^I * K_{h_n} - P\right) \Rightarrow G_I$ dans $l^\infty(\mathscr{F})$.

Théorème 7.2.7. *Soit $\mathscr{F}$ une classe P-Donsker. On suppose que*

$$\sup_{f \in \mathscr{F}} \sup_{x \in \mathbb{R}^d} \left| \int_{\mathbb{R}^d} (f(x + h_n u) - f(x)) K(u)\ du \right| = o\left(\frac{1}{\sqrt{n}}\right).$$

Alors

$$\sqrt{n}\left(\mathbb{P}_n^I * K_{h_n} - P\right) \Rightarrow G_I \ \ dans\ \ l^\infty(\mathscr{F}).$$

Démonstration.
Soit $f \in \mathscr{F}$. Calculons

$$\begin{aligned}
\mathbb{P}_n^I * K_{h_n}(f) &= \int_{\mathbb{R}^d} f(x)\ d\mathbb{P}_n^I * K_{h_n}(x) \\
&= \frac{1}{h_n^d} \sum_{i=1}^n q_i \int_{\mathbb{R}^d} f(x) K\left(\frac{x - X_i}{h_n}\right)\ dx.
\end{aligned}$$

En faisant le changement de variable $u = \frac{x-X_i}{h_n}$, *on a*

$$\frac{1}{h_n^d}\sum_{i=1}^{n} q_i \int_{\mathbb{R}^d} f(x)K\left(\frac{x-X_i}{h_n}\right)\,dx = \sum_{i=1}^{n} q_i \int_{\mathbb{R}^d} f(X_i+h_n u)K(u)\,du.$$

Posons pour tout $x \in \mathbb{R}^d$,

$$\Gamma_{n,f}(x) = \int_{\mathbb{R}^d} f(x+h_n u)K(u)\,du.$$

Ainsi

$$\sqrt{n}\left(\mathbb{P}_n^I * K_{h_n}(f) - Pf\right) = \sqrt{n}\left(\mathbb{P}_n^I(\Gamma_{n,f}-f)\right) + \sqrt{n}(\mathbb{P}_n^I f - Pf).$$

Puisque

$$\begin{aligned}\left|\sqrt{n}\left(\mathbb{P}_n^I(\Gamma_{n,f}-f)\right)\right| &\le \sqrt{n}\sup_{x\in\mathbb{R}^d}\left|\int_{\mathbb{R}^d}(f(x+h_n u)-f(x))K(u)\,du\right| \\ &\le \sqrt{n}\sup_{f\in\mathscr{F}}\sup_{x\in\mathbb{R}^d}\left|\int_{\mathbb{R}^d}(f(x+h_n u)-f(x))K(u)\,du\right|\end{aligned}$$

on obtient par hypothèse que

$$\sqrt{n}\sup_{f\in\mathscr{F}}\left|\left(\mathbb{P}_n^I(\Gamma_{n,f}-f)\right)\right| \le \sqrt{n}\sup_{f\in\mathscr{F}}\sup_{x\in\mathbb{R}^d}\left|\int_{\mathbb{R}^d}(f(x+h_n u)-f(x))K(u)\,du\right| = o(1).$$

Par conséquent, puisque $\mathscr{F}$ *est une classe* P*-Donsker et par le résultat de type Donsker pour la mesure empirique informée (théorème 3.1.2, 3.2.3 et 4.1.6) , on a*

$$\alpha_n^I = \sqrt{n}(\mathbb{P}_n^I - P) \Rightarrow G_I \ \ dans\ \ l^\infty(\mathscr{F}).$$

On en déduit que

$$\sqrt{n}\left(\mathbb{P}_n^I * K_{h_n} - P\right) \Rightarrow G_I \ \ dans\ \ l^\infty(\mathscr{F}).$$

□

Exemple 7.2.3. Supposons que $\sqrt{n}h_n \to 0$. Alors la classe $\mathscr{F}$ des fonctions lipschitziennes dont la constante de lipschitz est bornée par un certain réel $M > 0$ vérifie les hypothèses du théorème 7.2.7.

Dans [20], les auteurs proposent d'autres conditions à vérifier afin d'assurer la convergence en loi pour la mesure empirique lisse. Dans le théorème suivant, on supposera que l'information auxiliaire I est forte et est donnée par une fonctionnelle $g(P)$.

Théorème 7.2.8. *Soit $\mathscr{F}$ une classe de fonctions P-Donsker et invariante par translation. Posons la mesure suivante $\mu_n(dy) = \frac{1}{h_n} K\left(\frac{y}{h_n}\right)$. On suppose que $\int_{\mathbb{R}^d} K(y)\, dy = 1$ et que pour tout n, $\mathscr{F} \subset L^1(|\mu_n|)$ et $\int_{R^d} \|f(\cdot - y)\|_{L^2(P)} d|\mu_n|(y) < +\infty$. On suppose de plus que les conditions suivantes sont satisfaites*

$$\sqrt{n} \sup_{f \in \mathscr{F}} \left| \mathbb{E} \int_{\mathbb{R}^d} (f(X+y) - f(X)) d\mu_n(y) \right| \xrightarrow[n \to +\infty]{} 0\,, \tag{7.2}$$

$$\sup_{f \in \mathscr{F}} \mathbb{E} \left(\int_{\mathbb{R}^d} (f(X+y) - f(X)) d\mu_n(y) \right)^2 \xrightarrow[n \to +\infty]{} 0. \tag{7.3}$$

Alors

$$\sqrt{n} \left(\mathbb{P}_n^I * \mu_n - P \right) \Rightarrow G_I \ \ dans \ \ l^\infty(\mathscr{F}).$$

Remarque 7.2.9. *Faisons quelques remarques utiles pour la démonstration de ce théorème*

$$\begin{aligned}
\mathbb{P}_n^I * \mu_n(f) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x+y) d\mathbb{P}_n^I(x) d\mu_n(y) \\
&= \sum_{i=1}^n q_i \int_{\mathbb{R}^d} f(X_i + y) \frac{1}{h_n} K\left(\frac{y}{h_n}\right) dy \\
&= \sum_{i=1}^n q_i \int_{\mathbb{R}^d} f(X_i + h_n y) K\left(y\right) d \\
&= \mathbb{P}_n^I \Gamma_{n,f}
\end{aligned}$$

en reprenant les notations de la démonstration du théorème 7.2.7. Posons pour tout borélien A de $\mathbb{R}^d$, $\bar{\mu}_n(A) := \mu_n(-A)$. Ainsi, on a pour une mesure de probabilité Q,

$$\begin{aligned}
(Q * \mu_n - Q) f &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (f(x + h_n y) - f(x)) K(y) dy dQ(x) \\
&= \int_{\mathbb{R}^d} (\bar{\mu}_n * f - f)(x) dQ(x) \\
&= Q(\bar{\mu}_n * f - f)
\end{aligned}$$

où

$$\bar{\mu}_n * f(x) = \int_{\mathbb{R}^d} f(x-y) d\bar{\mu}_n(y) \tag{7.4}$$

$$= \int_{\mathbb{R}^d} f(x-y) \frac{1}{h_n} K\left(\frac{-y}{h_n}\right) dy \tag{7.5}$$

$$= \int_{\mathbb{R}^d} f(x + h_n y) K(y) dy. \tag{7.6}$$

Démonstration.
On reprend la démonstration faite dans [20] et on la modifie légèrement. Puisque

$$\sqrt{n}\left(\mathbb{P}_n^I * \mu_n - P\right) = \sqrt{n}\left(\mathbb{P}_n^I * \mu_n - \mathbb{P}_n^I\right) + \sqrt{n}\left(\mathbb{P}_n^I - P\right),$$

il suffit d'étudier

$$\mathbb{P}_n^I * \mu_n - \mathbb{P}_n^I = (\mathbb{P}_n^I - P) * \mu_n - (\mathbb{P}_n^I - P) + P * \mu_n - P.$$

Observons que

$$(P * \mu_n - P)(f) = \mathbb{E}\int_{\mathbb{R}^d} (f(X+y) - f(X))d\mu_n(y).$$

Alors par (7.2)*, on a*

$$\|P * \mu_n - P\|_{\mathscr{F}} = o\left(\frac{1}{\sqrt{n}}\right).$$

Par la remarque précédente, on obtient

$$\left((\mathbb{P}_n^I - P) * \mu_n - (\mathbb{P}_n^I - P)\right) f = (\mathbb{P}_n^I - P)(\bar{\mu}_n * f - f).$$

Par le lemme 2 *de [20],* $\mathscr{G} = \bigcup_n \{\bar{\mu}_n * f - f,\ f \in \mathscr{F}\}$ *est une classe* P*-Donsker. Par* (7.3)*,* $\sup_{f\in\mathscr{F}} P(\bar{\mu}_n * f - f)^2 \to 0$*. On en déduit comme dans [20],*

$$\sup_{f\in\mathscr{F}} |(\mathbb{P}_n - P)(\bar{\mu}_n * f - f)| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

De plus, remarquons que

$$(\mathbb{P}_n^I - P)(\bar{\mu}_n * f - f) = (\mathbb{P}_n - P)(\bar{\mu}_n * f - f) - cov_n(\bar{\mu}_n * f - f, \varphi_g(\mathbb{P}_n))(Var_n\varphi_g(\mathbb{P}_n))^{-1} Z_n$$

où $Z_n = g(\mathbb{P}_n) - g(P)$*. En utilisant le fait que* $\sqrt{n}Z_n$ *converge en loi et* $(Var_n\varphi_g(\mathbb{P}_n))^{-1} \xrightarrow[n\to+\infty]{p.s.} (Var_P\varphi_g(P))^{-1}$ *, on remarque qu'il suffit de montrer que*

$$\sup_{f\in\mathscr{F}} \|cov_n(\bar{\mu}_n * f - f, \varphi_g(\mathbb{P}_n))\| \xrightarrow[n\to+\infty]{p.s.} 0.$$

Observons que

$$\begin{aligned}\|cov_n(\bar{\mu}_n * f - f, \varphi_g(\mathbb{P}_n))\| &\le \sqrt{Var_n(\bar{\mu}_n * f - f)} \max_{1\le i\le m} \sqrt{Var_n\varphi_{g_i}(\mathbb{P}_n)} \\ &\le \|\bar{\mu}_n * f - f\|_{L^2(\mathbb{P}_n)} \max_{1\le i\le m} \sqrt{Var_n\varphi_{g_i}(\mathbb{P}_n)}\end{aligned}$$

et que

$$\begin{aligned}\mathbb{P}_n(\bar{\mu}_n * f - f)^2 &= (\mathbb{P}_n - P)(\bar{\mu}_n * f - f)^2 + P(\bar{\mu}_n * f - f)^2 \\ &\le \|\mathbb{P}_n - P\|_{\mathscr{G}^2} + \sup_{f\in\mathscr{F}} P(\bar{\mu}_n * f - f)^2.\end{aligned}$$

Pour le second terme, par (7.3), $\sup_{f\in\mathscr{F}} P(\bar{\mu}_n * f - f)^2 \to 0$. *Pour le premier terme, puisque* $\mathscr{G}$ *est* P*-Donsker, elle est donc* P*-Glivenko-Cantelli. Puisque* $\mathscr{G}$ *admet une fonction enveloppe de carré intégrable et par stabilité du caractère Glivenko-Cantelli par une application continue,* $\mathscr{G}^2$ *est* P*-Glivenko-Cantelli. Ainsi*

$$\|\mathbb{P}_n - P\|_{\mathscr{G}^2} \xrightarrow[n\to+\infty]{p.s.} 0.$$

D'où $\sup_{f\in\mathscr{F}} \mathbb{P}_n(\bar{\mu}_n * f - f)^2 \xrightarrow[n\to+\infty]{p.s.} 0$ *et*

$$\sup_{f\in\mathscr{F}} \|cov_n(\bar{\mu}_n * f - f, \varphi_g(\mathbb{P}_n))\| \xrightarrow[n\to+\infty]{p.s.} 0.$$

Enfin par ce qui précède, on en déduit que

$$\sup_{f\in\mathscr{F}} |(\mathbb{P}_n^I - P)(\bar{\mu}_n * f - f)| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

En conclusion, on obtient

$$\sqrt{n}\left(\mathbb{P}_n^I * \mu_n - P\right) \Rightarrow G_I \ \ dans \ \ l^\infty(\mathscr{F}).$$

□

Exemple 7.2.4. Nous souhaitons comparer $f_0, \hat{f}_n$ et $\hat{f}_n^I$ à l'aide d'une simulation. Pour cela, nous générons $p = 20$ échantillons *i.i.d.* de taille $n = 100$ de loi $P = \mathscr{N}(5,16)$. Nous prenons comme fenêtre $h_n = (1,06)\sigma n^{-\frac{1}{5}}$ avec $\sigma = 4$. Nous supposons qu'on dispose d'une information auxiliaire I donnée par le moment d'ordre 1 et 2 de P. Pour chaque échantillon, nous traçons les courbes représentatives de $\hat{f}_n$ et $\hat{f}_n^I$ (Figure 7.4 se trouvant à la page suivante). Nous observons qu'il est difficile de juger de la pertinence d'injecter une information auxiliaire (par cette méthode) dans l'estimation point par point de la densité f_0. Cependant l'estimation d'espérance de loi donnée par la densité f_0 est améliorée.

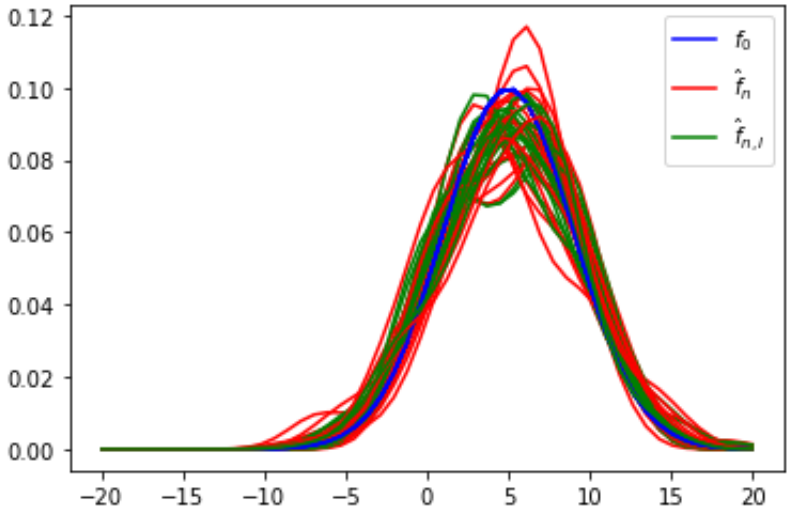

FIGURE 7.4 – Comparaison des estimateurs à noyau.

7.2.4 Quantiles empiriques informés

Dans cette sous-section, nous souhaitons estimer de manière plus précise le quantile d'une loi réelle en exploitant une information auxiliaire à notre disposition. Pour cela, nous allons définir le quantile empirique informé en utilisant la mesure empirique informée $\mathbb{P}_n^I$. Des travaux ont été menés autour des quantiles empiriques informés par *Zhang* dans [42] en utilisant la mesure empirique informée obtenue par vraisemblance empirique $\mathbb{P}_n^{(1)}$ définie dans (2.10). Nous montrons que le processus limite obtenu dans [42] est le même pour la mesure $\mathbb{P}_n^I$.

Soient $(X_n)_{n\in\mathbb{N}^*}$ une suite *i.i.d.* de loi P sur $\mathbb{R}$ et I une information auxiliaire. Notons F la fonction de répartition, $\mathbb{F}_n$ la fonction de répartition empirique et $\mathbb{F}_n^I$ la fonction de répartition empirique informée définie pour tout $t \in \mathbb{R}$ par

$$\mathbb{F}_{n,I}(t) = \mathbb{P}_n^I 1_{\cdot\leq t} = \sum_{i=1}^{n} q_i 1_{X_i\leq t}.$$

Exemple 7.2.5. Comparons la fonction de répartition empirique informée avec la fonction de répartition empirique non informée. Pour cela, on génère $n = 100$ variables aléatoires *i.i.d.* de loi $P = \mathcal{N}(0,1)$. On suppose qu'on dispose d'une information auxiliaire I donnée par $\big(\mathbb{E}(X), \mathbb{E}(X^2)\big)$. La Figure 7.5 est composée de deux graphiques. Dans le graphique de gauche, nous réalisons une seule fois l'expérience tandis que dans le graphique de droite nous réalisons 20 fois l'expérience et nous traçons sur le même graphique $\mathbb{F}_n^I$ et $\mathbb{F}_n$. Dans les deux graphiques, nous observons que la fonction de répartition informée $\mathbb{F}_{n,I}$ est plus proche de F que l'est $\mathbb{F}_n$. Dans le graphique de droite, nous remarquons que le faisceau vert est beaucoup plus resserré autour de la courbe bleue que ne l'est le faisceau rouge. Cela illustre les propriétés asymptotiques du processus empirique informé démontrées dans les chapitres précédents (Figure 7.5).

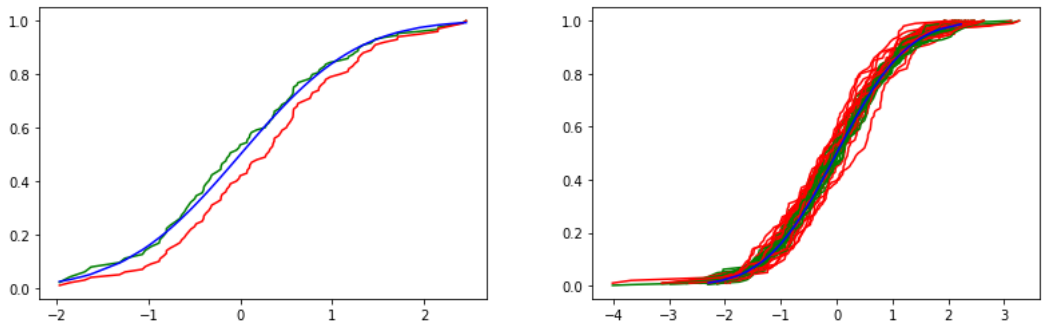

FIGURE 7.5 – Comparaison entre $\mathbb{F}_{n,I}$ (verte), $\mathbb{F}_n$ (rouge) et F (bleue). Le nombre de réalisations dans le graphe de droite est de 20 et de 1 pour celui de gauche.

Définissons la notion de quantile empirique informé.

Definition 7.2.5. Le **quantile empirique informé** d'ordre $\alpha \in]0,1[$ est défini de la manière suivante

$$q^I_{n,\alpha} = \mathbb{F}^{-1}_{n,I}(\alpha) = \inf\left\{t \in \mathbb{R},\ \mathbb{F}_{n,I}(t) \geq \alpha\right\}.$$

Remarque 7.2.10. *L'infimum est bien défini puisque pour tout* $t \in]-\infty, \min_{1\leq i\leq n} X_i[$, $F_{n,I}(t) = 0$ *et pour tout* $t \in [\max_{1\leq i\leq n} X_i, +\infty[$, $F_{n,I}(t) = 1$.

Afin d'obtenir des résultats asymptotiques à propos du quantile empirique informé, nous avons besoin du lemme technique suivant.

Lemme 7.2.11. *Soit* $\mathscr{S} = D(\mathbb{R}) \cap l^\infty(\mathbb{R})$ *l'ensemble des fonctions càdlàg bornées définies sur* $\mathbb{R}$ *muni de la norme uniforme. Soit* $\alpha \in]0,1[$ *et* $\phi_\alpha : D_\alpha \subset \mathscr{S} \to \mathbb{R}$ *définie par* $\phi_\alpha(F) = F^{-1}(\alpha)$ *pour tout* $F \in D_\alpha$ *où* D_α *contient l'ensemble des fonctions de répartition définies sur* $\mathbb{R}$ *et* $\mathbb{F}_{n,I}$ *pour tout* $n \in \mathbb{N}^*$. *Soit* F *une fonction de répartition.*

1. *On suppose que* F *est strictement croissante en* $F^{-1}(\alpha)$. *Alors* ϕ_α *est continue en* F.
2. *On suppose que* F *est différentiable en* $F^{-1}(\alpha)$ *et* $F'(F^{-1}(\alpha)) = f(F^{-1}(\alpha)) > 0$. *Alors* ϕ_α *est Hadamard-différentiable en* F *tangentiellement à*

$$D_0 = \{h \in \mathscr{S},\ h \text{ continue en } F^{-1}(\alpha)\}$$

de dérivée

$$\phi'_\alpha(h) = -\frac{h(\phi_\alpha(F))}{F'_{\phi(F)}} = -\frac{h(F^{-1}(\alpha))}{f(F^{-1}(\alpha))}.$$

Remarque 7.2.12. *Remarquons que nous pouvons prendre*

$$D_\alpha = \left\{ F \in \mathscr{S}, \lim_{t\to-\infty} F(t) = 0, \lim_{t\to+\infty} F(t) = 1 \right\}.$$

Démonstration.
La démonstration de la seconde assertion peut être trouvée dans [38] par exemple.

Démontrons la première assertion. Fixons $\varepsilon > 0$. Puisque F est strictement croissante en $F^{-1}(\alpha)$ alors

$$F(F^{-1}(\alpha) - \varepsilon) < \alpha < F(F^{-1}(\alpha) + \varepsilon).$$

Soit $(F_n)_{n\in\mathbb{N}^*} \subset D_\alpha$ telle que $\|F_n - F\|_\infty \to 0$. Alors il existe $N > 0$ telle que $\forall\, n \geq N$

$$F_n(F^{-1}(\alpha) - \varepsilon) < \alpha < F_n(F^{-1}(\alpha) + \varepsilon).$$

D'où par définition de la fonction quantile, cela implique que

$$F^{-1}(\alpha) - \varepsilon \leq F_n^{-1}(\alpha) \leq F^{-1}(\alpha) + \varepsilon.$$

Ainsi

$$|F_n^{-1}(\alpha) - F^{-1}(\alpha)| \leq \varepsilon.$$

On conclut que ϕ_α est continue en F. □

On en déduit le résultat suivant :

Théorème 7.2.13. *Soient F une fonction de répartition et q_α le quantile d'ordre $\alpha \in]0,1[$ de F. On note $q^I_{n,\alpha}$ le quantile empirique informée (Définition 7.2.5). Alors :*

1. *Si F est strictement croissante en $F^{-1}(\alpha)$ alors :*

$$q^I_{n,\alpha} \xrightarrow{p.s.} q_\alpha.$$

2. *Si F est différentiable en $F^{-1}(\alpha)$ et $F'(F^{-1}(\alpha)) = f(F^{-1}(\alpha)) > 0$ alors*

$$\sqrt{n}(q^I_{n,\alpha} - q_\alpha) \Rightarrow \mathcal{N}\left(0, \frac{Var_P G_I(F^{-1}(\alpha))}{\left(f(F^{-1}(\alpha)\right)^2}\right)$$

où pour tout $s \in \mathbb{R}$, $G_I(s) := G_I(1_{]-\infty,s]})$. Dans le cas où l'information auxiliaire est donnée par une fonctionnelle $g(P)$ (ou par des estimations issues de sources indépendantes), la variance s'exprime de la manière suivante

$$\sqrt{n}(q^I_{n,\alpha} - q_\alpha) \Rightarrow \mathcal{N}\left(0, \frac{\alpha(1-\alpha) - \tilde{I}}{\left(f(F^{-1}(\alpha)\right)^2}\right)$$

avec $\tilde{I} = cov_P(\varphi_g(P), 1_{]-\infty,F^{-1}(\alpha)]})^T \Sigma^{-1} cov_P(\varphi_g(P), 1_{]-\infty,F^{-1}(\alpha)]}) \geq 0$.

Démonstration.
Posons la classe $\mathscr{F} = \{f_s = 1_{]-\infty,s]},\ s \in \mathbb{R}\}$. Cette dernière est une classe P-Donsker. Alors

$$\|\mathbb{F}_{n,I} - F\|_\infty \xrightarrow[n\to+\infty]{p.s.} 0,$$
$$\sqrt{n}(\mathbb{F}_{n,I} - F) \Rightarrow G_I \text{ dans } l^\infty(\mathbb{R})$$

où $G_I(s) := G_I(f_s)$ dépend en s de seulement $G(s)$ et $cov_P(\varphi_g(P), 1_{]-\infty,s]})$. Observons que $G_I \in l^\infty(\mathbb{R})$. Par le point 1 du lemme 7.2.11, on a

$$\phi_\alpha(\mathbb{F}_{n,I}) \xrightarrow[n\to+\infty]{p.s.} \phi_\alpha(F).$$

D'où la première assertion. Concernant la seconde assertion, montrons que G_I est càdlàg sur $\mathbb{R}$ et continue en $F^{-1}(\alpha)$. Pour cela il suffit de montrer que les applications suivantes sont càdlàg et continues en $F^{-1}(\alpha)$

$$\psi_1 : s \mapsto G(s),$$
$$\psi_2 : s \mapsto cov_P(\varphi_g(P), 1_{]-\infty,s]}) = P\varphi_g(P)1_{]-\infty,s]} - P\varphi_g(P)F(s).$$

Concernant ψ_1, rappelons que G admet presque sûrement des trajectoires continues pour la semi-distance $\rho^2(f_1, f_2) = Var_P\left(f_1 - f_2\right)$ pour tout $f_1, f_2 \in L^2(P)$. Posons $\varphi : (\mathbb{R}, |\cdot|) \to \left(\mathscr{F}, \rho\right)$ définie pour tout $S \in \mathbb{R}$ par $\varphi(s) = f_s$ et observons que $\psi_1 = G \circ \varphi$. Il suffit de montrer que φ est càdlàg et continue en $F^{-1}(\alpha)$. En effet soient $s, s' \in \mathbb{R}$,

$$\rho(f_s, f_{s'}) = \sqrt{Var_P\left(f_s - f_{s'}\right)} \leq \|f_s - f_{s'}\|_{L^2(P)} = \sqrt{F(s) + F(s') - 2F(\min(s, s'))}$$

On peut donc en déduire que ψ_1 est càdlàg et continue en $F^{-1}(\alpha)$. Concernant ψ_2, il suffit de montrer que $s \mapsto P\varphi_g(P)1_{]-\infty,s]}$ est càdlàg et continue en $F^{-1}(\alpha)$. Soit $(s_n)_{n\in\mathbb{N}^*}$ une suite décroissante qui tend

vers s lorsque $n \to +\infty$. Alors puisque $\varphi_g(P) \in L^1(P)$, on peut en déduire par théorème de convergence dominée

$$P\varphi_g(P)1_{]-\infty,s_n]} \xrightarrow[n\to\infty]{} P\varphi_g(P)1_{]-\infty,s]}.$$

De même par théorème de convergence dominée, on peut montrer que

$$P\varphi_g(P)1_{]-\infty,s']} \xrightarrow[s'\to s^-]{} P\varphi_g(P)1_{]-\infty,s[}.$$

Enfin notons que puisque F est continue en $F^{-1}(\alpha)$, G_I est continue en $F^{-1}(\alpha)$. Il suffit maintenant d'appliquer la méthode Delta fonctionnelle et de conclure

$$\sqrt{n}(q^I_{n,\alpha} - q_\alpha) = \sqrt{n}(\phi_\alpha(\mathbb{F}_{n,I}) - \phi_\alpha(F_\alpha)) \Rightarrow \phi'_\alpha(G_I) = -\frac{G_I(F^{-1}(\alpha))}{f(F^{-1}(\alpha))}$$

en remarquant que $Var_P(1_{]-\infty,F^{-1}(\alpha)]}) = \alpha(1-\alpha)$ car $F(F^{-1}(\alpha)) = \alpha$. $\square$.

Exemple 7.2.6. Dans cet exemple, on suppose que l'information auxiliaire est donnée par le quantile d'ordre $\beta \in]0,1[$ de la loi P. On souhaite calculer explicitement $\widetilde{I}$ et la variance du processus limite données dans le théorème 7.2.13. Ainsi dans notre cas $g(x) = 1_{]-\infty,F^{-1}(\beta)]}(x)$ avec $x \in \mathbb{R}$. La quantité $\widetilde{I}$ est alors

$$\widetilde{I} = \frac{\left(cov_P(g, 1_{]-\infty,F^{-1}(\alpha)]})\right)^2}{Var_P g} = \frac{\left(\min(\alpha,\beta) - \alpha\beta\right)^2}{\beta(1-\beta)}.$$

Ainsi

$$\alpha(1-\alpha) - \widetilde{I} = \max\left((\alpha-\beta)\frac{1-\alpha}{1-\beta}, (\beta-\alpha)\frac{\alpha}{\beta}\right).$$

On en déduit que

$$\sqrt{n}(q^I_{n,\alpha} - q_\alpha) \Rightarrow \mathcal{N}\left(0, \frac{\max\left((\alpha-\beta)\frac{1-\alpha}{1-\beta}, (\beta-\alpha)\frac{\alpha}{\beta}\right)}{\left(f(F^{-1}(\alpha)\right)^2}\right).$$

Exemple 7.2.7. Illustrons cela par une simulation. On génère $n = 210$ variables aléatoires *i.i.d.* de loi $P = \mathcal{N}(0,1)$. On souhaite estimer la médiane de P qui est dans ce cas 0. On suppose que l'information auxiliaire disponible est donnée par le moment d'ordre 1 et le moment d'ordre 2. On calcule alors pour chaque $n \in [\![2, 210]\!]$, la médiane empirique informée et la médiane empirique classique. On observe qu'à partir d'un certain rang (ici à partir de $n = 70$), la médiane empirique informée converge plus rapidement vers 0 que la médiane empirique classique.

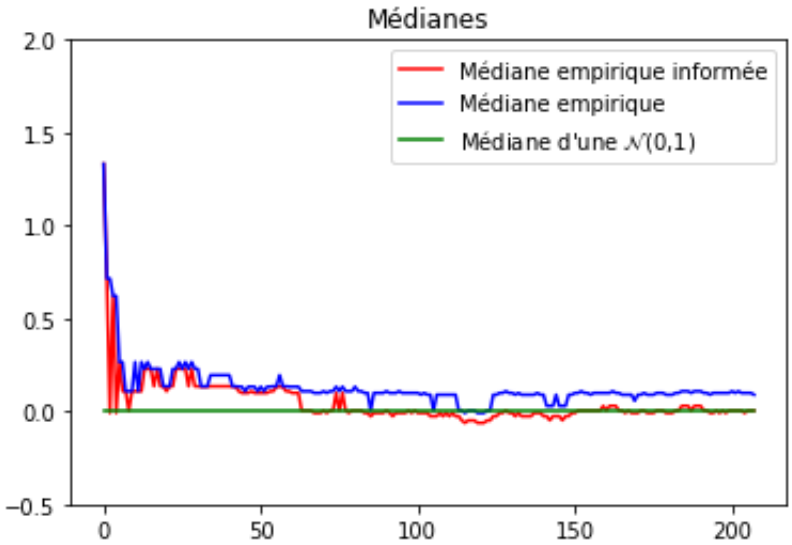

FIGURE 7.6 – Comparaison entre la médiane empirique informée et la médiane empirique classique.

7.3 Estimateur des moindres carrés avec informations auxiliaires et modèle linéaire informé

7.3.1 Estimateur des moindres carrés avec informations auxiliaires

Nous souhaitons injecter une information auxiliaire dans le cadre de la régression. L'information auxiliaire I portera sur la loi du couple (X,Y). Rappelons le cadre de la régression linéaire. On souhaite estimer

$$\theta_0 = arg \min_{\theta \in \Theta \subset \mathbb{R}^k} \mathbb{E}(Y - \langle X, \theta \rangle)^2$$

avec $k \in \mathbb{N}^*$. Notons $P = \mathbb{P}^{(X,Y)}$ la loi du couple et $h(\theta) = \mathbb{E}(Y - \langle X, \theta \rangle)^2$ pour tout $\theta \in \Theta$. Alors

$$\nabla h(\theta) = -2\mathbb{E}(Y - \langle X, \theta \rangle)X.$$

Ainsi

$$\begin{aligned} \nabla h(\theta) = 0 &\Leftrightarrow \mathbb{E}(XY) = \mathbb{E}X\langle X, \theta \rangle \\ &\Leftrightarrow \mathbb{E}(XY) = \mathbb{E}\left(XX^T\right)\theta. \end{aligned}$$

Sous réserve que $\mathbb{E}\left(XX^T\right)$ soit inversible, on a

$$\begin{aligned} \theta_0 &= \mathbb{E}\left(XX^T\right)^{-1}\mathbb{E}(XY) \\ &= (Pxx^T)^{-1}(Pxy). \end{aligned}$$

Puisque P est inconnu, l'estimateur des moindres carrés ordinaires (MCO) est défini de la manière suivante,

$$\hat{\theta}_n = arg\min_{\theta\in\Theta} \mathbb{P}_n(Y - \langle X,\theta\rangle)^2$$

où $\mathbb{P}_n$ est la mesure empirique associée à l'échantillon *i.i.d.* $(X_1, Y_1), \cdots, (X_n, Y_n)$ de taille $n \in \mathbb{N}^*$. De manière analogue, on peut en déduire la définition de l'estimateur informé des moindres carrés ordinaires.

Definition 7.3.1. Soit I une information auxiliaire. **L'estimateur informé des moindres carrés ordinaires** (IMCO) est défini par

$$\hat{\theta}_n^I = arg\min_{\theta\in\Theta} \mathbb{P}_n^I(Y - \langle X,\theta\rangle)^2 = arg\min_{\theta\in\Theta} \sum_{i=1}^n q_i(Y_i - \langle X_i,\theta\rangle)^2.$$

Sous réserve d'inversibilité,

$$\hat{\theta}_n^I = \left(\sum_{i=1}^n q_i X_i X_i^T\right)^{-1} \sum_{i=1}^n q_i Y_i X_i.$$

Remarque 7.3.1. *Posons $\psi(x,y) = xx^T$ pour tout $(x,y) \in \mathbb{R}^k \times \mathbb{R}$. Observons que*

$$\hat{\theta}_n^I = \left(\mathbb{P}_n^I \psi\right)^{-1} \mathbb{P}_n^I xy$$

où xy représente l'application $(x,y) \mapsto xy$. On peut écrire $\hat{\theta}_n^I$ sous forme matricielle en posant $Q = Diag(q_1, \cdots, q_n)$ la matrice des poids et $\mathbb{X}^T = (X_1, \cdots, X_n)$. On a alors

$$\hat{\theta}_n^I = \left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T QY.$$

Remarquons qu'en posant $Q = \frac{1}{n} Id$, on retrouve l'estimateur des moindres carrés ordinaires (MCO).

7.3.2 Propriétés asymptotiques de l'estimateur des moindres carrés informé

Afin de quantifier le gain d'information qu'on obtient en injectant de l'information auxiliaire, nous devons faire quelques hypothèses. Supposons que

$$\mathbb{E}(Y|X) = \langle X, \beta_0\rangle$$

pour un certain $\beta_0 \in \mathbb{R}^p$ avec $p \in \mathbb{N}^*$. Ainsi on peut écrire Y comme

$$Y = \langle X, \beta_0\rangle + \varepsilon$$

où $\varepsilon = Y - \mathbb{E}(Y|X)$ est le terme d'erreur. On remarque que l'hypothèse d'exogénéité est automatiquement satisfaite puisque

$$\mathbb{E}(\varepsilon|X) = \mathbb{E}(Y|X) - \mathbb{E}(Y|X) = 0$$

Ainsi $\mathbb{E}(\varepsilon X) = 0$. En fait par définition de l'espérance conditionnelle comme projection orthogonale dans L^2, on a $\mathbb{E}(\varepsilon\gamma(X)) = 0$ pour toute fonction mesurable γ. Sous reserve d'inversibilité $\mathbb{E}\left(XX^T\right)$, le modèle est identifiable puisque

$$\begin{aligned}\theta_0 &= \mathbb{E}\left(XX^T\right)^{-1}\mathbb{E}(XY)\\ &= \mathbb{E}\left(XX^T\right)^{-1}\mathbb{E}\left(XX^T\right)\beta_0 + \mathbb{E}\left(XX^T\right)^{-1}\mathbb{E}(\varepsilon X)\\ &= \beta_0.\end{aligned}$$

Soit $((X_n, Y_n))_{n\in\mathbb{N}^*}$ une suite *i.i.d.* de loi P. L'estimateur informé des moindres carrés ordinaires (IMCO) est

$$\hat{\theta}_n^I = \left(\sum_{i=1}^n q_i X_i X_i^T\right)^{-1} \sum_{i=1}^n q_i Y_i X_i.$$

Le résultat suivant porte sur la consistance forte de $\hat{\theta}_n^I$.

Théorème 7.3.2. *Supposons que* $\mathbb{E}(XX^T)$ *est inversible. Alors*

$$\hat{\theta}_n^I \xrightarrow[n\to+\infty]{p.s.} \beta_0.$$

Démonstration.
En effet

$$\begin{aligned}\hat{\theta}_n^I &= \left(\sum_{i=1}^n q_i X_i X_i^T\right)^{-1} \sum_{i=1}^n q_i Y_i X_i\\ &= \left(\sum_{i=1}^n q_i X_i X_i^T\right)^{-1} \left(\sum_{i=1}^n q_i X_i X_i^T\right)\beta_0 + \left(\sum_{i=1}^n q_i X_i X_i^T\right)^{-1} \sum_{i=1}^n q_i (Y_i - \langle X_i, \beta_0\rangle) X_i\\ &= \beta_0 + \left(\sum_{i=1}^n q_i X_i X_i^T\right)^{-1} \sum_{i=1}^n q_i (Y_i - \langle X_i, \beta_0\rangle) X_i.\end{aligned}$$

Par résultat de type Glivenko-Cantelli pour la mesure empirique informée (théorème 3.1.1, 3.2.2 et 4.1.5), le second terme tend p.s. vers 0 *car* $\mathbb{E}(\varepsilon X) = 0$. □

De plus, nous avons aussi un résulat de normalité asymptotique :

Théorème 7.3.3. *Supposons que* $\mathbb{E}(XX^T)$ *est inversible. Posons* $f_{\beta_0}(x,y) = (y - \langle x, \beta_0\rangle)x$ *et* $U = \mathbb{E}(XX^T)^{-1}$. *Alors*

$$\sqrt{n}(\hat{\theta}_n^I - \beta_0) \Rightarrow \mathcal{N}\left(0, UVar_P\, G_I\,(f_{\beta_0})U^T\right).$$

De plus, on a l'inégalité suivante

$$UVar_P\, G_I\,(f_{\beta_0})U^T \leq UVar_P\, G\,(f_{\beta_0})U^T.$$

Dans le cas où l'information auxiliaire est donnée par une fonctionnelle $g(P)$ *(ou par des estimations issues de sources indépendantes), la variance s'exprime de la manière suivante,*

$$\begin{aligned} Var_P\, G_I\,(f_{\beta_0}) &= Var_P\, f_{\beta_0} - cov_P(\varphi_g(P), f_{\beta_0})^T \Sigma^{-1} cov_P(\varphi_g(P), f_{\beta_0}) \\ &= Var_P\, f_{\beta_0} - \mathbb{E}(\varepsilon X \varphi_g(P)(X,Y)^T)\Sigma^{-1}\mathbb{E}(\varepsilon X \varphi_g(P)(X,Y)). \end{aligned}$$

Remarque 7.3.4. *Dans le cas où l'information auxiliaire est donnée par une fonctionnelle* $g(P)$ *(ou par des estimations issues de sources indépendantes), on peut remarquer que si l'information auxiliaire ne dépend que de* X *alors cela n'améliore pas l'estimation car*

$$\mathbb{E}(\varepsilon \varphi_g(P)(X)^T) = \mathbb{E}(\mathbb{E}(\varepsilon|X)\varphi_g(P)(X)^T) = 0.$$

Démonstration.
Remarquons que

$$\sqrt{n}(\hat{\theta}_n^I - \beta_0) = \left(\sum_{i=1}^n q_i X_i X_i^T\right)^{-1} \sqrt{n}\sum_{i=1}^n q_i(Y_i - \langle X_i, \beta_0\rangle)X_i.$$

Par un résultat du type Donsker pour la mesure empirique informée (théorème 3.1.2, 3.2.3 et 4.1.6) et en observant que $\mathbb{E}(\varepsilon X) = 0$ *où* $\varepsilon = Y - \langle X, \beta_0\rangle$, *on a*

$$\sqrt{n}\sum_{i=1}^n q_i(Y_i - \langle X_i, \beta_0\rangle)X_i = \sqrt{n}\mathbb{P}_n^I f_{\beta_0} \Rightarrow G_I(f_{\beta_0}).$$

Par résultat du type Glivenko-Cantelli pour la mesure empirique informée (théorème 3.1.1, 3.2.2 et 4.1.5) et par continuité de l'inverse, on a

$$\left(\sum_{i=1}^n q_i X_i X_i^T\right)^{-1} \xrightarrow[n\to+\infty]{p.s.} \left(\mathbb{E}(XX^T)\right)^{-1} = U.$$

Par le lemme de Slutsky, on obtient finalement que

$$\sqrt{n}(\hat{\theta}_n^I - \beta_0) \Rightarrow UG_I(f_{\beta_0}) = \mathcal{N}\left(0, UVar_P\, G_I\,(f_{\beta_0})U^T\right).$$

□

Exemple 7.3.1. On se place dans le cadre d'une régression linéaire simple. Un organisme d'étude dispose de $N = 530$ données portant sur des appartements. Plus précisément, ces données sont de la forme (x, y) où x représente la surface en m^2 et y le loyer de cet appartement. Cependant, nous n'avons accès qu'aux $n = 250$ premières données de cette base de données. De plus, nous disposons aussi d'une information auxiliaire I concernant le loyer moyen de l'ensemble des N appartements. À partir de ces données, on peut calculer l'estimateur non informé $\hat{\theta}_n$ et l'estimateur informé $\hat{\theta}_n^I$. Ainsi on peut tracer la droite de régression associée à chacun de ces deux estimateurs (Figure 7.7). On observe que la droite de régression informée a tendance à passer par les valeurs extrêmes.

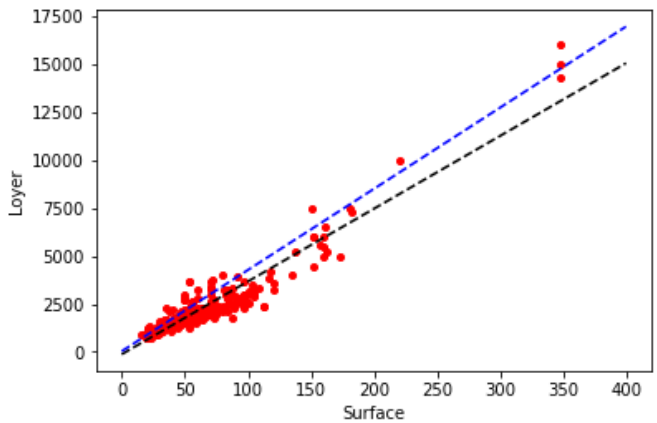

FIGURE 7.7 – La droite de régression pour l'estimateur informé (resp. non informé) est tracée en bleu (resp. en noir).

7.3.3 Interprétation géométrique et qualité d'ajustement dans le cadre d'un modèle linéaire avec informations auxiliaires

On souhaite interpréter géométriquement $\hat{Y} = \mathbb{X}\hat{\theta}_n^I$ comme le projeté orthogonale de Y sur $Vect(X^1, \cdots, X^p)$ pour un certain produit scalaire avec $p \in \mathbb{N}^*$. Notons $\mathbb{X} = \left(X^1, \cdots, X^p\right)$ avec $X^1 := \mathbf{1} := (1, \cdots, 1)^T$ et $(X^j)_{j \in [\![2,p]\!]}$ sont les variables du modèle. Pour cela, posons la forme bilinéaire symétrique suivante

$$\langle x, y \rangle_Q = x^T Q y, \ x, y \in \mathbb{R}^n. \tag{7.7}$$

Par la proposition 2.3.3, la matrice Q est définie positive presque sûrement à partir d'un certain rang. Ainsi cette forme bilinéaire est un produit scalaire presque sûrement à partir d'un certain rang.

Proposition 7.3.5. *Presque sûrement à partir d'un certain rang, le vecteur $\hat{Y} = \mathbb{X}\hat{\theta}_n^I \in \mathbb{R}^n$ est le projeté orthogonale de Y sur $Vect(X^1, \cdots, X^p)$ pour le produit scalaire $\langle \cdot, \cdot \rangle_Q$.*

Démonstration.
En posant $H = \mathbb{X}\left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q$, on a que $\hat{Y} = HY$. Montrons que H est la matrice de projection orthogonale sur $Vect(X^1, \cdots, X^p)$ pour le produit scalaire $\langle \cdot, \cdot \rangle_Q$. Tout d'abord, observons

$$\begin{aligned} H^2 &= \mathbb{X}\left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q \mathbb{X}\left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q \\ &= \mathbb{X}\left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q \\ &= H. \end{aligned}$$

Donc H est une matrice de projection. H est une projection orthogonale si pour tout $x \in \mathbb{R}^p$

$$\langle Hx, x - Hx \rangle_Q = 0 \Leftrightarrow x^T H^T Q H x = x^T Q H x.$$

Autrement dit montrons que $H^T Q H = QH$. En effet,

$$\begin{aligned} H^T Q H &= Q\mathbb{X}\left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q \mathbb{X}\left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q \\ &= Q\mathbb{X}\left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q \\ &= QH. \end{aligned}$$

Ainsi $\hat{Y}$ est le projeté orthogonal de Y sur $Vect(X^1, \cdots, X^p)$ pour le produit scalaire $\langle \cdot, \cdot \rangle_Q$. □

Dans le cadre d'un modèle linéaire informé, on peut de la même manière mesurer la qualité d'ajustement par le coefficient R^2,

$$R^2 = \frac{SSE}{SST}$$

où SSE est la variance expliquée, SST la variance totale et SSR la variance résiduelle **sous** $\mathbb{P}_n^I$. Plus précisément, notons $\overline{Y}_n^I = (\sum_{i=1}^n q_i Y_i)\mathbf{1}$ et

$$\begin{aligned} SST &= \sum_{i=1}^n q_i (Y_i - \overline{Y}_n^I)^2 = \| Y - \overline{Y}_n^I \|_Q^2, \\ SSE &= \sum_{i=1}^n q_i (\hat{Y}_i - \overline{Y}_n^I)^2 = \| \hat{Y} - \overline{Y}_n^I \|_Q^2, \\ SSR &= \sum_{i=1}^n q_i (Y_i - \hat{Y}_i)^2 = \| Y - \hat{Y} \|_Q^2. \end{aligned}$$

Puisque $Y - \overline{Y}_n^I \in Vect(X^1, ..., X^p)$ (car $X^1 = \mathbf{1}$) et

$$Y - \hat{Y} \perp_Q Vect(X^1, ..., X^p)$$

alors par le théorème de Pythagore, on a

$$\|Y - \overline{Y}_n^I\|_Q^2 = \|\hat{Y} - \overline{Y}_n^I\|_Q^2 + \|Y - \hat{Y}\|_Q^2.$$

Autrement dit

$$SST = SSE + SSR.$$

Ainsi ce coefficient R^2 est entre 0 et 1 et il est interprétable.

7.3.4 Variables instrumentales et informations auxiliaires

En économétrie et en statistique de manière générale, il arrive que les erreurs du modèle soient endogènes. Plus précisément X est **endogène** si $\mathbb{E}(\varepsilon|X) \neq 0$ ou plus faiblement $cov(\varepsilon, X) \neq 0$. Cela peut venir du fait qu'il y a des erreurs dans les mesures de X et Y ou suite à des équations simultanées par exemple. Ainsi le modèle n'est pas adapté aux données. De plus, l'estimateur n'est pas consistant car $cov(X, \varepsilon) \neq 0$. Afin de contourner ce problème, on introduit la notion d'**instrument** Z.

Definition 7.3.2. Un instrument Z est une variable aléatoire dans $\mathbb{R}^p$ telle que

$$\mathbb{E}(\varepsilon|Z) = 0 \text{ et } cov(X, Z) \neq 0.$$

En posant $\varepsilon = Y - X^T\theta_0$ on a

$$\mathbb{E}(Z(Y - X^T\beta_0)) = 0 \Leftrightarrow \mathbb{E}(ZY) = \mathbb{E}(ZX^T)\beta_0.$$

Sous réserve d'inversibilité de $\mathbb{E}(ZX^T)$, on a

$$\beta_0 = \mathbb{E}(ZX^T)^{-1}\mathbb{E}(ZY).$$

Soit $(X_n, Y_n, Z_n)_{n\in\mathbb{N}^*}$ une suite *i.i.d.* de loi $P = \mathbb{P}^{(X,Y,Z)}$. On dispose d'une information auxiliaire I portant sur au moins une des variables X, Y, Z. L'estimateur à variables instrumentales informé est donné par

$$\hat{\theta}_n^I = \left(\mathbb{Z}^T Q \mathbb{X}\right)^{-1} \mathbb{Z}^T Q Y$$

où $\mathbb{Z} = (Z^1, \cdots, Z^p)$. La proposition suivante porte sur la consistance forte de cet estimateur.

Théorème 7.3.6. *Supposons que* $\mathbb{E}(ZX^T)$ *est inversible. Alors*

$$\hat{\theta}_n^I \xrightarrow[n\to+\infty]{p.s.} \beta_0.$$

Démonstration.
Observons que

$$\begin{aligned}\hat{\theta}_n^I &= \left(\sum_{i=1}^n q_i Z_i X_i^T\right)^{-1} \sum_{i=1}^n q_i Y_i Z_i \\ &= \left(\sum_{i=1}^n q_i Z_i X_i^T\right)^{-1} \left(\sum_{i=1}^n q_i Z_i X_i^T\right)\beta_0 + \left(\sum_{i=1}^n q_i Z_i X_i^T\right)^{-1} \sum_{i=1}^n q_i (Y_i - \langle X_i, \beta_0\rangle) Z_i \\ &= \beta_0 + \left(\sum_{i=1}^n q_i Z_i X_i^T\right)^{-1} \sum_{i=1}^n q_i (Y_i - \langle X_i, \beta_0\rangle) Z_i.\end{aligned}$$

Par résultat du type Glivenko-Cantelli pour la mesure empirique informée (théorème 3.1.1, 3.2.2 et 4.1.5), le second terme tend p.s. vers 0 *puisque* $\mathbb{E}(\varepsilon Z) = 0$. □

De plus, nous avons aussi un résulat de normalité asymptotique pour l'estimateur $\hat{\theta}_n^I$.

Théorème 7.3.7. *Supposons que* $\mathbb{E}(ZX^T)$ *est inversible. Posons* $f_{\beta_0}(x, y, z) = (y - \langle x, \beta_0\rangle)z$ *et* $U = \mathbb{E}(XZ^T)^{-1}$. *Alors on a*

$$\sqrt{n}(\hat{\theta}_n^I - \beta_0) \Rightarrow \mathcal{N}\left(0, UVar_P\, G_I\,(f_{\beta_0})U^T\right).$$

De plus, on obtient l'inégalité suivante

$$UVar_P\, G_I\,(f_{\beta_0})U^T \leq UVar_P\, G\,(f_{\beta_0})U^T.$$

Dans le cas où l'information auxiliaire est donnée par une fonctionnelle $g(P)$ *(ou par des estimations issues de sources indépendantes), la variance s'exprime de la manière suivante*

$$\begin{aligned}Var_P\, G_I\,(f_{\beta_0}) &= Var_P\, f_{\beta_0} - cov_P(\varphi_g(P), f_{\beta_0})^T \Sigma^{-1} cov_P(\varphi_g(P), f_{\beta_0}) \\ &= Var_P\, f_{\beta_0} - \mathbb{E}(\varepsilon Z \varphi_g(P)(X, Y, Z)^T)\Sigma^{-1}\mathbb{E}(\varepsilon \varphi_g(P)(X, Y, Z)Z^T).\end{aligned}$$

Remarque 7.3.8. *Toute information auxiliaire portant sur* Z *n'améliore pas l'estimation. En effet cela vient du fait que* $\mathbb{E}(\varepsilon|Z) = 0$. *Cependant cette fois-ci une information* ***portant sur*** X *peut améliorer l'estimation puisque cette variable est endogène.*

Démonstration.
Remarquons que

$$\sqrt{n}(\hat{\theta}_n^I - \beta_0) = \left(\sum_{i=1}^n q_i Z_i X_i^T\right)^{-1} \sqrt{n} \sum_{i=1}^n q_i (Y_i - \langle X_i, \beta_0\rangle) Z_i.$$

Par un résulat du type Donsker pour la mesure empirique informée (théorème 3.1.2, 3.2.3 et 4.1.6) et en utilisant le fait que $\mathbb{E}(\varepsilon Z) = 0$ *où* $\varepsilon = Y - \langle X, \beta_0 \rangle$*, on a*

$$\sqrt{n}\sum_{i=1}^{n} q_i(Y_i - \langle X_i, \beta_0\rangle)Z_i = \sqrt{n}\mathbb{P}_n^I f_{\beta_0} \Rightarrow G_I(f_{\beta_0}).$$

Par résultat du type Glivenko-Cantelli pour la mesure empirique informée (théorème 3.1.1, 3.2.2 et 4.1.5) et par continuité de l'inverse, on a

$$\left(\sum_{i=1}^{n} q_i Z_i X_i^T\right)^{-1} \xrightarrow[n\to+\infty]{p.s.} \left(\mathbb{E}(ZX^T)\right)^{-1} =: U.$$

Par le lemme de Slutsky, on obtient finalement que

$$\sqrt{n}(\hat{\theta}_n^I - \beta_0) \Rightarrow UG_I(f_{\beta_0}) = \mathcal{N}\left(0, UVar_P\, G_I\,(f_{\beta_0})U^T\right).$$

□

7.3.5 Tests statistiques asymptotiques sur le modèle informé (avec ou sans instrument)

Dans cette sous-section, nous proposons des tests statistiques asymptotiques afin de tester la significativité globale du modèle. De plus, nous ne faisons pas d'hypothèse sur la gaussianité des erreurs. Par le théorème 7.3.3 et 7.3.7, la matrice de covariance limite est

$$V = UVar_P\, G_I\,(f_{\beta_0})U^T$$

où $U = (\mathbb{E}(XX^T))^{-1}$ ou $(\mathbb{E}(ZX^T))^{-1}$ et $f_{\beta_0}(x,y) = (y - \langle x, \beta_0\rangle)x$ ou $f_{\beta_0}(x,y,z) = (y - \langle x, \beta_0\rangle)z$ dans le cas d'un modèle linéaire avec ou sans variable instrumentale. On note V_n l'estimateur empirique de V en remplaçant P par $\mathbb{P}_n$. Ainsi remarquons que V_n est une matrice symétrique. Posons $\hat{\beta}_n^I := \hat{\theta}_n^I$.

Test de significativité d'un paramètre

On s'intéresse au problème de test suivant

$$\mathcal{H}_0 : \beta_j = 0 \text{ contre } H_1 : \beta_j \neq 0$$

pour $j \in [\![1, p]\!]$. Par le théorème 7.3.3 et 7.3.7, on a

$$\sqrt{n}(\hat{\beta}_{n,j}^I - \beta_j) \Rightarrow \mathcal{N}(0, V_{jj})$$

où V_{jj} est j-ème coordonnée sur la diagonale de V. On suppose $V_{jj} \neq 0$ et on pose

$$L_n = V_{n,jj}^{-\frac{1}{2}}\sqrt{n}\hat{\beta}_{n,j}^I.$$

Observons que sous $\mathcal{H}_0$ par le lemme de Slutsky $L_n \Rightarrow \mathcal{N}(0,1)$ et sous $\mathcal{H}_1$, $|L_n| \xrightarrow[n\to+\infty]{p.s.} +\infty$. Fixons $\alpha \in]0,1[$. Alors le test défini par

$$T_n = 1_{|L_n| > z_{1-\alpha/2}}$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une $\mathcal{N}(0,1)$ est un test asymptotique de niveau α.

Test de significativité de plusieurs paramètres

Quitte à réindicer, on suppose que les paramètres à tester commencent à partir de p_0+1 pour un $p_0 \in [\![1, p-1]\!]$ fixé. On s'intéresse donc au problème de test suivant

$$\mathscr{H}_0 : \beta_{p_0+1} = ... = \beta_p = 0 \text{ contre } H_1 : \exists j \in \{p_0+1, ..., p\},\ \beta_j \neq 0.$$

Par le théorème 7.3.3 et 7.3.7 et la méthode delta pour la fonction $h(x_1, ..., x_p) = (x_{p_0+1}, ..., x_p)^T$ avec $(x_1, ..., x_p) \in \mathbb{R}^p$, on a

$$\sqrt{n}(h(\hat{\beta}_n^I) - h(\beta)) \Rightarrow \mathscr{N}(0, Dh(\beta)VDh(\beta)^T)$$

avec $Dh(\beta) = (\mathbf{0}_{p-p_0,p_0} | Id_{p-p_0})$. Ainsi

$$Dh(\beta)VDh(\beta)^T =: \widetilde{V}$$

où $\widetilde{V} = \left(V_{i,j}\right)_{i,j \in [\![p_0+1,p]\!]}$. On note $\widetilde{V}_n$ l'estimateur empirique de la matrice carrée $\widetilde{V}$. Supposons que par $\tilde{V}$ est inversible. Alors puisque $\widetilde{V}_n$ convergence p.s. vers $\widetilde{V}$ alors p.s. pour n suffisamment grand $\widetilde{V}_n$ est aussi inversible. On pose

$$L_n := \| \tilde{V}_n^{-\frac{1}{2}} \sqrt{n} h(\hat{\beta}_n^I) \|^2 = nh(\hat{\beta}_n^I)^T \tilde{V}_n^{-1} h(\hat{\beta}_n^I).$$

Alors sous $\mathscr{H}_0$

$$L_n \Rightarrow \chi^2(p-p_0),$$

et sous $\mathscr{H}_1$

$$L_n \xrightarrow[n \to +\infty]{p.s.} +\infty.$$

Ainsi pour $\alpha \in]0,1[$ le test défini par

$$T_n = 1_{L_n > c_{1-\alpha}}$$

où $c_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une $\chi^2(p-p_0)$ est un test asymptotique de niveau α.

Remarque 7.3.9. *Dans le cadre d'un modèle linéaire gaussien, il est difficile de déterminer la loi de* $\hat{\theta}_n^I$

$$\begin{aligned}\hat{\theta}_n^I &= \left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q Y \\ &= \theta + \left(\mathbb{X}^T Q \mathbb{X}\right)^{-1} \mathbb{X}^T Q \varepsilon\end{aligned}$$

du fait que les poids $(q_i)_{i \in [\![1,n]\!]}$ *sont aléatoires et dépendent de* X *et* Y*. Ainsi il est difficile d'obtenir des tests non asymptotiques dans le cadre gaussien pour l'estimateur des moindres carrés informés.*

7.4 Pistes méthodologiques : Analyse des données et apprentissage statistique avec informations auxiliaires

Dans cette section, nous proposons des pistes méthodologiques pour injecter de l'information auxiliaire en analyse des données et en apprentissage statistique.

7.4.1 Problème de bandits avec information auxiliaire

Nous souhaitons injecter une information auxiliaire dans le contexte d'un bandit stochastique. Tout d'abord, énonçons le cadre. On se donne un ensemble de bras $\mathscr{A}$. Le choix du bras $a \in \mathscr{A}$ à l'instant $t \in \mathbb{N}^*$ nous permet d'observer la variable $X_{a,t} \sim P_a$ où P_a est un loi de probabilité sur $\mathscr{X}$ inconnue et nous accorde la récompense $\psi(X_{a,t})$ où $\psi : \mathscr{X} \to \mathbb{R}$ une fonction intégrable. On suppose que les variables aléatoires $(X_{a,t})_{t\in\mathbb{N}^*}$ sont indépendantes. L'objectif est de choisir le bras optimal $a^* \in \mathscr{A}$ c'est à dire celui tel que

$$a^* = arg \max_{a\in\mathscr{A}} P_a\psi$$

afin de maximiser la récompense cumulée

$$\mathbb{E}S_n = \mathbb{E}\left(\sum_{t=1}^{T} \psi(X_{A_t,t})\right).$$

Cela revient à minimiser le regret

$$R_n = T \max_{a\in\mathscr{A}} P_a\psi - \mathbb{E}S_n.$$

Dans le cas où $\mathscr{X} = \mathbb{R}$, généralement $\psi = Id$. Dans la suite, on considérera le cas réel.

Nous devons nous doter d'une *stratégie optimale* concernant le choix du bras à l'étape t en utilisant **l'information disponible avant l'étape** t. C'est à dire une politique $\phi = (\phi_t)_{t\geq 1}$ telle que

$$A_t = \phi_t\left(X_{A_1,1}, \cdots, X_{A_{t-1},t-1}\right).$$

Cependant ici on n'exploite pas l'information auxiliaire mais seulement l'information séquentielle des données. L'idée maintenant est d'incorporer de l'information auxiliaire en plus de l'information des données afin d'avoir une politique beaucoup plus optimale c'est à dire

$$A_t = \phi_t\left(X_{A_1,1}, \cdots, X_{A_{t-1},t-1}, \widetilde{I}_1, \cdots, \widetilde{I}_{t-1}, \widetilde{I}_t\right)$$

où $\widetilde{I}_s$ est l'information auxiliaire donnée au temps $s \in [\![1, t]\!]$.

Remarque 7.4.1. *Plusieurs travaux ont été menés afin d'incorporer de l'information auxiliaire dans le problème du bandit stochastique par exemple [14],[36]. Néanmoins, l'information auxiliaire est différente que celle traitée dans cette thèse. Par exemple :*

1. *On pose un ensemble contexte $\mathscr{C}$ et on suppose que $X_{a,t}$ s'écrit comme une fonction du contexte et du bras a. A cela on rajoute généralement un bruit. Ainsi on traite différents types de fonctions (linéaire, logistique etc).*
2. *Ils supposent que périodiquement ou aléatoirement, l'agent peut connaître la récompense d'au moins 2 bras.*

Dans notre cadre on considérera les informations auxiliaires suivantes :

- On dispose d'une information auxiliaire forte ou faible portant sur un certain nombre de lois P_a, $a \in \mathscr{A}$.
- On suppose que cette même expérience a été effectuée par d'autres entités (joueurs etc) et on a accès à leurs données (récompenses) et leurs stratégies employées.

Premier cas : Fixons $t \in [\![1, T]\!]$. À l'étape t, nous disposons d'une information auxiliaire $\widetilde{I}$ portant sur $P_{a_1}, ..., P_{a_r}$ pour $r \in \mathbb{N}^*$. Ainsi l'information auxiliaire I peut être décomposée en r informations auxiliaires $I_1, ..., I_r$ portant respectivement sur $P_{a_1}, \cdots, P_{a_r}$. À l'étape t, nous avons exploré $T_{a_k}(t-1)$ fois le bras a_k pour $k \in [\![1, r]\!]$. Pour chaque $k \in [\![1, r]\!]$, on dispose de l'information I_k et on observe les récompenses notées $\tilde{X}_{k,1}, ..., \tilde{X}_{k,T_{a_k}(t-1)}$ qui forment un échantillon *i.i.d.* de loi P_{a_k}. Les stratégies π utilisent comme estimation de $\mu_{a_k} = P_{a_k} x$,

$$\frac{1}{T_{a_k}(t-1)} \sum_{i=1}^{T_{a_k}(t-1)} \tilde{X}_{k,i}$$

afin de pouvoir décider quel bras choisir à l'étape t. Afin d'améliorer l'efficacité de ces stratégies π, on propose de remplacer cette estimation par

$$\mathbb{P}^{I_k}_{T_{a_k}(t-1)} x.$$

Second cas : Il faut construire une stratégie π^* en fonction des stratégies *auxiliaires* afin d'exploiter cette information auxiliaire. Par exemple, plaçons nous dans le modèle à 2 bras. Supposons qu'on ait accès à une stratégie π d'un autre joueur qui a exploré majoritairement le bras 1 et peu le bras 2. Dans ce cas, la stratégie informée explorera dans un premier temps le bras 2 et prendra plus rapidement la décision de conserver le bras gagnant.

7.4.2 Analyse en composantes principales avec informations auxiliaires

L'**Analyse en Composantes Principales** (ACP) consiste à projeter les variables sur un espace de dimension plus petit tout en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales. Le cadre est le suivant. Soit $X_1, ..., X_n$ un échantillon i.i.d de taille $n \in \mathbb{N}^*$ à valeurs dans $\mathbb{R}^p$ où $p \in \mathbb{N}^*$ est le nombre de variables. On note P la loi de X.

Remarque 7.4.2. *En statistique descriptive, on a accès à l'ensemble de la population finie* $\Omega = [\![1, N]\!]$ *avec* $N \in \mathbb{N}^*$ *et* $X_i := X(i)$ *pour* $i \in \Omega$. *Ainsi* $P = \frac{1}{n}\sum_{i=1}^{n} \delta_{X(i)}$.

Nous souhaitons trouver une droite affine dans $\mathbb{R}^p$ sur laquelle le projeté de X est le plus dispersé. Plus précisément on cherche $v^* \in \mathbb{R}^p$ tel que

$$v^* \in arg \max_{\|v\|=1} Var_P(\langle v, X\rangle) = arg \max_{\|v\|=1} v^T \Sigma v.$$

On peut montrer que la solution est un vecteur propre unitaire associé à la plus grande valeur propre λ^* de la matrice de covariance Σ. Puis on itère ce procédé en cherchant w tels que $\|w\| = 1$ et w linéairement indépendant de v. Si l'on souhaite projeter sur un espace de dimension $q < p$, il suffit de prendre une base orthonormée de chaque espace propre classé par ordre décroissant jusqu'à obtenir q vecteurs.

En statistique descriptive, on a accès à la matrice de covariance Σ et donc toute notion d'information auxiliaire perd son sens. Cependant en statistique inférentielle, on a rarement accès à la matrice de covariance Σ. De ce fait, on remplace Σ par sa version empirique Σ_n puis on cherche

$$v_n^* \in arg \max_{\|v\|=1} Var_{\mathbb{P}_n}(\langle v, X\rangle) = arg \max_{\|v\|=1} v^T \Sigma_n v.$$

Dans le cas où on dispose d'une information auxiliaire I, nous proposons de remplacer Σ par Σ_n^I (matrice de covariance sous $\mathbb{P}_n^I$) à la place de Σ_n et de chercher

$$v_{n,I}^* \in arg \max_{\|v\|=1} Var_{\mathbb{P}_n^I}(\langle v, X\rangle) = arg \max_{\|v\|=1} v^T \Sigma_n^I v.$$

7.4.3 Réseaux de neurones

On souhaite injecter une information auxiliaire dans le cadre des réseaux de neurones. Notre échantillon d'entraînement de taille $n \in \mathbb{N}^*$, noté $(X_1, Y_1), ..., (X_n, Y_n)$, est à valeurs dans le produit cartésien de deux ensembles mesurables $\mathscr{X} \times \mathscr{Y}$. Un réseau de neurones prend en entrée des données X_i et sort en sortie un vecteur Y qui est une variable réponse. Ce réseau est composé de couches cachées qui sont reliées par des fonctions. Ces dernières dépendent de divers paramètres (matrice de poids, biais et fonction d'activation). En général, la fonction d'activation est fixée à l'avance. Ainsi la fonction résultante de ce réseau de neurones dépend d'une matrice de poids W et de biais b. On la notera $f_{W,b}$. Les poids et le biais des couches cachées sont obtenus par étude d'un problème d'optimation. Plus précisément pour une fonction de coût $c : \mathscr{X} \to \mathscr{Y}$ donnée, le réseau de neurones est entraîné idéalement en minimisant la fonction

$$(W, b) \mapsto \mathbb{E}\left(c(Y, f_{W,b}(X))\right).$$

Cependant nous n'avons pas accès à la loi $P = \mathbb{P}^{(X,Y)}$. Ainsi les réseaux de neurones sont entraînés en résolvant le problème d'optimisation suivant

$$arg \min_{W,b} \mathbb{P}_n\left(c(y, f_{W,b}(x)\right)$$

où $\mathbb{P}_n$ est la mesure empirique associée à l'échantillon d'entraînement. On pourra consulter à ce sujet [7].

Dans le cas où on dispose d'une information auxiliaire I concernant la loi P, nous proposons de s'intéresser au problème d'optimisation suivant

$$arg \min_{W,b} \mathbb{P}_n^I\left(c(y, f_{W,b}(x))\right)$$

où $\mathbb{P}_n^I$ est la mesure empirique informée.

7.4.4 Bagging informé

D'un point de vue méthodologique, nous pouvons utiliser la mesure empirique informée afin de tirer des échantillons bootstrap informés. Le bagging (bootstrap aggregating) est un méta-algorithme qui a pour objectif d'améliorer la stabilité et la précision des algorithmes d'apprentissage. Nous souhaitons modifier ce méta-algorithme afin d'incorporer une information auxiliaire I.

On dispose d'un échantillon $S_n = \{(X_1, Y_1), \cdots, (X_m, Y_m)\}$ de taille $n \in \mathbb{N}^*$ à valeurs dans $\mathscr{X} \times \mathscr{Y}$ et de règles d'apprentissage $\phi_n : S_n \mapsto h_n$ où $h_n : \mathscr{X} \to \mathscr{Y}$ est une règle de décision. On suppose que $\mathbb{P}_n^I$ est une mesure de probabilité (proposition 2.3.3). Les étapes de ce méta-algorithme sont les suivantes :

1. Tirer M échantillons bootstrap $\tilde{S}_n^1, \cdots, \tilde{S}_n^M$ selon $\mathbb{P}_n^I$.
2. Pour chaque $1 \leq i \leq M$, calculer le classifieur $\hat{h}_n^i = \phi_n(\tilde{S}_n^i)$.
3. Agréger ces classifieurs afin d'obtenir un meilleur classifieur. Cela se fait de la manière suivante :
   - Pour la classification, cela se fait par vote majoritaire
   $$\hat{h}_n(x) = arg \max_{y \in \mathscr{Y}} \sum_{i=1}^{M} 1_{\hat{h}_n^i(x)=y}.$$
   - Pour la régression
   $$\hat{h}_n(x) = \frac{1}{M} \sum_{i=1}^{M} \hat{h}_n^i(x).$$

Ce méta-algorithme s'applique à tous les modèles d'apprentissage (réseaux de neurones, forêts d'arbres décisionnels etc).

7.4.5 Algorithme des k-plus proches voisins avec informations auxiliaires

Soit $(X_1, Y_1), ..., (X_n, Y_n)$ un échantillon *i.i.d.* de taille $n \in \mathbb{N}^*$ de loi P à valeurs dans l'espace $\mathscr{X} \times \mathscr{Y}$. Fixons $k \in \mathbb{N}^*$ et rappelons la règle des k plus proches voisins. Pour tout $x \in \mathscr{X}$, on trouve ses k plus proches voisins pour une distance d. On note ces k plus proches voisins $X_{(1),x}, ..., X_{(k),x}$ où

$$d(X_{(1),x}, x) \leq \cdots \leq d(X_{(k),x}, x).$$

Enfin notons $Y_{(j),x}$ le "label" associé $X_{(j),x}$. On peut alors calculer la statistique suivante en fonction du contexte :

- Dans un contexte de classification, la statistique est
$$\hat{f}_n(x) = arg \max_{y \in \mathscr{Y}} \frac{1}{k} \sum_{j=1}^{k} 1_{Y_{(j),x}=y}, \ x \in \mathscr{X}.$$
- Dans un contexte de régression, la statistique est
$$\hat{f}_n(x) = \frac{1}{k} \sum_{j=1}^{k} Y_{(j),x}, \ x \in \mathscr{X}.$$

On suppose désormais qu'on dispose d'une information auxiliaire I sur P. On note $\mathbb{P}_n^I = \sum_{i=1}^n q_i \delta_{(X_i,Y_i)}$ la mesure empirique informée. On pose pour $j \in [\![1,k]\!]$

$$w_{j,x} = \frac{q_{(j),x}}{\sum_{l=1}^k q_{(l),x}}$$

où $q_{(j),x}$ est le poids associé à $(X_{(j),x}, Y_{(j),x})$. Ainsi la règle des k plus proches voisins avec informations auxiliaires est définie comme :

- Dans un contexte de classification, la statistique est

$$\hat{f}_n^I(x) = arg\max_{y\in\mathscr{Y}} \sum_{j=1}^k w_{j,x} 1_{Y_{(j),x}=y},\ x \in \mathscr{X}.$$

- Dans un contexte de régression, la statistique est

$$\hat{f}_n^I(x) = \sum_{j=1}^k w_{j,x} Y_{(j),x},\ x \in \mathscr{X}.$$

On remarque que si pour tout $i \in [\![1,n]\!]$, $q_i = q = \frac{1}{n}$ alors $w_j = \frac{1}{k}$.

Remarque 7.4.3. *Dans le cas d'ajout de données auxiliaires, c'est le classement des k plus proches voisins qui diffère.*

7.4.6 Apprentissage non supervisé : k-moyennes avec informations auxiliaires

Cette sous-section est inspirée d'un article d'Aurélie Fischer [19] intitulé *Deux méthodes d'apprentissage non supervisé : synthèse sur la méthode des centres mobiles et présentation des courbes principales.*

Énonçons le principe de la **quantification**. Fixons $d \in \mathbb{N}^*$. On se donne un vecteur aléatoire X de $\mathbb{R}^d$ de loi P tel que $\mathbb{E}\|X\|^2 < +\infty$. Soit $k \in \mathbb{N}^*$ un entier. Un k-**quantificateur** est une application borélienne $q : \mathbb{R}^d \to \mathscr{C}$ où $\mathscr{C} = \{c_1, \cdots, c_l\}$ avec $l \in [\![1,k]\!]$ est un sous ensemble de $\mathbb{R}^d$ appelé **table de codage**. L'idée de la quantification consiste à limiter l'erreur commise en remplaçant X par $q(X)$. Cette erreur est évaluée par la **distorsion**

$$W(q) = \mathbb{E}\left(\|X - q(X)\|^2\right).$$

Remarque 7.4.4. *Cela se généralise à une divergence de type Bergman par exemple. À ce sujet, on pourra consulter la thèse d'Aurélie Fischer [18].*

Chercher le meilleur k-quantificateur revient à minimiser cette distorsion c'est à dire

$$W^* = \inf_{q \in Q_k} W(q)$$

où Q_k est l'ensemble de tous les k-quantificateurs. Un k-quantificateur q^* tel que $W(q^*) = W^*$ est dit optimal.

Dans le contexte statistique, la loi P est inconnue. De ce fait, on remplace P par la mesure empirique

$$W_n(q) = \mathbb{P}_n \|x - q(x)\|^2$$

et on recherche

$$W_n^* = \inf_{q \in Q_k} W_n(q).$$

Remarquons que chaque k-quantificateur est déterminé par sa table de codage $\mathscr{C}$ et une partition de $\mathbb{R}^d$ en cellules $S_j = \{x \in \mathbb{R}^d,\ q(x) = c_j\}$. De ce fait, il suffit de donner sa table de codage et sa partition. Un type de partition important est la **partiton de Voronoi** définie comme suit

$$S_1 = \{x \in \mathbb{R}^d,\ \|x - c_1\| \leq \|x - c_p\|\ p = [\![1, l]\!]\}$$

et pour $j = [\![2, l]\!]$

$$S_j = \left\{x \in \mathbb{R}^d,\ \|x - c_j\| \leq \|x - c_p\|\ p = 1, \ldots, l\right\} \setminus \cup_{m=1}^{j-1} S_m.$$

Autrement dit x est affecté à la cellule S_j si, et seulement si, cet élément est plus proche de c_j que de tous les autres centres, et en cas d'égalité, la cellule de plus petit indice est choisie. Un quantificateur associé à la partition de Voronoi est appelé **quantificateur des plus proches voisins.** Ce dernier minimise la distorsion sur l'ensemble des k-quantificateurs ayant même table de code comme l'énonce le lemme suivant.

Lemme 7.4.5. *Soit q un k-quantificatueur de table de codage $\mathscr{C}$ où $l \leq k$ et q' le quantificateur des plus proches voisins ayant même table de codage. Alors, on a*

$$W(q') \leq W(q).$$

S'il existe un k-quantificateur optimal, c'est donc nécessairement un quantificateur des plus proches voisins. Pour construire un bon quantificateur, il suffit par conséquent de considérer cette classe particulière de quantificateurs. Ainsi, trouver un quantificateur optimal, c'est trouver la table de codage optimale minimisant la distorsion réécrite en fonction de $\mathscr{C}$

$$W(\mathscr{C}) = \mathbb{E}\left(\min_{j=1,\cdots,k} \|X - c_j\|\right).$$

D'autre part, on sait décrire la meilleure table de codage pour un quantificateur de partition donnée :

Lemme 7.4.6. *Soit q un quantificateur de partition associée $\{S_j\}_{j=1,\dots,l}$ où $l \leq k$ avec $P(S_j) > 0$ pour tout j. Si q' est un quantificateur de même partition dont la table de codage $\{c'_1, \cdots, c'_l\}$ est définie par*

$$c'_j = \mathbb{E}(X|X \in S_j)$$

alors

$$W(q') \leq W(q).$$

Ce lemme nous donne une méthode algorithmique pour mettre à jour les centres. L'existence de quantificateurs optimaux a été démontrée par Pollard [33] :

Théorème 7.4.7. *Il existe une table de codage optimale $\mathscr{C}^*$ telle que $W(\mathscr{C}^*) = W^*$, et donc un quantificateur des plus proches voisins q^* tel que $W(q^*) = W^*$. En particulier il existe $\mathscr{C}_n^*$ tel que $W_n(\mathscr{C}_n^*) = W_n^*$.*

Enfin on a aussi un résultat de consistance

$$W(\mathscr{C}_n^*) \xrightarrow[n\to+\infty]{p.s.} W^*$$

En pratique, trouver un minimiseur exact de la distorsion est un problème que l'on ne peut résoudre en temps polynomial, mais les lemmes précédents montrent que la solution peut être approchée à l'aide d'un algorithme itératif, reposant sur deux étapes, au cours desquelles la table de codage et la partition sont actualisées successivement. Puisque la loi P est inconnue, on la remplace par la mesure empirique $\mathbb{P}_n$ et l'algorithme est :

- Choisir k centres $c_1^{(1)}, \cdots, c_k^{(1)}$.
- Répéter jusqu'à ce qu'il y ait convergence (centres inchangés) :
  1. Affecter chaque observation à la partition la plus proche

     $$\{X_1, \cdots, X_n\} \cap S_j^{(t)}$$

     où $(S_j^{(t)})_{j\in[\![1,k]\!]}$ représente la partition de Voronoi associée aux centres $c_1^{(t)}, \cdots, c_k^{(t)}$.
  2. Mettre à jour les centres

     $$c_j^{(t+1)} = \frac{\mathbb{P}_n x 1_{S_j^{(t)}}}{\mathbb{P}_n(S_j^{(t)})} = \frac{1}{|S_j^{(t)}|} \sum_{X_i \in S_j^{(t)}} X_i,\ j \in [\![1,k]\!].$$

L'injection d'information auxiliaire se fera en remplaçant la mesure empirique par la mesure empirique informée. Ainsi **l'algorithme des k-moyennes avec informations auxiliaires** est le suivant :

- Choisir k centres $c_1^{(1)}, \cdots, c_k^{(1)}$.

- Répéter jusqu'à ce qu'il y ait convergence (centres inchangés) :
  1. Affecter chaque observation à la partition la plus proche

$$\{X_1, \cdots, X_n\} \cap S_j^{(t)}$$

où $(S_j^{(t)})_{j \in [\![1,k]\!]}$ représente la partition de Voronoi associée aux centres $c_1^{(t)}, \cdots, c_k^{(t)}$.

Remarque 7.4.8. *Dans le cas d'ajout de données auxiliaires, il faut affecter les données auxiliaires à la cellule correspondante.*

  2. Mettre à jour les centres

$$c_j^{(t+1)} = \frac{\mathbb{P}_n^I x 1_{S_j^{(t)}}}{\mathbb{P}_n^I (S_j^{(t)})} \; j \in [\![1, k]\!].$$

Appendices

Annexe A : Rudiments de géométrie différentielle et riemannienne

L'objectif de cette annexe est d'introduire les objets classiques de la géométrie différentielle des variétés afin de pouvoir les utiliser librement en géométrie de l'information. Cette annexe est un résumé du polycopié [21].

A.1 Variétés topologiques et différentielles

Tout d'abord définissons la notion de **variété topologique**.

Definition A.1.1. Soit $(M, \mathscr{T})$ un espace topologique séparé ayant une base dénombrable d'ouverts. M est une **variété topologique** s'il existe un entier $m \in \mathbb{N}$ tel que pour tout $p \in M$, il existe un voisinage ouvert U de p homéormorphe à un ouvert de $\mathbb{R}^m$. Autrement dit, il existe un homéomorphisme $\phi : U \to \phi(U) \subset \mathbb{R}^m$.

On appelle (U, ϕ) une **carte locale** ou **coordonnées locales** ou **système de coordonnées locales**. L'indice m est appelé la dimension de M. Elle est bien définie par le théorème de Brouwer. ($\mathbb{R}^n$ est homéomorphe à $\mathbb{R}^m$ si et seulement si $n = m$, le cas difféomorphe est beaucoup plus facile à traiter). On note M^n une variété de dimension n.

Remarquons qu'on peut définir la variété topologique sans topologie de départ. La topologie sera construite à l'aide de cartes et elle rendra homéomorphe les applications de carte. Une variété topologique a quelques propriétés intéressantes. Elle est séparable, métrisable, localement compact et connexe par arcs. Néanmoins, on a besoin d'une structure différentielle sur cette variété afin de pouvoir mesurer des longueurs par exemple. Cette structure se définit à l'aide de cartes.

Un C^r-atlas est une collection (U_α, x_α) de cartes locales qui recouvrent la variété topologique M de dimension m et pour tout α, β

$$x_\alpha \circ x_\beta^{-1}$$

sont C^r. On les appelle **les changements de cartes** ou **fonctions de transition** ou encore **changement de coordonnées**. Une carte locale (U, x) est dite compatible avec un C^r-atlas $\mathscr{A}$ si $\mathscr{A} \cup (U, x)$

est un C^r-atlas. Un atlas est dit **maximal** s'il contient toutes les cartes locales compatibles avec lui. Pour tout atlas, il existe un atlas maximal qui le contient. Il suffit de prendre la réunion. Un C^r **atlas maximal** sur M est appelée une C^r-structure differentielle sur M. Donc **une variété différentiable de classe** C^r $(M, \mathscr{A})$ est une variété topologique de dimension m avec $\mathscr{A}$ est une structure différentielle de classe C^r.

Exemple A.1.1. Donnons quelques exemples.

- Soit U un ouvert de $\mathbb{R}^N$. On prend alors la carte (U, Id) qui recouvre bien. Cela définit une structure différentielle sur $\mathbb{R}^N$. Prenons maintenant la sphère S^N dans $\mathbb{R}^{N+1}$. On munit la sphère de la topologie induite de $\mathbb{R}^{N+1}$. On note $N = (1,0) \in \mathbb{R} \times \mathbb{R}^N$ le pôle nord et notons $S = (-1,0)$ le pôle sud. On définit les applications suivantes

$$\begin{aligned} x_N : & U_N \to \mathbb{R}^N \\ & p \mapsto \frac{1}{1-p_1}(p_2, \cdots, p_N), \end{aligned}$$

$$\begin{aligned} x_S : & U_S \to \mathbb{R}^N \\ & p \mapsto \frac{1}{1+p_1}(p_2, \cdots, p_N) \end{aligned}$$

  où $U_N = (S^N)^{\backslash N}$ et de même pour U_S. Les changements de cartes sont donnés par

$$x_N \circ x_S^{-1}, x_S \circ x_N^{-1} : x \mapsto \frac{x}{|x|^2}$$

  pour tout $x \in \mathbb{R}_*^{N+1}$. Donc l'atlas $\{(U_N, x_N), (U_S, x_S)\}$ est C^∞ et définit une structure C^∞ sur la sphère.
- Un autre exemple important est celui de l'espace projectif $\mathbb{P}^n(\mathbb{R})$ qui est une variété C^∞ de dimension n. Remarquons qu'il existe une unique structure différentielle sur $\mathbb{R}^n$ et S^1. De plus le produit cartésien de deux variétés est une variété dont la dimension est la somme des deux dimensions. Ainsi le tore $\mathbb{T}^2 = S^1 \times S^1$ est une variété de dimension deux.
- L'espace des paramètres Θ d'une famille $(P_\theta)_{\theta \in \Theta}$ est généralement une variété. Par exemple pour les lois normales $\Theta = \mathbb{R} \times \mathbb{R}_+^*$, loi exponentielle $\mathbb{R}_+^*$ etc. L'idée de la géométrie de l'information va consister à utiliser une distance *intrinsèque* à la géométrie du problème.

A.2 Différentiabilité, sous variétés et plongements

La différentiabilité d'une fonction entre variétés différentielles s'établit par les cartes. On entend par **différentiable** au moins C^1. Soit $f : M^m \to N^n$ une application. On dira que f est différentiable au point $p \in M^m$ s'il existe une carte (U, ϕ) au voisinage de p et une carte (V, ψ) au voisinage de $f(p)$

telles que $f(U) \subset V$ et

$$\psi \circ f \circ \phi^{-1} : \phi(U) \to \psi(V)$$

est différentiable. Une application f est un **difféomorphisme** si f est différentiable, bijective et sa réciproque est différentiable. Notons que la carte locale ψ est une difféomorphisme, en effet $\phi \circ \phi^{-1}$ est différentiable du fait de la structure différentielle sur la variété. Remarquons que cette définition est indépendante du choix de cartes. En effet on peut le voir en écrivant

$$f^{\phi\psi} = \psi_1 \circ f \circ \phi_1^{-1} = \psi_1 \circ \psi^{-1} \circ \psi \circ f \circ \phi^{-1} \circ \phi \circ \phi_1^{-1}$$

et en utilisant que les changements de cartes sont réguliers. Enfin mentionnons que si $M = I$ un intervalle ouvert, on obtient la notion de **courbe tracée sur** N. On peut définir la notion de **rang** d'une application différentiable f par le rang de $\psi \circ f \circ \phi^{-1}$ (qui ne dépend pas du choix de cartes en utilisant la régularité des changements de cartes et du théorème de différentiabilité des fonctions composées!). Faisons une remarque évidente, une appliation différentiable est continue. Pour le voir il suffit d'écrire dans des cartes (ou une si $\phi = \psi$) $f = \psi^{-1}(\psi f \phi^{-1})\phi$. Définissons la notion **d'immersion** et de **submersion**. On dit que f est une **immersion** si elle est différentiable et $rg(f) = dim(M)$, autrement dit f différentiable et $Df^{\phi\psi}(x)$ est injective pour tout x. De même f est une **submersion** si f différentiable et $Df^{\phi\psi}(x)$ est surjective pour tout x. Dans ce cas $\dim N \leq \dim M$. Notons qu'une immersion n'est pas forcément injective. En effet posons $f(t) = (\cos(2\pi t), \sin(2\pi t))$. Cela nous amène à définir la notion de **plongement**. Une application f est un **plongement** si f est une immersion et un homéomorphisme sur son image $f(M)$ pour la topologie induite. Remarquons qu'une immersion injective est déjà une bijection sur son image, pour qu'elle soit un plongement il suffit de montrer que f^{-1} est continue. Par exemple $t \in \mathbb{R} \mapsto f(t) = (\cos t, \sin t, t)$ est un plongement. Enfin énonçons un lemme utile concernant la caractérisation des difféomorphismes par le rang :

Lemme A.2.1. *Une application $f : M \to N$ différentiable est un difféomorphisme si et seulement si f est bijective et de rang $n = \dim M = \dim N$.*

Démonstration.
On note $f^{\phi\psi} = \psi \circ f \circ \phi^{-1}$. Supposons que f est un difféomorphisme alors f est bijective par définition. De plus pour tout $p \in M$, $D(f^{\phi\psi})(\phi(p)) : \mathbb{R}^m \to \mathbb{R}^n$ est un isomorphisme. Donc $n = m = rg(f^{\phi\psi})$.

Maintenant supposons que f est bijective, différentiable et vérifiant la condition de rang. Alors $f^{\phi\psi}$ est bijective et différentiable (au moins C^1). De plus par la condition de rang, pour tout p (on choisit les cartes pour chaque p), $D(f^{\phi\psi})(\phi(p))$ est un isomorphisme. Par le théorème d'inversion globale, $f^{\phi\psi}$ est un difféomorphisme. Donc f^{-1} est différentiable car $\phi \circ f^{-1} \circ \psi^{-1} = (f^{\phi\psi})^{-1}$. □

Parlons maintenant de **sous variété** d'une variété M. Une partie N d'une variété M de dimension n est une **sous variété** de dimension $k \leq n$ si pour tout $q \in N$, il existe une carte (U, ϕ) de M contenant q telle que

$$\phi(U \cap N) = \phi(U) \cap \widetilde{\mathbb{R}^k}$$

où $\widetilde{\mathbb{R}^k} = \mathbb{R}^k \cap \{0\}^{n-k}$ (notation). Remarquons qu'une sous variété est une variété de dimension k en posant l'atlas $\{(U \cap N, \pi \circ \phi_{U \cap N})\}$ où $\pi : \mathbb{R}^k \times \mathbb{R}^{n-k} \to \mathbb{R}^k$ est la projection. Énonçons deux lemmes utiles pour générer des sous variétés :

Lemme A.2.2. *Soient M, N deux variétés de dimension respectives m, n. Soit $F : N \to M$ un plongement. Alors $W = F(N)$ est une sous variété de M de dimension n.*

Lemme A.2.3. *Soient M, N deux variétés de dimension respectives m, n. Soit $F : M \to N$ une submersion et soit $y \in F(M)$. Alors $W = F^{-1}(y)$ est une sous variété de M de dimension $m - n$.*

On peut donner plusieurs applications de ces lemmes : sphère, groupe orthogonale, groupe spécial linéaire etc. Dans le cas des sous variétés de $\mathbb{R}^n$, la sous variété N est plongée. En effet soit $i : N \to \mathbb{R}^n$ définie par $i(x) = x$, $x \in \mathbb{N}$. Montrons que i est un plongement. Tout d'abord elle est bijective sur son image. Et sa réciproque est aussi l'identité. Montrons que i est différentiable. Notons $\phi = \phi_{U \cap N}$ on a $i \circ \phi^{-1} = \phi^{-1}$. En utilisant que les cartes sont différentiables, on obtient que i est différentiable et comme ϕ est un difféomorphisme, i est une immersion injective. L'homéo de i sur son image est évident ($\psi \circ i \circ \phi^{-1} = \psi \circ \phi^{-1}$). D'où une sous variété de $\mathbb{R}^n$ est plongé. Plus généralement toute sous variété N de M est plongée dans M par les même arguments ($\psi \circ \phi^{-1}$ est un difféomorphisme). On peut alors se poser la question suivante : Est ce que toute variété peut être plongée dans $\mathbb{R}^N$ pour un certain N? Soit $f : M \to \mathbb{R}^N$ un plongement. Alors $f(M)$ est une sous variété de $\mathbb{R}^N$. Ainsi M et $f(M)$ sont difféomorphes par le lemme A.2.2. La réciproque est vérifiée par le théorème suivant :

Théorème A.2.4. *(Plongement de Whitney) Toute variété de dimension n admet un plongement sur une sous variété fermée de $\mathbb{R}^{2n+1}$.*

A.3 Espaces tangents : courbes et dérivations

Définissons maintenant la notion de **vecteur tangent**. Soit $p \in M$. Un vecteur tangent à M au point p est une classe d'équivalence $[c]_p$ de courbes lisses tracées sur M et passant par le point p pour la relation d'équivalence suivante : Deux courbes lisses $c_i :]-\alpha_i, +\alpha_i[\to N$ avec $i = 1, 2$ telles que $c_i(0) = p$ sont dites équivalentes si pour une carte locale (U, ϕ) on a

$$\frac{d}{dt}(\phi \circ c_1)(0) = \frac{d}{dt}(\phi \circ c_2)(0) \in \mathbb{R}^m.$$

En utilisant le théorème de différentiation des fonctions composées et en écrivant

$$\psi \circ c = \psi \circ \phi^{-1} \circ \phi \circ c,$$

on s'aperçoit que cette relation d'équivalence est indépendante du choix de cartes. On note T_pM **l'ensemble des vecteurs tangents à M au point p**. Dans $\mathbb{R}^n$, deux courbes sont tangentes en un

point x quelconque ($c(0) = x$) dès que $c_1'(0) = c_2'(0)$. Il y a donc un isomorphisme **canonique** entre $T_p\mathbb{R}^n$ et $\mathbb{R}^n$ et l'ensemble des direction $c'(0)$. Cet isomorphisme ne dépend pas de x ce qui est propre à $\mathbb{R}^n$ car la structure différentielle est $(\mathbb{R}^n, id)$. Concernant la structure d'espace vectoriel de l'espace tangent, on peut l'obtenir en montrant que l'application $\gamma : [c]_p \to \frac{d}{dt}(\phi \circ c_1)(0)$ est une bijection entre T_pM et $\mathbb{R}^m \simeq T_{\phi(p)}\mathbb{R}^m$. Cette bijection permet de définir une structure d'espace vectoriel de la manière suivante

$$\begin{aligned}[c]_p + [d]_p &= \gamma^{-1}(v) + \gamma^{-1}(w) = \gamma^{-1}(v + w)\\ \lambda[c]_p &= \gamma^{-1}(\lambda v)\end{aligned}$$

qui transforme γ en un isomorphisme. Sauf que l'isomorphisme quand $M = \mathbb{R}^n$ est canonique (il ne dépend pas de x, du choix de carte, car il y a une seule carte et une seule structure différentielle sur $\mathbb{R}^n$). De ce fait, on peut identifier $\mathbb{R}^n$ à $T_p\mathbb{R}^n$ de manière canonique.

Cependant le maniement de ces classes d'équivalence n'est pas aisé. Donnons une autre définition équivalente de l'espace tangent par la notion de *dérivation*. Notons $C^\infty(p)$ l'ensemble des fonctions à valeurs réelles de classe C^∞ sur un ouvert de M contenant p dans lequel on identifie les fonctions qui sont égales sur un voisinage de p (on obtient la notion de *germe*). Cet ensemble est une algèbre c'est à dire un espace vectoriel muni d'une opération interne qui est la multiplication.

Une **dérivation en** p est une application linéaire $D_p : C^\infty(p) \to \mathbb{R}$ qui vérifie la règle de Leibniz c'est à dire

$$\begin{aligned}D_p(\alpha f + \beta g) &= \alpha D_p(f) + \beta D_p(g)\\ D_p(fg) &= g(p)D_p(f) + f(p)D_p(g)\end{aligned}$$

pour tout $f, g \in C^\infty(p)$. On remarque que $D_p(cste) = 0$ car $D_p(1) = D_p(1 \times 1) = 0$. On note $\mathscr{D}(p)$ l'ensemble des dérivations en p. On remarque que $\mathscr{D}(p)$ est clairement un espace vectoriel. Nous allons montrer que T_pM s'identifie à $\mathscr{D}(p)$. Pour cela, énonçons un lemme utile :

Lemme A.3.1. *(Hadamard) Soit (U, ϕ) une carte locale centrée en p (c'est à dire $\phi(p) = 0$). Pour toute fonction $g \in C^\infty(p)$, il existe $\chi_1, \cdots, \chi_n \in C^\infty(p)$ tels que*

$$g = g(p) + \sum_{i=1}^{n} x^i \chi_i$$

où $\phi = (x^1, \cdots, x^n)$.

En utilisant ce lemme, on peut caractériser les éléments de $\mathscr{D}(p)$. Fixons une carte (U, ϕ) centrée en p. Une dérivation s'écrit alors

$$\begin{aligned}D_p(g) &= D_p(g(p)) + \sum_{i=1}^{n}(\chi_i(p)D_p(x^i) + x^i(p)D_p(\chi_i))\\ &= \sum_{i=1}^{n} \chi_i(p)D_p(x^i).\end{aligned}$$

D'où D_p est entièrement déterminée par les réels $D_p(x^i)$ pour $i = 1, \cdots, n$. Ainsi nous avons

$$\dim \mathscr{D}(p) = n = \dim M.$$

Remarquons que dans $\mathbb{R}^n$, toute dérivation est une dérivation directionnelle. En effet posons

$$v \mapsto \partial_v$$

où $v \in \mathbb{R}^n$ et $\partial_v g = Dg(x)(v)$. L'ensemble des dérivées directionnelles en x, notée $\mathscr{D}'(x)$ est un sous espace vectoriel de $\mathscr{D}(x)$. Montrons cette application est un isomorphisme entre $\mathbb{R}^n$ et $\mathscr{D}'(x)$. Prouvons la partie injective. Soient $v \neq w$. Alors il existe u tel que $\langle u, v\rangle \neq \langle u, w\rangle$ (sinon on pose $u = v - w$ et on obtient une contradiction). On pose $f(x) = \langle u, x\rangle$. Alors $\partial_v f = \langle u, v\rangle \neq \langle u, w\rangle = \partial_w f$. D'où l'injectivité, la partie surjective est triviale par corestriction de l'espace d'arrivée. De plus, remarquons que les dérivées partielles en x forment une base de $\mathscr{D}'(x)$. Ainsi $\mathscr{D}'(x)$ est un sous espace vectoriel de dimension n. On en déduit alors que $\mathscr{D}' = \mathscr{D}$. On obtient alors un isomorphisme entre $\mathbb{R}^n$ et $\mathscr{D}(x)$. Enfin notons que les dérivées partielles en x forment une base de $\mathscr{D}(x)$.

Faisons maintenant le lien entre les vecteurs tangents et les dérivations :

Proposition A.3.2. *Soit $g \in C^\infty(p)$ et X_p un vecteur tangent en p. Alors la dérivée $\frac{d}{dt}(g \circ c)(0)$ est la même pour tout $c \in X_p$ (classe d'équivalence).*

Démonstration.
Choisissons des coordonnées locales ϕ et posons $c^\phi = \phi c$ et $g^\phi = g\phi^{-1}$. Alors $gc = g^\phi c^\phi$ et

$$\frac{d}{dt}(g \circ c)(0) = D(g\phi^{-1})(g(p))(\frac{d}{dt}(\phi \circ c)(0)).$$

Puisque $\frac{d}{dt}(g \circ c)(0)$ ne dépend que de classe d'équivalence, on obtient le résultat voulu. □

On note $X_p \cdot g = X_p(g) = \frac{d}{dt}(g \circ c)(0)$ On s'aperçoit que $g \mapsto X_p(g)$ est une dérivation. Pour le voir il suffit d'écrire

$$X_p(g) = \frac{d}{dt}(g^\phi \circ c^\phi)(0).$$

La dérivation $X_p(g)$ est souvent appelée dérivée directionnelle de g dans la direction X_p. Par cette dérivation, on débouche au théorème suivant :

Théorème A.3.3. *L'ensemble des vecteurs tangents T_pM s'identifie à l'espace vectoriel $\mathscr{D}(p)$ de dimension n des dérivations en p.*

Par ce théorème on peut aussi munir T_pM d'une structure vectorielle. De plus un vecteur tangent peut être vu soit comme une classe d'équivalence de courbes, soit comme une dérivation.

A.4 Différentielle, théorème d'inversion locale et théorème des fonctions implicites

On peut maintenant définir la notion de **différentielle**. Soient $F : M \to N$ une application différentiable et $g : N \to \mathbb{R}$. On définit *l'image réciproque de g par F* comme suit :

$$\begin{aligned} F^* : C^\infty(F(p)) &\to C^\infty(p) \\ g &\mapsto g \circ F \end{aligned}$$

La **différentielle** de F en p est définie par $DF_p : T_pM \to T_{F(p)}N$

$$DF_p(X_p) \cdot g = X_p(F^* g) = X_p(g \circ F).$$

On remarque que DF_p est linéaire grâce à la structure d'espace vectoriel de l'ensemble des dérivations en p. On notera la différentielle soit D, soit d. On appelle aussi DF_p *l'application linéaire tangente en p*. Soient $\phi : M_1 \to M_2$ et $\psi : M_2 \to M_3$, calculons la différentielle de $\psi \circ \phi$ en p. On a

$$\begin{aligned} (D\psi_{\phi(p)} \circ D\phi_p)(X_p)(f) &= D\psi_{\phi(p)}(D\phi_p(X_p))(f) \\ &= D\phi_p(X_p)(f \circ \psi) \\ &= X_p(f\psi\phi) \\ &= D(\psi \circ \phi)_p(X_p)(f). \end{aligned}$$

Ainsi si ϕ est un difféomorphisme et $\psi = \phi^{-1}$. Alors $D\phi_p : T_pM \to T_{\phi(p)}N$ est un isomorphisme et $D\psi_{\phi(p)} = (D\phi_p)^{-1}$.

Soit M une variété différentielle de dimension m. On note $(e_k)_{k\in[\![1,m]\!]}$ la base canonique de $\mathbb{R}^m$. Choisissons un système de coordonnées locales en $p \in M : (U, x)$. Alors la carte $x : U \to x(U)$ est un difféomorphisme, par ce qui précède x^{-1} est un difféomorphisme et $d(x^{-1})_{x(p)} : T_{x(p)}\mathbb{R}^m \to T_pM$ est un isomorphisme. Puisque $\{\partial_{e_k},\ k \in [\![1, m]\!]\}$ est une base de $T_{x(p)}\mathbb{R}^m$, alors $(d(x^{-1})_{x(p)}(\partial_{e_k}))_k$ est une base de T_pM. On a

$$\begin{aligned} d(x^{-1})_{x(p)}(\partial_{e_k})(f) &= \partial_{e_k}(f \circ x^{-1})(x(p)) \\ &:= \frac{\partial f}{\partial x_k}(p) := \left(\frac{\partial}{\partial x_k}\right)_p (f). \end{aligned}$$

L'ensemble $\left\{\left(\frac{\partial}{\partial x_k}\right)_p, k \in [\![1, m]\!]\right\}$ est une base de T_pM. On remarque que

$$\frac{\partial x_i}{\partial x_k}(p) = \delta_{i,k}$$

car

$$\begin{aligned} \frac{\partial x_i}{\partial x_k} &= \partial_{e_k}(\pi_i \circ x \circ x^{-1}) \\ &= \partial_{e_k}(\pi_i) \end{aligned}$$

où $\pi_i : (a_1, \cdots, a_m) \mapsto a_i$. Ainsi pour tout $a \in \mathbb{R}^m$, $\partial_{e_k}(\pi_i)(a) = \delta_{i,k}$. Le calcul s'opère de la manière suivante pour $\phi : M \to N$, $g \in C^\infty(\phi(p))$ et $X_p \in T_pM$,

$$\begin{aligned} D\phi(p)(X_p) \cdot g &= D\phi(p)(\sum_{i=1}^m v_i(p)(\frac{\partial}{\partial x_i})_p) \cdot g \\ &= \sum_{i=1}^m v_i(p) D\phi(p)(\frac{\partial}{\partial x_i})_p) \cdot g \\ &= \sum_{i=1}^m v_i(p)(\frac{\partial}{\partial x_i})_p(\phi^* \circ g) \\ &= \sum_{i=1}^m v_i(p) \frac{\partial(g \circ \phi)}{\partial x_i}(p). \end{aligned}$$

De plus, on a une généralisation des théorèmes classiques en calcul différentiel :

Théorème A.4.1. *(Théorème d'inversion locale) Soit $\phi : M \to N$ une application différentiable avec $\dim M = \dim N$. Si p est un point de M tel que $d\phi_p$ est un isomorphisme. Alors ϕ est un difféomorphisme local (existence d'un ouvert autour de p et un autre autour de $\phi(p)$ tel que ϕ restreint est un difféomorphisme).*

Théorème A.4.2. *(Théorème des fonctions implicites) Soit $\phi : M \to N$ une application différentiable tq $m \geq n$. Si $q \in \phi(M)$ est une valeur régulière (i.e pour tout $p \in \phi^{-1}(q)$, $d\phi_p$ est surjective (de rang max)), alors $\phi^{-1}(q)$ est une sous variété de M de dimension $m - n$. De plus l'espace tangent $T_p\phi^{-1}(q)$ de $\phi^{-1}(q)$ en p est le noyau de $d\phi_p$ i.e :*

$$T_p\phi^{-1}(q) = \{X \in T_pM, d\phi_p(X) = 0\}$$

Par exemple, si ϕ est une submersion (pour tout p, $d\phi_p$ est surjective), le théorème est vérifié pour tout $q \in \phi(M)$.

A.5 Fibré vectoriel et fibré tangent

Introduisons maintenant la notion de **fibré vectoriel**. Soient E et M deux variétés topologiques et $\pi : E \to M$ une application **continue et surjective**. Le triplet (E, M, π) est un **fibré vectoriel au dessus de** M de dimension n si :

1. Pour tout $p \in M$, la **fibre** $E_p = \pi^{-1}(\{p\})$ est un espace vectoriel de dimension n.
2. Pour tout $p \in M$, il existe une **carte locale fibrée** $(\pi^{-1}(U), \psi)$ où U est un voisinage ouvert de p et $\psi : \pi^{-1}(U) \to U \times \mathbb{R}^n$ un homéomorphisme tel que pour tout $q \in U$, $\psi_q = \psi_{E_q} : E_q \to \{q\} \times \mathbb{R}^n$ est un isomorphisme d'espace vectoriel.

On remarque que $\{q\} \times \mathbb{R}^n$ hérite de la structure d'espace vectoriel grâce à la bijection canonique entre $\{q\} \times \mathbb{R}^n$ et $\mathbb{R}^n$. Un **atlas fibré** est une collection

$$\mathscr{B} = \left\{(\pi^{-1}(U_\alpha), \psi_\alpha), \alpha \in I\right\}$$

de cartes locales fibrées telles que $M = \bigcup_\alpha U_\alpha$ et pour tout $\alpha, \beta \in I$, il existe une application $A_{\alpha,\beta}$: $U_\alpha \cap U_\beta \to GL_n(\mathbb{R})$ telle que :

$$\begin{aligned} \psi_\beta \circ \psi_\alpha^{-1} : (U_\alpha \cap U_\beta) \times \mathbb{R}^n &\to (U_\alpha \cap U_\beta) \times \mathbb{R}^n \\ (p, v) &\mapsto (p, A_{\alpha,\beta}(p)v). \end{aligned}$$

Les éléments de $\left\{A_{\alpha,\beta}, \alpha, \beta \in I\right\}$ sont appelés **des applications de transitions** de l'atlas fibré $\mathscr{B}$.

Une **section** du fibré vectoriel (E, M, π) est une application continue $v : M \to E$ telle que pour tout $p \in M, \pi \circ v(p) = p$. Enfin un fibré vectoriel topologique (E, M, π) est dit **trivial** s'il existe une carte globale fibré $\psi : E \to M \times \mathbb{R}^n$.

Exemple A.5.1. Donnons quelques exemples. Soit $M = S^1$ le cercle unité de $\mathbb{R}^2$ et soit $E = S^1 \times \mathbb{R}$ le cylindre. On note $\pi : (z, t) \mapsto z$ la projection canonique sur le cercle. π est clairement continue et surjective. Alors on peut voir facilement que (E, M, π) est une droite fibrée triviale puisque l'application identité $\psi : D^1 \times \mathbb{R} \to S^1 \times \mathbb{R}$ est une carte fibrée globale. Plus généralement $(M \times \mathbb{R}^n, M, \pi)$ est un fibré vectoriel trivial (M est une variété de dimension n). Mais il existe des fibrés non triviaux. Par exemple le cercle plongé dans $\mathbb{R}^4$ et le ruban de Möbius dans $\mathbb{R}^4$.

Dorénavant, on travaillera qu'avec des **fibrés vectoriels lisses** c'est à dire qu'on suppose E et M sont des variétés différentielles, π est une application différentiable et qu'il existe un atlas fibré **maximal** différentiable (autrement dit que les applications de transitions sont différentiables (lisses) et maximal (notion définie au début)). Ici on entend par application **différentiable** une application C^∞. Notons $C^\infty(E)$ l'ensemble de toutes les **sections lisses** du fibré vectoriel (E, M, π). On peut définir une structure de module sur l'anneau $C^\infty(M)$ et d'espace vectoriel sur le corps $\mathbb{R}$ de la manière suivante :

1. $(v + w)_p = v_p + w_m$,
2. $(f \cdot v)_p = f(p) \cdot v_p$

pour tout $p \in M, v, w \in C^\infty(E), f \in C^\infty(M)$. Si U est un ouvert de M alors l'ensemble $\{v^1, ..., v^n\}$ de sections lisses est un **repère local** pour E si pour tout $p \in U$, l'ensemble $\{v_p^1, ..., v_p^n\}$ est une base de $E_p = \pi^{-1}(\{p\})$ (par la définition de section, $v_p^i \in E_p$).

On peut maintenant définir la notion de **fibré tangent**. Soit M une variété différentielle de dimension m. Définissons l'ensemble

$$TM = \left\{(p, v), p \in M, v \in T_pM\right\}$$

et soit $\pi : TM \to M$ **l'application projection** qui à $(p, v) \mapsto p$. La **fibre** est alors $\pi^{-1}(\{p\})$ est $\{p\} \times T_pM \simeq T_pM$ (où la bijection est *canonique*). De ce fait $\{p\} \times T_pM$ a la même structure d'espace vectoriel que T_pM. Le triplet (TM, M, π) est appelé le **fibré tangent**. On peut l'équiper d'une structure

différentielle de fibré. Pour cela, on peut définir, pour toute carte locale de M, $x : U \to \mathbb{R}^m$, une carte locale sur le fibré tangent

$$x^* : \pi^{-1}(U) \to \mathbb{R}^m \times \mathbb{R}^m$$
$$(p, \sum_{k=1}^{m} v_k(p)(\frac{\partial}{\partial x_k})_p) \mapsto (x(p), (v_1(p), ..., v_m(p))).$$

Puisque $\left(\frac{\partial}{\partial x_k}\right)_p$ pour tout k forme une base de T_pM, l'application est bien définie (bijective dans $x(U) \times \mathbb{R}^m$). Remarquons que

$$\begin{aligned}(v_1(p), \cdots, v_m(p)) &= \sum_{k=1}^{m} v_k(p)(\frac{\partial}{\partial x_k})_p(x) \\ &= (\sum_{k=1}^{m} v_k(p)(\frac{\partial x_i}{\partial x_k})(p))_{i\in[\![1,m]\!]}.\end{aligned}$$

D'où

$$x^*(p, \sum_{k=1}^{m} v_k(p)(\frac{\partial}{\partial x_k})_p) = (x(p), \sum_{k=1}^{m} v_k(p)(\frac{\partial}{\partial x_k})_p(x)).$$

On note $\mathscr{A}$ la structure différentielle de M. La collection

$$\{(x^*)^{-1}(W), (U, x) \in \mathscr{A}, W \subset x(U) \times \mathbb{R}^m \text{ouvert}\}$$

définit une base d'ouvert. La topologie engendrée par cette base est la plus petite topologie qui rend continue x^*. Ainsi $(\pi^{-1}(U), x^*)$ est une carte locale de TM. Donc TM est une variété topologique de dimension $m + m = 2m$. Pour obtenir la régularité de la variété, nous avons besoin de regarder de plus près les changements de cartes. Tout d'abord donnons la réciproque de x^*

$$(x^*)^{-1}(a, b) = (x^{-1}(a), \sum_{k=1}^{m} b_k(\frac{\partial}{\partial x_k})_{x^{-1}(a)}).$$

Ainsi

$$y^* \circ (x^*)^{-1}(a, b) = (y \circ x^{-1}(a), \sum_{k=1}^{m} v_k(p)\frac{\partial y}{\partial x_k}(x^{-1}(a)))$$

où $y = (y_1, \cdots, y_m)$. Si l'on suppose que $y \circ x^{-1}$ est différentiable alors $y^* \circ (x^*)^{-1}$ est aussi différentiable. Alors

$$\mathscr{A}^* := \left\{(\pi^{-1}(U), x^*), (U, x) \in \mathscr{A}\right\}$$

est une structure différentielle sur la variété TM (quitte à prendre l'atlas maximal). Enfin remarquons que la restriction de x^* à la fibre $E_p = \{p\} \times T_pM$ est clairement un isomorphisme pour la structure d'espace vectoriel associée à $\{p\} \times T_pM$. D'où le fibré tangent est un fibré vectoriel. Par ce qui précède, on remarque que les cartes locales sont des cartes locales fibrées et donc l'ensemble est un atlas fibrée. De plus les applications de transitions sont régulières car l'application $A : p \mapsto A(p) = \left(\frac{\partial y_i}{\partial x_j}(p)\right)_{i,j\in[\![1,m]\!]}$ est bien définie et est régulière (car $y \circ x^{-1}$ est un difféomorphisme). On peut conclure que le fibré vectoriel a une structure de fibré vectoriel lisse.

A.6 Champ de vecteurs et crochet de Lie

Introduisons maintenant la notion de **champ de vecteurs**. Soit M une variété différentielle. Alors une section $X : M \to TM$ du fibré tangent est appelée un **champ de vecteurs**. Pour un champ de vecteur X, on peut définir l'application $X(f) : p \mapsto X_p(f) = X_p \cdot f$ pour tout $f \in C^\infty(M)$. L'ensemble des champs de vecteurs lisses est noté $C^\infty(TM)$. De plus, chaque champ de vecteurs lisse s'écrit dans une carte (U, x)

$$X = \sum_{i=1}^{m} X^i \frac{\partial}{\partial x_i}$$

car pour tout $p \in M$, $X(p) = (p, X_p) \simeq X_p = \sum_{i=1}^{m} X^i(p)(\frac{\partial}{\partial x_i})_p \in T_pM$. Remarquons que $\left\{\frac{\partial}{\partial x_i}, i \in [\![1, m]\!]\right\}$ est un repère local sur la carte (U, x). Introduisons une opération importante sur les champs de vecteurs : **le crochet de Lie**. Pour deux champs de vecteurs $X, Y \in C^\infty(TM)$, le crochet de Lie en p est défini par

$$[X, Y]_p(f) = X_p(Y(f)) - Y_p(X(f))$$

et $[X, Y]_p : C^\infty(M) \to \mathbb{R}$ et $[X, Y] = X \circ Y - Y \circ X : p \mapsto [X, Y]_p$. En utilisant la définition de dérivation, on remarque que

$$\begin{aligned} [X, Y]_p(\lambda f + \mu g) &= \lambda [X, Y]_p(f) + \mu [X, Y]_p(g), \\ [X, Y]_p(fg) &= [X, Y]_p(f)g(p) + f(p)[X, Y]_p(g). \end{aligned}$$

D'où $[X, Y]_p$ est une dérivation. Ainsi $[X, Y] : p \mapsto (p, [X, Y]_p) \simeq [X, Y]_p$ est une section du fibré tangent. Il est possible de montrer que cette section est lisse. Pour montrer qu'une section est lisse sur une variété lisse, on utilise les équivalences suivantes :

1. La section X est lisse.
2. Si (U, x) est une carte locale alors les fonctions $a_1, ..., a_m$ données par :

$$X_{|U} = \sum_{i=1}^{m} a_i \frac{\partial}{\partial x_i}$$

   sont lisses.
3. Si $f : V \to \mathbb{R}$ où V est un ouvert de M et f est lisse alors la fonction $X(f) : V \to \mathbb{R}$ définie par $X(f)(p) = X_p(f)$ est lisse.

On peut faire agir $C^\infty(M)$ sur $C^\infty(TM)$ de la manière suivante

$$f \cdot Y : p \mapsto f(p) \cdot Y_p : g \mapsto f(p) Y_p(g)$$

Après calcul, on obtient des égalités suivantes

$$\begin{aligned} [X, f \cdot Y] &= X(f) \cdot Y + f \cdot [X, Y], \\ [f \cdot X, Y] &= f \cdot [X, Y] - Y(f) \cdot X \end{aligned}$$

pour tout $X, Y \in C^\infty(TM)$, $f \in C^\infty(M)$.

Un concept fondamental en géométrie différentielle est celui **d'algèbre de Lie**. Un espace vectoriel $(V, +, \cdot)$ muni d'une opération $[,] : V \times V \to V$ est une algèbre de lie si les rélations suivantes sont vérifiées :

1. $[\lambda X + \mu Y, Z] = \lambda[X, Z] + \mu[Y, Z]$.
2. $[X, Y] = -[Y, X]$.
3. $[X, [Y, Z]] + [Z, [X, Y]] + [Y, [Z, X]] = 0$.

La dernière égalité est appelée **l'identité de Jacobi**. Un exemple d'algèbre de lie est $\mathbb{R}^3$ muni de l'application bilinéaire *produit vectoriel* définie par

$$\begin{aligned} X \wedge Y &= Z, \\ Z \wedge X &= Y, \\ Y \wedge Z &= X. \end{aligned}$$

Un autre exemple important est le suivant. Soit M une variété lisse. L'espace vectoriel $C^\infty(TM)$ (sur $\mathbb{R}$) muni du crochet de Lie $[,] : C^\infty(TM) \times C^\infty(TM) \to C^\infty(TM)$ est une algèbre de Lie.

Intéressons nous maintenant à la notion de *différentielle d'une fonction en un champ de vecteur*. Soit $\phi : M \to N$ une application différentiable et surjective entre deux variétés différentielles. Alors deux champs de vecteurs $X \in C^\infty(TM)$ et $Y \in C^\infty(TN)$ sont dits ϕ-**reliés** si $d\phi_p(X_p) = Y_{\phi(p)}$ pour tout $p \in M$. On note

$$d\phi(X) = Y$$

Faisons une remarque intéressante permettant de montrer que $d\phi(X)$ est un champ de vecteur sur la variété d'arrivée. Soit $\phi : M \to N$ une application différentiable entre deux variétés différentielles. Alors pour tout $p \in M$, $d\phi_p(X_p)(f) = X_p(f \circ \phi)$ et

$$d\phi(X)(f)(\phi(p)) = X(f \circ \phi)(p).$$

En réecrivant

$$d\phi(X)(f) \circ \phi = X(f \circ \phi),$$

on peut montrer que si ϕ est surjective et différentiable alors

$$d\phi([X, Y]) = [d\phi(X), d\phi(Y)].$$

Si de plus, ϕ est un difféomorphisme alors $d\phi(X)$ est un champ de vecteur sur N (car ϕ bijective donc $d\phi(X)$ est une section $(q \in N \mapsto d\phi_{\phi^{-1}(q)}(X_\phi^{-1}(q))))$. De plus $d\phi : C^\infty(TM) \to C^\infty(TN)$ est un morphisme d'algèbre de lie car

$$d\phi([X, Y]) = [d\phi(X), d\phi(Y)].$$

On dit que $X, Y \in C^\infty(TM)$ **commutent** si $[X, Y] = 0$. On peut montrer qu'un repère local du fibré tangent commute :

Proposition A.6.1. *Soit M une varité différentielle. Soit (U, x) une carte locale sur M et notons*

$$\left\{ \frac{\partial}{\partial x_k}, k \in [\![1, m]\!] \right\}$$

le repère local induit par la carte pour le fibré tangent TM. Alors on a

$$\left[\frac{\partial}{\partial x_k}, \frac{\partial}{\partial x_l}\right] = 0$$

pour tout $k, l \in [\![1, m]\!]$.

Démonstration.
Fixons $k, l \in [\![1, m]\!]$. Par ce qui précède (au début) $x : U \to x(U)$ est un difféomorphisme. Puisque $dx_p((\frac{\partial}{\partial x_k})_p) = (\partial_{e_k})_{x(p)}$ pour tout $p \in U$, on a

$$dx(\frac{\partial}{\partial x_k}) = \partial_{e_k} \in C^\infty(Tx(U)).$$

D'où $\frac{\partial}{\partial x_k}$ et ∂_{e_k} sont x-reliés. Comme x est un difféomorphisme,

$$dx([\frac{\partial}{\partial x_k}, \frac{\partial}{\partial x_l}]) = [\partial_{e_k}, \partial_{e_l}].$$

Par le théorème de Schwarz,

$$[\partial_{e_k}, \partial_{e_l}](f) = \partial_{e_k}(\partial_{e_l})(f) - \partial_{e_l}(\partial_{e_k})(f) = 0$$

pour tout $f \in C^2(x(U))$. Puisque x est un difféomorphisme, dx_p est une application linéaire bijective pour tout $p \in U$ donc

$$[\frac{\partial}{\partial x_k}, \frac{\partial}{\partial x_l}]_p = (dx_p)^{-1}([\partial_{e_k}, \partial_{e_l}]_x(p)) = 0.$$

Ainsi

$$[\frac{\partial}{\partial x_k}, \frac{\partial}{\partial x_l}] = 0.$$

□

A.7 Variété riemannienne, longueurs et fibré normal

Introduisons maintenant la notion de **variété riemannienne**. Soit M une variété lisse. Notons $C^\infty(M)$ l'anneau des fonctions lisses sur M et $C^\infty(TM)$ l'ensemble des champs de vecteurs lisses qui forme un module sur l'anneau $C^\infty(M)$. On pose :

$$C_0^\infty(TM) = C^\infty(M)$$

et

$$C_r^\infty(TM) = C^\infty(TM) \otimes \ldots \otimes C^\infty(TM)$$

le produit tensoriel $r \in \mathbb{N}^*$ fois. Une propriété remarquable du produit tensoriel est que si E, F sont des espaces vectoriels de dimension finie alors $\dim(E \otimes F) = \dim(E)\dim(F)$. Une généralisation est possible avec la notion de *longueur de module*.

Soit M une variété différentielle. Un (champ de) tenseur lisse A de type (r,s) est une application

$$A : C_r^\infty(TM) \to C_s^\infty(TM)$$

qui est multilinéaire sur l'anneau $C^\infty(M)$. On note $A(X_1,...,X_r)$ à la place de $A(X_1 \otimes .... \otimes X_r)$. Un résultat fondamental est le suivant. Si l'on prend $X_1,...,X_r$ et $Y_1,...,Y_r$ des champs de vecteurs tels que

$$(X_k)_p = (Y_k)_p$$

alors on a

$$A(X_1,\cdots,X_r)(p) = A(Y_1,\cdots,Y_r)(p)$$

Ainsi cela dépend seulement de la valeur des champs de vecteur en p. Pour un tenseur A, on note A_p la restriction du tenseur aux espaces tangents

$$A_p : ((X_1)_p,...,(X_r)_p) \mapsto A(X_1,...,X_r)(p).$$

Une **métrique riemannienne** est un tenseur $g : C_2^\infty(TM) \to C_0^\infty(TM) = C^\infty(M)$ tel que pour tout $p \in M$, la restriction g_p au produit $T_pM \times T_pM$ (à valeurs dans $\mathbb{R}$) est un **produit scalaire réel** sur l'espace tangent T_pM. La paire (M,g) est appelée une **variété riemannienne**.

Exemple A.7.1. Donnons quelques exemples :

- Le produit scalaire euclidien : $E^n = (\mathbb{R}^n, \langle , \rangle)$.
- Sur $\mathbb{R}^n$, on peut poser la métrique suivante

$$g_p(X,Y) = \frac{4}{(1+|p|^2)^2}\langle X,Y\rangle.$$

On note $\Sigma^n = (\mathbb{R}^n, g)$.

La **longueur d'une courbe** en géométrie riemannienne est définie pour une courbe $\gamma : I \to M$ C^1 de la manière suivante

$$L(\gamma) = \int_I \sqrt{g_{\gamma(t)}(\dot\gamma(t),\dot\gamma(t))}\, dt.$$

Exemple A.7.2. Prenons la courbe $\gamma(t) = (t,0,...,0) \in \mathbb{R}^n$ pour $t \in \mathbb{R}_+$. Dans l'espace Σ^n,

$$\begin{aligned} L(\gamma) &= 2\int_0^\infty \frac{\sqrt{\langle\dot\gamma(t),\dot\gamma(t)\rangle}}{1+|\gamma|^2}\, dt \\ &= 2\int_0^\infty \frac{1}{1+t^2}\, dt \\ &= 2[\arctan(t)]_0^\infty = \pi. \end{aligned}$$

Dans l'espace E^n, $L(\gamma) = +\infty$.

On peut munir une variété riemannienne d'une structure d'espace métrique de manière naturelle. Soit (M, g) une variété riemannienne connexe par arcs. Notons C_{pq} l'ensemble des courbes $\gamma : [0,1] \to M$ C^1 telles que $\gamma(0) = p$ et $\gamma(1) = q$. Définissons la **distance**

$$d(p,q) = \inf\{L(\gamma), \gamma \in C_{pq}\}, p, q \in M.$$

La topologie induite sur M par la distance d est la même que la topologie de M comme variété topologique.

De plus, nous pouvons définir une structure de variété riemannienne pour une sous variété d'une variété riemannienne de la manière suivante. Soient (N, h) une variété riemannienne et M une sous variété de N. On définit le tenseur $g : C_2^\infty(TM) \to C^\infty(M)$ par

$$g(X,Y) : p \mapsto h_p(X,Y).$$

Mentionnons un résultat important :

Théorème A.7.1. *Toute variété différentielle peut être équipée d'une structure de variété riemannienne (autrement dit il existe une métrique riemannienne sur cette variété).*

Introduisons maintenant la notion de **fibré normal**. Soit (N, h) une variété riemannienne et M une sous variété. Pour tout point $p \in M$, on définit l'espace normal N_pM en p par

$$N_pM = \left\{X \in T_pN, h_p(X,Y) = 0 \ \ \forall Y \in T_pM\right\}.$$

Pour tout $p \in M$, on a la décomposition suivante $T_pN = T_pM \oplus N_pM$. On définit le **fibré normal** par

$$NM = \left\{(p,X), p \in M, X \in N_pM\right\}.$$

Le fibré normal est un fibré vectoriel au dessus de M de dimension $n - m$.

Exemple A.7.3. Soit S^m la sphère dans $\mathbb{R}^{m+1}$ équipée de la structure euclidienne $<,>$. Si $p \in S^m$, alors l'espace tangent en p est

$$T_pS^m = \left\{X \in \mathbb{R}^{m+1}, < p, X >= 0\right\}$$

et l'espace normal en p est la droite engendrée

$$N_pS^m = \left\{\lambda p, \lambda \in \mathbb{R}\right\}.$$

Le fibré normal est alors

$$NS^m = \left\{(p, \lambda p), p \in S^m, \lambda \in \mathbb{R}\right\}.$$

A.8 Connexions et symboles de Christoffel

Introduisons maintenant un objet central en géométrie différentielle : **la connexion**.

Definition A.8.1. Soit M une variété différentielle, et soit (E, M, π) un fibré vectoriel lisse au dessus de M. Alors une connexion ∇ sur (E, M, π) est un opérateur

$$\nabla : C^\infty(TM) \times C^\infty(E) \to C^\infty(E)$$

tel que pour tout $\lambda, \mu \in \mathbb{R}$, $f, g \in C^\infty(M)$, $X, Y \in C^\infty(TM)$ et pour toutes sections lisses $v, w \in C^\infty(E)$ nous avons :

1. $\nabla_X(\lambda \cdot v + \mu \cdot w) = \lambda \cdot \nabla_X v + \mu \cdot \nabla_X w$,
2. $\nabla_X(f \cdot v) = X(f) \cdot v + f \cdot \nabla_X v$,
3. $\nabla_{(f \cdot X + g \cdot Y)} v = f \cdot \nabla_X v + g \cdot \nabla_Y v$.

Une section $v \in C^\infty(E)$ du fibré vectoriel (E, M, π) est dite **parallèle** par rapport à la connexion ∇ si et seulement si pour tout champ de vecteurs $X \in C^\infty(TM)$

$$\nabla_X v = 0.$$

Pour $M = \mathbb{R}^n$ ou un ouvert, on a que $T_p\mathbb{R}^n \simeq \mathbb{R}^n$ de manière canonique, un exemple de connexion dans ce cas est

$$\nabla_X Y(x) = \lim_{t \to 0} \frac{Y(x + tX(x)) - Y(x)}{t}.$$

Soit M une variété lisse et soit ∇ une connexion sur le fibré tangent (TM, M, π). On peut définir la **torsion** de la connexion ∇

$$\begin{aligned} T : C^\infty(TM) \times C^\infty(TM) &\to C^\infty(TM) \\ (X, Y) &\mapsto \nabla_X Y - \nabla_Y X - [X, Y] \end{aligned}$$

où $[,]$ est le crochet de lie sur $C^\infty(TM)$.

Une connexion ∇ est dite **sans torsion** si $T = 0$, autrement dit pour tout $X, Y \in C^\infty(TM)$

$$\nabla_X Y - \nabla_Y X = [X, Y].$$

Lorsque M est une **variété riemannienne** de métrique g alors nous pouvons définir la notion de connexion **compatible avec la métrique** g ou **de connexion métrique**. Une connexion sur le fibré tangent est dite **métrique** si pour tout $X, Y, Z \in C^\infty(TM)$

$$X(g(Y, Z)) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z).$$

Faisons une petite remarque. Soit (U, x) une carte locale de la variété riemannienne M. Soit $p \in U$. Alors $\left\{ (\frac{\partial}{\partial x_i})_p \right\}_{i \in [\![1,m]\!]}$ est une base de T_pM. Sur T_pM, on a un produit scalaire réel g_p. Ainsi à partir

de cette base par Gram-Schmidt, on peut construire une base orthonormée pour le produit scalaire g_p. Notons $\{(E_1)_p, \cdots, (E_m)_p\}$ cette b.o.n. On a alors définit un repère local orthonormé $\{E_1, \cdots, E_m\}$ tel que

$$g(E_i, E_j)(p) = g_p((E_i)_p, (E_j)_p) = \delta_{i,j},\ i, j \in [\![1, m]\!].$$

Puisque $\nabla_X Y \in C^\infty(TM)$, on a

$$\nabla_X Y = \sum_{i=1}^{m} \alpha_i E_i.$$

Montrons que $\alpha_i = g(\nabla_X Y, E_i)$. Puisque g est un tenseur alors

$$g(\nabla_X Y, E_i) = \sum_{j=1}^{m} \alpha_j g(E_j, E_i).$$

En utilisant que $g(E_j, E_i) = \delta_{i,j}$, on a le résultat voulu. Ainsi

$$\nabla_X Y = \sum_{i=1}^{m} g(\nabla_X Y, E_i) E_i.$$

Definition A.8.2. Soit (M, g) une variété riemannienne. La **connexion de Levi-Civita** est l'opérateur

$$\nabla : C^\infty(TM) \times C^\infty(TM) \to C^\infty(TM)$$

défini pour tout $X, Y, Z \in C^\infty(TM)$ par

$$\begin{aligned} g(\nabla_X Y, Z) = \frac{1}{2}[&X(g(Y, Z)) + Y(g(X, Z)) - Z(g(X, Y)) \\ &+ g([Z, X], Y) + g([Z, Y], X) + g(Z, [X, Y])]. \end{aligned}$$

Si l'on prend une connexion métrique sans torsion alors elle vérifie cette relation (appelée relation de **Koszul**). De plus l'écriture dans un repère local orthonormé de la connexion $\nabla_X Y$ montre qu'il y a au plus une connexion métrique sans torsion sur le fibré tangent (TM, M, π). La connexion de Levi-Civita est un objet intrinsèque de la variété riemannienne. De plus c'est **l'unique** connexion métrique sans torsion sur la variété riemannienne (M, g).

Enfin mentionnons une autre notion importante : **la dérivée covariante du premier ordre**. La dérivée covariante (du premier ordre) dans la direction de X

$$\nabla_X : C^\infty(TM) \to C^\infty(TM)$$

est donnée par

$$\nabla_X : Y \mapsto \nabla_X Y.$$

Introduisons maintenant les **symboles de Christoffel**. Soit (M, g) une variété riemannienne de dimension $m \in \mathbb{N}^*$ équipée de la connexion de Levi-Civita ∇. Prenons une carte locale (U, x) de M,

sur celle ci on peut définir un repère local de $TM : \{X_1, \cdots, X_m\}$ où $X_i = \frac{\partial}{\partial x_i} \in C^\infty(TU)$. On définit les **symboles de Christoffel** $\Gamma^k_{i,j} : U \to \mathbb{R}$ de la connexion ∇ par rapport à (U, x) de la manière suivante,

$$\nabla_{X_i} X_j = \sum_{k=1}^m \Gamma^k_{i,j} X_k.$$

Sur l'ouvert $x(U)$ de $\mathbb{R}^m$, on définit la métrique riemannienne $\tilde{g}(e_i, e_j) = g_{i,j} = g(X_i, X_j)$ où $(e_i)_{i\in[\![1,m]\!]}$ est la base canonique. Notons que puisque x est un difféomorphisme alors

$$dx : C^\infty(TM) \to C^\infty(TM)$$

est bijective. Remarquons aussi que pour tout $i \in [\![1, m]\!]$, $p \in U$ on a $dx(X_i) = \partial_i$ car $(X_i)_p(g \circ x) = \partial_i(g \circ x \circ x^{-1})(x(p)) = \partial_i(g)(x(p))$ et $dx(X_i)(p) = (X_i)_p$. Par la propriété portant sur les crochets de Lie, on a

$$dx([X_i, X_j]) = [dx(X_i), dx(X_j)] = [\partial_i, \partial_j] = 0$$

avec $i, j \in [\![1, m]\!]$. En utilisant la bijectivité de dx, on en déduit que $[X_i, X_j] = 0$.

Par la définition des symboles de Christoffel et la relation de Koszul et par le fait $[X_i, X_j] = 0$ pour tout $i, j \in [\![1, m]\!]$, on a que pour tout $l \in [\![1, m]\!]$,

$$\begin{aligned}
\sum_{k=1}^m g_{k,l}\Gamma^k_{i,j} &= \sum_{k=1}^m g(X_k, X_l)\Gamma^k_{i,j} \\
&= g(\sum_{k=1}^m \Gamma^k_{i,j} X_k, X_l) \\
&= g(\nabla_{X_i} X_j, X_l) \\
&= \frac{1}{2}\left\{X_i(g(X_j, X_l)) + X_j(g(X_l, X_i)) - X_l(g(X_i, X_j))\right\} \\
&= \frac{1}{2}\left\{\frac{\partial g_{j,l}}{\partial x_i} + \frac{\partial g_{l,i}}{\partial x_j} - \frac{\partial g_{i,j}}{\partial x_l}\right\}.
\end{aligned}$$

Et cela **pour tout** $l \in [\![1, m]\!]$. Ainsi pour chaque paire $(i, j) \in [\![1, m]\!]^2$ on a un système linéaire de m équations pour m variables $\Gamma^k_{i,j}$. La métrique g est entièrement déterminée par le repère local donc $g = (g_{i,j})_{i,j\in[\![1,m]\!]}$. Puisque g est définie positive, on peut trouver une solution à ce système. Notons $g^{k,l} = (g^{-1})_{k,l}$ pour tout $k, l \in [\![1, m]\!]$. On obtient alors

$$\Gamma^k_{i,j} = \frac{1}{2}\sum_{l=1}^m g^{k,l}\left(\frac{\partial g_{j,l}}{\partial x_i} + \frac{\partial g_{l,i}}{\partial x_j} - \frac{\partial g_{i,j}}{\partial x_l}\right). \tag{A.1}$$

Pour une sous variété M d'une variété riemanienne N, il existe des notions d'**extension locale ou globale** d'un champ de vecteur. On peut décomposer $Z \in C^\infty(TN)$ comme $Z = Z^\perp + Z^\top$ où $Z^\perp$ est une section du fibré normal de M et $Z^\top$ est une section du fibré tangent de M.

A.9 Géodésiques et champ parallèle le long d'une courbe

Parlons maintenant de **géodésique**. Soit (TM, M, π) le fibré tangent d'une variété lisse M. **Un champ de vecteurs le long d'une courbe** $\gamma : I \to M$ est une application $X : I \to TM$ telle que $\pi \circ X = \gamma$. On note $C^\infty_\gamma(TM)$ l'ensemble des champs de vecteurs le long de la courbe γ lisses. On définit les opérations suivantes :

1. $(X+Y)(t) = X(t) + Y(t)$,
2. $(fX)(t) = f(t)X(t)$

pour $t \in I$, $f \in C^\infty(I)$ et $X, Y \in C^\infty_\gamma(TM)$. Ainsi C^∞_γ a une structure de module sur $C^\infty(I)$ et une structure d'espace vectoriel sur $\mathbb{R}$.

Pour une courbe γ de classe C^1, on appelle **champ tangent le long de** γ le champ de vecteur $X(t) = (\gamma(t), \dot\gamma(t))$ pour tout $t \in I$. De plus, on peut définir un opérateur différentiel le long d'une courbe qui est relié à la connexion :

Proposition A.9.1. *Soit (M, g) une variété riemannienne lisse et soit $\gamma : I \to M$ une courbe dans M. Alors il existe **un unique** opérateur différentiel*

$$\frac{D}{dt} : C^\infty_\gamma(TM) \to C^\infty_\gamma(TM)$$

tel que pour tout $\lambda, \mu \in \mathbb{R}$ et $f \in C^\infty(M)$

1. $D(\lambda X + \mu Y)/dt = \lambda \frac{DX}{dt} + \mu \frac{DY}{dt}$,
2. $D(fX)/dt = \frac{df}{dt} X + f \frac{DX}{dt}$,
3. *Pour tout $t_0 \in I$, il existe un intervalle $J \subset I$ tel que $t_0 \in J$ et si $X \in C^\infty(TM)$ tel que $X_{\gamma(t)} = Y(t)$ pour tout $t \in J$ alors*
$$\frac{DY}{dt}(t_0) = (\nabla_{\dot\gamma} X)_{\gamma(t_0)}.$$

Démonstration.
Prouvons l'unicité. Soit $t_0 \in I$, on choisit une carte locale (U, x) et un sous intervalle J tel que $\gamma(J) \subset U$. Pour tout $i \in [\![1, m]\!]$, on note $X_i = \frac{\partial}{\partial x_i}$. Soit Y un champ de vecteur le long de la restriction de γ à J. On peut écrire Y de la manière suivante

$$Y(t) = \sum_{j=1}^m \alpha_j(t)(X_j)_{\gamma(t)}$$

où $\alpha_j \in C^\infty(J)$ et $j \in [\![1, m]\!]$. Les conditions 1) et 2) impliquent

$$DY/dt(t) = \sum_{k=1}^m \dot\alpha_k(t)(X_k)_{\gamma(t)} + \sum_{j=1}^m \alpha_j(t)(\frac{DX_j}{dt})_{\gamma(t)}$$

où $Y \in C^\infty_\gamma(TM)$. Pour une carte (U,x), on note

$$x \circ \gamma(t) = (\gamma_1(t), ..., \gamma_m(t)) = \sum_{i=1}^m \gamma_i(t) e_i.$$

Puisque x est un difféomorphisme, $dx : C^\infty(TU) \to C^\infty(T\mathbb{R}^m)$ est un isomorphisme (en particulier linéaire) tel que $i \in [\![1,m]\!]$

$$dx(X_i) = e_i$$

Puisque $\frac{d}{dt}(x \circ \gamma)(t) = dx_{\gamma(t)}(\dot\gamma(t)) := dx(\dot\gamma)(t)$, on a

$$\begin{aligned}\dot\gamma &= dx^{-1}(dx(\dot\gamma)) \\ &= dx^{-1}(\frac{d}{dt}(x \circ \gamma)) \\ &= dx^{-1}(\sum_{i=1}^m \dot\gamma_i e_i) \\ &= \sum_{i=1}^m \dot\gamma_i \ (X_i)_\gamma.\end{aligned}$$

Donc $\dot\gamma(t) = \sum_{i=1}^m \dot\gamma_i(t) \ (X_i)_{\gamma(t)}$. Par 3), on a

$$\left(\frac{DX_j}{dt}\right)_{\gamma(t)} = (\nabla_{\dot\gamma} X_j)_{\gamma(t)} = \sum_{i=1}^m \dot\gamma_i(t) \ (\nabla_{X_i} X_j)_{\gamma(t)}.$$

En utilisant l'expression de $(\nabla_{X_i} X_j)_{\gamma(t)}$ avec les symboles de christoffel et l'expression de $\frac{dY}{dt}$, puis en simplifiant (réarrangement des sommes), on obtient

$$\left(\frac{dY}{dt}\right)(t) = \sum_{k=1}^m \left(\dot\alpha_k(t) + \sum_{i,j=1}^m \alpha_j(t)\dot\gamma_i(t)\Gamma^k_{i,j}(\gamma(t))\right) (X_k)_{\gamma(t)}. \tag{A.2}$$

Ainsi l'opérateur différentiel $\frac{d}{dt}$ est entièrement déterminé de manière unique par γ et les symboles de christoffel $(\Gamma^k_{i,j})_{i,j,k \in [\![1,m]\!]}$. □

La formule (A.2) est très importante pour le calcul de $\nabla_{\dot\gamma} X$. Donnons quelques définitions :

Definition A.9.1. Soit (M,g) une variété riemannienne et soit $\gamma : I \to M$ une courbe C^1.

1. Un champ de vecteur X le long de γ est dit **parallèle** si

$$\nabla_{\dot\gamma} X = 0.$$

   En géométrie de l'information, X est dit parallèle à M si il est parallèle à toute courbe C^1 de M.

2. On suppose que γ est une courbe de classe C^2. Alors γ est une **géodésique** si son champ de tangent est parallèle le long de γ autrement dit

$$\nabla_{\dot\gamma} \dot\gamma = 0.$$

Le résultat suivant nous montre qu'on peut trouver un unique champ de vecteur le long de γ parallèle passant par un point donné :

Théorème A.9.2. *Soit (M,g) une variété riemannienne et soit $I =]a,b[$ un intervalle ouvert de $\mathbb{R}$. De plus, soient $\gamma : I \to M$ une courbe de classe C^1, $t_0 \in I$ et $v \in T_{\gamma(t_0)}M$. Alors il existe un **unique** champ de vecteur le long de γ **parallèle** tel que $Y(t_0) = v$.*

La preuve est assez similaire à la preuve de la proposition précédente. Par ce théorème, on peut définir la notion de **transport parallèle**. On note $\Pi_\gamma(v) = Y(1) \in T_{\gamma(1)}M$ si $\gamma : [0,1] \to M$ et $v \in T_{\gamma(0)}(M)$. C'est une application qui relie les espaces tangents.

Pour les champs de vecteurs parallèles, un autre résultat important est le suivant :

Proposition A.9.3. *Soient (M,g) une variété riemannienne, $\gamma : I \to M$ une courbe C^1 et X, Y deux champs de vecteurs parallèles le long de γ. Alors la fonction $g(X,Y) : I \to \mathbb{R}$ donnée par*

$$t \mapsto g_{\gamma(t)}(X_{\gamma(t)}, Y_{\gamma(t)})$$

*est **constante**. En particulier, si γ est une géodésique alors $g(\dot\gamma,\dot\gamma)$ est constante le long de γ.*

Démonstration.
En utilisant le fait que la connexion de Levi-Civita est métrique, on a

$$\begin{aligned}\frac{d}{dt}(g(X,Y)) &= g(\nabla_{\dot\gamma}X, Y) + g(X, \nabla_{\dot\gamma}Y)\\ &= 0.\end{aligned}$$

D'où $g(X,Y)$ est constante le long de γ. □

Un corollaire très utile de cette proposition :

Corollaire A.9.4. *Soient (M,g) une variété riemannienne, $p \in M$ et $\{v_1, \cdots, v_m\}$ une base orthonormée de T_pM. Et soient $\gamma : I \to M$ une courbe C^1 telle que $\gamma(0) = p$ et $X_1, \cdots, X_m$ des champs de vecteurs parallèles le long de γ tels que $X_k(0) = v_k$ pour $k \in [\![1,m]\!]$. Alors, pour tout $t \in I$, $\{X_1(t), \cdots, X_m(t)\}$ est une base orthonormée de $T_{\gamma(t)}M$.*

Démonstration.
Cela est une conséquence directe de la proposition précédente. En effet, soient $i, j \in [\![1,m]\!]$, $g(X_i, X_j)(t) = g(X_i, X_j)(0) = \delta_{i,j}$. □

De plus, il existe un résultat d'existence et d'unicité des géodésiques dans une variété riemannienne :

Théorème A.9.5. *Soit (M,g) une variété riemannienne. Si $p \in M$ et $v \in T_pM$ alors il existe un intervalle ouvert $]-\varepsilon,\varepsilon[$ avec $\varepsilon \in]0,1[$ et une* ***unique*** *géodésique $\gamma : I \to M$ telle que $\gamma(0)$ et $\dot\gamma(0) = v$.*

Démonstration.

Donnons les grandes idées de la preuve. Soit $\gamma : I \to M$ une courbe de classe C^2 telle que $\gamma(0) = p$ et $\dot\gamma(0) = v$. Soient (U,x) une carte locale contenant p et $J \subset I$ tels que $\gamma(J) \subset U$. On note $X_i = \frac{\partial}{\partial x_i}$ pour tout $i \in [\![1,m]\!]$. On peut alors écrire

$$\dot\gamma(t) = \sum_{i=1}^{m} \dot\gamma_i(t)\ (X_i)_{\gamma(t)}.$$

En utilisant (A.2)*, on a*

$$(\nabla_{\dot\gamma}\dot\gamma)(t) = \sum_{k=1}^{m}\left(\ddot\gamma_k(t) + \sum_{i,j=1}^{m} \dot\gamma_i(t)\dot\gamma_j(t)\Gamma_{i,j}^k(\gamma(t))\right)(X_k)_{\gamma(t)}.$$

Ainsi γ géodésique si et seulement si $\nabla_{\dot\gamma}\dot\gamma = 0$ autrement dit pour tout $k \in [\![1,m]\!]$,

$$\ddot\gamma_k(t) + \sum_{i,j=1}^{m} \dot\gamma_i(t)\dot\gamma_j(t)\Gamma_{i,j}^k(\gamma(t)) = 0. \tag{A.3}$$

On reconnaît une équation différentielle et on peut utiliser le théorème de Cauchy-Lispchitz afin de trouver une unique solution locale $(\gamma_1,...,\gamma_m)$. Enfin il suffit seulement de définir γ à l'aide de $(\gamma_1,\cdots,\gamma_m)$. □

Exemple A.9.1. On peut alors calculer les géodésiques de certaines variétés. Par exemple si $(M,g) = (\mathbb{R}^m, \langle,\rangle)$. Pour la carte globale $(\mathbb{R}^m, Id)$, $g_{i,j} = \delta_{i,j}$. Ainsi par (A.1), $\Gamma_{i,j}^k = 0$ pour tout $i,j,k \in [\![1,m]\!]$. Ainsi par (A.3), $\ddot\gamma = 0$. On peut définir $\gamma_{(p,v)}(t) = p + tv$. Elle vérifie bien les conditions initiales et l'équation $\ddot\gamma = 0$. Par unicité des géodésiques vérifiant ces conditions initiales, les géodésiques sont alors les lignes droites.

A.10 Courbure et torsion

Parlons maintenant de la notion de **courbure** et de **torsion**. Soit (M,g) une variété riemannienne muni de la connexion de Levi-Civita. On définit la **dérivée covariante du premier ordre** pour un

champ de vecteur $X \in C^\infty(TM)$ par

$$\begin{aligned}\nabla_X : C^\infty(TM) &\to C^\infty(TM)\\ Z &\mapsto \nabla_X Z.\end{aligned}$$

On généralise cela pour les tenseurs de type $(r,0)$ et $(r,1)$. Soit $A : C^\infty_r(TM) \to C^\infty(M)$ un tenseur du type $(r,0)$. On définit la **dérivée covariante**

$$\begin{aligned}\nabla A : C^\infty_{r+1}(TM) &\to C^\infty_0(TM) = C^\infty(M)\\ (X, X_1, ..., X_r) &\mapsto (\nabla_X A)(X_1, \cdots, X_r)\end{aligned}$$

par

$$(\nabla_X A)(X_1, \cdots, X_r) = X(A(X_1, \cdots, X_r)) - \sum_{k=1}^{r} A(X_1, \cdots, X_{k-1}, \nabla_X X_k, ..., X_r).$$

Un tenseur du type $(r,0)$ est dit **parallèle** si $\nabla A = 0$.

Proposition A.10.1. *La métrique g est un tenseur parallèle de type* $(2,0)$.

Pour chaque $Z \in C^\infty(TM)$, on peut définir un **tenseur naturel** $\Delta_Z = \Delta : C^\infty(TM) \to C^\infty(TM)$ de la manière suivante

$$\Delta(X) = \nabla_X Z.$$

En utilisant que la connexion est multilinéaire, on a

$$\Delta(fX + gY) = f\Delta(X) + g\Delta(Y).$$

Ainsi Δ est un tenseur du type $(1,1)$. De la même manière, on peut définir la dérivée covariante pour les tenseurs du type $(r,1)$. Soit $B : C^\infty_r(TM) \to C^\infty_1(TM) = C^\infty(TM)$, on définit la **dérivée covariante** de B

$$\nabla B : C^\infty_{r+1}(TM) \to C^\infty_1(TM)$$

est

$$(\nabla_X B)(X_1, \cdots, X_r) = \nabla_X(B(X_1, \cdots, X_r)) - \sum_{k=1}^{r} B(X_1, \cdots, X_{k-1}, \nabla_X X_k, \cdots, X_r). \tag{A.4}$$

Un tenseur du type $(r,1)$ est parallèle si $\nabla B = 0$. Enfin définissons maintenant la **dérivée covariante du second ordre**. Soient $X, Y \in C^\infty(TM)$. La dérivée covariante du second ordre est définie par

$$\begin{aligned}\nabla^2_{X,Y} : C^\infty(TM) &\to C^\infty(TM)\\ Z &\mapsto (\nabla_X \Delta_Z)(Y)\end{aligned}$$

où Δ_Z est le tenseur naturel de Z. On remarque que par (A.4), on a

$$\nabla^2_{X,Y} Z = \nabla_X(\Delta_Z(Y)) - \Delta_Z(\nabla_X Y) \tag{A.5}$$
$$= \nabla_X \nabla_Y Z - \nabla_{\nabla_X Y} Z. \tag{A.6}$$

On peut maintenant définir **l'opérateur de courbure** (riemannien) par

$$R(X,Y)Z = \nabla^2_{X,Y} Z - \nabla^2_{Y,X} Z$$

avec $X, Y, Z \in C^\infty(TM)$ et $R : C^\infty(TM) \times C^\infty(TM) \times C^\infty(TM) \to C^\infty(TM)$. Une remarque importante est que l'opérateur de courbure est un tenseur de type $(3,1)$:

Théorème A.10.2. *Soit (M,g) une variété riemannienne muni de la connexion de Levi-Civita. L'opérateur de courbure*

$$R : C^\infty_3(TM) \to C^\infty(TM)$$

satisfait pour tout $X, Y, Z \in C^\infty(TM)$

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

et l'opérateur est un tenseur du type $(3,1)$.

Citons quelques propriétés de cet opérateur de courbure. Soient $W, X, Y, Z \in C^\infty(TM)$:

1. $R(X,Y)Z = -R(Y,X)Z$,
2. $g(R(X,Y)Z,W) = -g(R(X,Y)W,Z)$,
3. Identitié de Bianchi : $R(X,Y)Z + R(Z,X)Y + R(Y,Z)X = 0$,
4. $g(R(X,Y)Z,W) = g(R(Z,W)X,Y)$.

Avant de définir la notion de **courbure sectionnelle**, on a besoin de définir la notion de section.

Definition A.10.1. Soient (M,g) une variété riemannienne et $p \in M$. Une section V en p est un sous espace vectoriel de dimension 2 de T_pM. On note

$$G_2(T_pM) = \{V,\ \text{V section de } T_pM\}.$$

Cet ensemble est appelé la **grassmannienne** des 2-plans.

Enonçons un lemme utile.

Lemme A.10.3. *Soient $X, Y, Z, W \in T_pM$ tels que $Vect(X,Y) = Vect(Z,W)$. Alors*

$$\frac{g(R(X,Y)Y,X)}{|X|^2|Y|^2 - g(X,Y)^2} = \frac{g(R(Z,W)W,Z)}{|Z|^2|W|^2 - g(Z,W)^2}.$$

On peut maintenant définir **la courbure sectionnelle**.

Definition A.10.2. Soient (M,g) une variété riemannienne et $p \in M$. Alors la fonction $K_p : G_2(T_pM) \to \mathbb{R}$ définie par

$$Vect(X,Y) \to \frac{g(R(X,Y)Y,X)}{|X|^2|Y|^2 - g(X,Y)^2}$$

est appelée la **courbure sectionnelle** en p. On note $K(X,Y)$ à la place de $K(Vect(X,Y))$.

Définissons maintenant deux fonctions utiles

$$\phi : p \mapsto \min_{V \in G_2(T_pM)} K_p(V),$$
$$\psi : p \mapsto \max_{V \in G_2(T_pM)} K_p(V).$$

Alors la variété riemannienne (M,g) est dite :

1. à **courbure positive** si $\phi(p) \geq 0$ pour tout $p \in M$,
2. à **courbure strictement positive** si $\phi(p) > 0$ pour tout $p \in M$,
3. à **courbure négative** si $\psi(p) \leq 0$ pour tout $p \in M$,
4. à **courbure strictement négative** si $\psi(p) < 0$ pour tout $p \in M$,
5. à **courbure constante** si $\phi = \psi$ est constant,
6. **plat** si $\phi = \psi = 0$.

On peut exprimer les coefficients de la courbure par les symboles de Christoffel. Soit (U,x) une carte locale de M. Notons comme d'habitude $X_i = \frac{\partial}{\partial x_i}$. Pour tout $i,j,k,l \in [\![1,m]\!]$:

$$g_{i,j} = g(X_i, X_j), \quad R_{ijkl} = g(R(X_i,X_j)X_k, X_l).$$

On peut montrer que :

$$R_{ijkl} = \sum_{s=1}^{m} g_{sl}\left(\frac{\partial \Gamma^s_{j,k}}{\partial x_i} - \frac{\partial \Gamma^s_{i,k}}{\partial x_j} + \sum_{r=1}^{m}(\Gamma^r_{j,k}\Gamma^s_{i,r} - \Gamma^r_{i,k}\Gamma^s_{j,r})\right). \tag{A.7}$$

Pour cela, on utilise le fait $[X_i, X_j] = 0$ pour tout $i,j \in [\![1,m]\!]$ et

$$\begin{aligned} R(X_i,X_j)X_k &= \nabla_{X_i}\nabla_{X_j}X_k - \nabla_{X_j}\nabla_{X_i}X_k \\ &= \sum_{s=1}^{m}\left(\nabla_{X_i}(\Gamma^s_{jk}X_s) - \nabla_{X_j}(\Gamma^s_{ik}X_s)\right). \end{aligned}$$

En développant et en utilisant la définition de connexion et des symboles de Christoffel, on obtient

$$R(X_i, X_j)X_k = \sum_{s=1}^{m} \left(\frac{\partial \Gamma^s_{j,k}}{\partial x_i} - \frac{\partial \Gamma^s_{i,k}}{\partial x_j} + \sum_{r=1}^{m} (\Gamma^r_{j,k}\Gamma^s_{i,r} - \Gamma^r_{i,k}\Gamma^s_{j,r}) \right) X_s.$$

Exemple A.10.1. Soit $(M, g) = (\mathbb{R}^m, <,>)$. L'ensemble $\{\partial/\partial x_i\}_{i\in[\![1,m]\!]}$ est un repère global du fibré tangent $T\mathbb{R}^m$. On a que $g_{i,j} = 0$ donc $\Gamma^k_{i,j} = 0$ pour tout $i, j, k \in [\![1, m]\!]$. Par (A.7), $R_{ijkl} = 0$ pour tout $i, j, k, l \in [\![1, m]\!]$. On en déduit que l'espace est **plat**.

Enfin notons au passage que les coefficients $\Gamma_{ij,k} = g(\tilde{\nabla}_{X_i} X_j, X_k)$ avec $i, j, k \in [\![1, m]\!]$ détermine la connexion métrique $\tilde{\nabla}$. En effet montrons que

$$\partial_k g_{i,j} = \Gamma_{ki,j} + \Gamma_{kj,i}.$$

Puisque la connexion ∇ est métrique et $X_k = \partial_k$ avec $k \in [\![1, m]\!]$

$$\begin{aligned} X_k(g(X_i, X_j)) &= g(\nabla_{X_k} X_i, X_j) + g(\nabla_{X_k} X_j, X_i) \\ &= \Gamma_{ki,j} + \Gamma_{kj,i}. \end{aligned}$$

Les coefficients $\Gamma_{ij,k}$ caractérisent la connexion. En effet pour tout $k \in [\![1, m]\!]$

$$\begin{aligned} \Gamma_{ij,k} &= g(\nabla_{X_i} X_j, X_k) \\ &= \sum_{s=1}^{m} \Gamma^s_{i,j} g_{s,k}. \end{aligned}$$

D'où

$$\Gamma_{i,j} = g^{-1}\Gamma_{ij,\cdot}$$

où $\Gamma_{i,j} = \left(\Gamma^k_{i,j}\right)_{k\in[\![1,m]\!]}$.

Enfin énonçons un théorème fondamental permettant de relier la planéité et la nullité des symboles de Christoffel est le suivant :

Théorème A.10.4. *Une variété riemannienne (M, g) est plate si et seulement si près de chaque point il existe un système de coordonnées locales tel que $g = \sum_{i=1}^{m} (dx^i)^2$. Ainsi pour ce système de coordonnées locales, les symboles de Christoffel sont nuls.*

De manière générale, il est possible de démontrer cela pour une connexion ∇ générale dès que sa torsion et sa courbure sont nulles. C'est une condition nécessaire et suffisante.

Annexe B : Optimal use of auxiliary information : information geometry and empirical process

Le chapitre 1, 2 et les sections 3.1, 5.1 ont fait l'objet d'un article prépublié [8]. Je joins dans cet annexe l'article.

B.1 Introduction

We call auxiliary information –or side information– any information external to an observed statistical experiment that concerns the underlying distribution. For instance, in order to improve the quality of a survey analysis, it is customary to incorporate any reliable auxiliary information available at the time of the survey such as the knowledge of one or more parameters of this population determined exactly by an exhaustive census. This principle finds its origin several centuries ago, according to [12]. Indeed, around 1740 the magistrate Jean-Baptiste François de La Michodière wanted to estimate the size of the French population by assuming that the number of marriages, births and deaths is proportional to the size of the population. He then introduced the ratio estimator which was validated by Laplace [24]. This method is for instance detailed in [23]. We note that it turns out to be a special case of [37]. More recently, several authors in survey analysis and statistics have worked on the incorporation of auxiliary information after or before sampling – see [16], [27].

In this article, we focus on the auxiliary information which concerns the underlying distribution of the data. More precisely, we assume that the auxiliary information is given by a finite collection of expectations. In the above mentioned case of a survey the side information can be given by the expectation of a random variable on the population. Rather few systematic analysis have been carried out on how to use at best such an auxiliary information in a general setting, despite the fact that in many case studies such a methodology is used. In [5], the raking-ratio method allows to incorporate the auxiliary information of the probabilities of one or many partitions of a set. This is not a perfect projection of the empirical measure on the set of constraints since it is a sequential procedure, incorporating each information after the other, not simultaneously. However the variance of large classes of estimators simultaneously decreases as the sample size tends to infinity faster than the number of successive partitions. The case of an independent empirical information,

as for distributed data, is investigated in [2]. In [37], a general method is proposed to incorporate a general auxiliary information. This approach consists in minimizing the variance over a class of unbiased estimators, that is equivalent to find the smallest dispersion ellipsoid. In [4] this method is applied to an auxiliary information brought by a finite collection of expectations. In particular, it is shown that this method is better than the raking-ratio [5] with respect to variance reduction. Moreover, in [32] a different method based on empirical likelihood is developed to also incorporate an auxiliary information given by a finite collection of expectations. The latter two methods will be compared and connected through our definition of an informed empirical measure. In [41] the Glivenko-Cantelli and Donsker theorems are established for the empirical likelihood method – in the special case of the class of functions $\mathscr{F} = \left\{1_{]-\infty,t]},\ t \in \mathbb{R}\right\}$. For more general classes, the Glivenko-Cantelli and Donsker theorems are established in [4] under stronger assumptions by using the strong approximation results of [9]. In addition, some works have also been done on U-statistics in the presence of auxiliary information [39] and, more recently, on informed statistical tests [1], [3].

Our contribution is to define and study an informed empirical measure supported by the sample that is optimal in the sense of information geometry. We thus intend to incorporate the auxiliary information given by expectations into the empirical measure itself by defining properly the geometrical setting in which the latter can be projected. This leads to define two projection measures, the first of which satisfies the same optimization problem as that of the empirical likelihood [31]. Next we prove that it is possible to approximate these two projection measures by a common measure we call the informed empirical measure. This informed empirical measure is far easier to compute numerically than the true projections and turns out to coincide with the adaptive estimator of the measure with auxiliary information defined in [4]. Furthermore we show that these three measures are so close that they share the same asymptotic properties in the sense of empirical process theory – in particular the same limiting Gaussian process. This allows to unify several methods aiming to incorporate a side information, among which those mentioned above. We establish under minimal assumptions the limit theorems for the informed empirical measure indexed by a general class of functions. As a by product this extends the asymptotic result of [41]. Moreover we derive a concentration result which shows that the informed empirical process is always more concentrated than the classical empirical process when the sample size is large enough.

The paper is structured as follows. In Section B.3, we introduce the geometrical framework and prove that the set of constraints associated to the auxiliary information has a submanifold structure. Then we show that an optimal method is to minimize the Kullback-Leibler divergence and its dual version on the set of constraints. The readers less familiar with geometrical notions can find reminders on information geometry in the book [6] – or could refer directly to Corollary 1. In Section B.4, we study these two optimization problems. More precisely, we discuss the existence and uniqueness of their solution and we give an asymptotic theorem for the Lagrange multipliers. In Section B.5, we prove that there exists a common approximation of these solutions which allows to define the informed empirical measure – see Definition 1. Then we study the informed empirical measure's weights. In Section B.6 we establish the Glivenko-Cantelli and Donsker theorems for a general class of functions $\mathscr{F}$ about the informed empirical process under minimal assumptions. We also quantify the asymptotic uniform variance reduction, which justifies the use of auxiliary information. In Section B.6.2 we derive a concentration result about the informed empirical process.

Finally in Section B.7, as an illustration of the variance reduction, we apply these results to the informed empirical quantile and prove that the informed estimator is asymptotically more efficient than the classical empirical quantile. As a special case, we find the same asymptotic result as in [42].

B.2 Framework

Let $(X_n)_{n\in\mathbb{N}^*}$ be a sequence of independent and identically distributed random variables (*i.i.d.*) defined on a probability space $(\Omega,\mathscr{B},\mathbb{P})$ and taking values in a measurable space $(\mathscr{X},\mathscr{A})$. The distribution of X_1 is denoted $P := \mathbb{P}^{X_1}$. Moreover, we assume that an auxiliary information I about P is available. In this article, we focus on the following particular case. We suppose that I is the information brought by a finite collection of expectations with respect to P. More precisely, let $m \in \mathbb{N}^*$ and $g = \left(g_1, \cdots, g_m\right)^T : \mathscr{X} \to \mathbb{R}^m$ be an integrable function with respect to P. We assume that I is given by

$$Pg := \int_{\mathscr{X}} g dP := \left(\int_{\mathscr{X}} g_1 dP, \cdots, \int_{\mathscr{X}} g_m dP\right)^T.$$

We shall assume at times that $Pg = 0$ – otherwise set $h = g - Pg$.

We denote $[\![n,p]\!]$ the set of integers between n and p where $n < p$ are two integers. For each $n \in \mathbb{N}^*$, the random data set is denoted $\mathscr{Z}_n = \{X_1, \cdots, X_n\}$ and $\mathbb{P}_n$ the empirical measure defined by

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

If $n \in \mathbb{N}^*$ is fixed then we denote $\mathscr{Z} = \mathscr{Z}_n$ and $\mathscr{P}(\mathscr{Z})$ the set of probability measures on $\mathscr{Z}$ with positive weights. Moreover the informed empirical measure is denoted $\mathbb{P}_n^I$ and will be defined at Section B.5 in Definition 1.

Our notation for stochastic convergences are as follows. Let $(Y_n)_{n\in\mathbb{N}^*}$ a sequence of random variables with values in $\mathbb{R}^m$ with $m \in \mathbb{N}^*$ and let $(a_n)_{n\in\mathbb{N}^*} \subset \mathbb{R}^*$ be a real valued sequence. We write $Y_n = o_p(a_n)$ (resp. $o_{a.s.}(a_n)$) if $(Y_n/a_n)_{n\in\mathbb{N}^*}$ tends to 0 in probability (resp. almost surely) as $n \to +\infty$. We write $Y_n = O_p(a_n)$ if $(Y_n/a_n)_{n\in\mathbb{N}^*}$ is tight. The fact that $(Y_n)_{n\in\mathbb{N}^*}$ converges in distribution to a random variable Y as $n \to +\infty$ is denoted by $Y_n \Rightarrow Y$.

B.3 A geometrical approach of auxiliary information

We intend to incorporate optimally an auxiliary information into the empirical measure. The notion of optimality may be debated but the information geometry approach seems to be a coherent and interesting answer to the problem of incorporating an auxiliary information.

Assume that an auxiliary information I about P is available. For $Q \in \mathscr{P}(\mathscr{Z})$, we denote $Q \sim I$ the fact that Q satisfies the auxiliary information I. The set of probability measures on $\mathscr{Z}$ which satisfy I is defined by

$$\mathscr{P}^I(\mathscr{Z}) = \left\{Q \in \mathscr{P}(\mathscr{Z}),\ \ Q \sim I\right\}.$$

As mentioned in Section B.2, the weights of Q are positive and I is given by a finite collection of expectations Pg. However Section B.3.2 remains valid for more general definitions of auxiliary

information I. We have

$$\mathscr{P}^I(\mathscr{Z}) = \left\{Q \in \mathscr{P}(\mathscr{Z}),\ \ Qg = Pg\right\}.$$

Assuming that the basic notions of information geometry are known – such as connection, geodesic etc – we only recall the notion of autoparallel submanifold. Let S be a manifold, M be a submanifold of S and ∇ a connection on S. We say that M is ∇-autoparallel if for every vector fields X, Y on M, $\nabla_X Y$ is also a vector field on M. In this context, since $\mathscr{P}(\mathscr{Z})$ is a finite mixture model it is also a differential manifold endowed with the dually flat structure $(\mathscr{P}(\mathscr{Z}), g_F, \nabla^{(1)}, \nabla^{(-1)})$ where g_F is the Fisher metric, $\nabla^{(1)}$ is the 1-connection and $\nabla^{(-1)}$ is the (-1)-connection. Moreover, the canonical divergence associated to $\mathscr{P}(\mathscr{Z})$ is the Kullback–Leibler divergence KL – see [30].

In order to use the auxiliary information I let project $\mathbb{P}_n \in \mathscr{P}(\mathscr{Z})$ on $\mathscr{P}^I(\mathscr{Z})$ in the sense of information geometry. To define properly this projection we first show that $\mathscr{P}^I(\mathscr{Z})$ is a submanifold, then we recall the projection theorem in information geometry and formulate our existence result.

B.3.1 Submanifold structure of $\mathscr{P}^I(\mathscr{Z})$

The following result states that $\mathscr{P}^I(\mathscr{Z})$ is a $(n-2)$-dimensional submanifold in the case $m = 1$.

Proposition 1. *Assume that there exists $i \neq j$ such that $g(X_i) \neq g\left(X_j\right)$ and Pg belongs to the convex hull of $\left\{g(X_1), \cdots, g(X_n)\right\}$. Then $\mathscr{P}^I(\mathscr{Z})$ is a $(n-2)$-dimensional submanifold of $\mathscr{P}(\mathscr{Z})$.*

Proof. *Remind that $]0,1[^n$ is a submanifold of $\mathbb{R}^n$ as it is an open set of $\mathbb{R}^n$. Set*

$$\mathscr{S} = \left\{q \in]0,1[^n,\ \sum_{i=1}^n q_i = 1\right\}.$$

We first show that $\mathscr{S}$ is a $(n-1)$-dimensional submanifold of $]0,1[^n$. Define the function $\theta :]0,1[^n \to \mathbb{R}$ by $\theta(q) = \sum_{i=1}^n q_i$.

Observe that $\mathscr{S} = \theta^{-1}(\{1\})$. So it is enough to prove that θ is a submersion. Indeed, θ is differentiable and for all $q \in]0,1[^n$,

$$D\theta(q)(h) = \theta(h),\ h \in \mathbb{R}^n.$$

So $D\theta(q)$ is surjective and θ is a submersion, hence $\mathscr{S}$ is a $(n-1)$-dimensional submanifold of $]0,1[^n$. Let endow $\mathscr{S}$ with the following global chart $(\mathscr{S}, \pi)$

$$\begin{aligned} \pi : \mathscr{S} &\to U \subset \mathbb{R}^{n-1} \\ q &\mapsto (q_1, \cdots, q_{n-1}) \end{aligned}$$

where $U = \left\{(q_1, \cdots, q_{n-1}) \in \mathbb{R}^{n-1},\ q_i > 0,\ i \in [\![1, n-1]\!],\ \sum_{i=1}^{n-1} q_i < 1\right\}$. Similarly endow $\mathscr{P}(\mathscr{Z})$ with the following global chart $\left(\mathscr{P}(\mathscr{Z}), \varphi\right)$

$$\begin{aligned} \varphi : \mathscr{P}(\mathscr{Z}) &\to U \subset \mathbb{R}^{n-1} \\ Q &\mapsto (q_1, \cdots, q_{n-1}). \end{aligned}$$

Next consider the following one to one mapping

$$\begin{aligned}\psi : \mathscr{P}(\mathscr{Z}) &\to \mathscr{S}\\ Q &\mapsto q.\end{aligned}$$

Notice that $\psi = \pi^{-1} \circ \varphi$. *Since* π *and* φ *are diffeomorphims and* $\dim \mathscr{P}(\mathscr{Z}) = \dim \mathscr{S} = n-1$, *we deduce that* ψ *is a diffeomorphim. So* ψ^{-1} *is also a diffeomorphism and thus an embedding. Observe that* $\mathscr{P}^I(\mathscr{Z}) = \psi^{-1}(\mathscr{E})$ *where* $\mathscr{E} = \left\{q \in \mathscr{S},\ \sum_{i=1}^n q_i g(X_i) = Pg\right\}$ *is not empty because* Pg *belongs to the convex hull of* $\{g(X_1), ..., g(X_n)\}$. *So it is enough to prove that* $\mathscr{E}$ *is a* $(n-2)$*-dimensional submanifold of* $\mathscr{S}$. *For this, it is sufficient to verify that*

$$\begin{aligned}f : \mathscr{S} &\to \mathbb{R}\\ q &\mapsto \sum_{i=1}^n q_i g(X_i)\end{aligned}$$

is a submersion. Let $\gamma := f \circ \pi^{-1} : U \to \mathbb{R}$ *be defined, for all* $q \in U$, *by*

$$\gamma(q) = \sum_{i=1}^{n-1} q_i g(X_i) + \left(1 - \sum_{i=1}^{n-1} q_i\right) g(X_n).$$

So γ *is differentiable and for all* $q \in U$

$$D\gamma(q) = (g(X_1) - g(X_n), \cdots, g(X_{n-1}) - g(X_n)).$$

Since there exists $i \neq j$ *such that* $X_i \neq X_j$, *we deduce that* $D\gamma(q)$ *is surjective. Therefore* f *is a submersion and* $\mathscr{E}$ *is a* $(n-2)$*-dimensional submanifold. We conclude that* $\mathscr{P}^I(\mathscr{Z}) = \psi^{-1}(\mathscr{E})$ *is a* $(n-2)$*-dimensional submanifold, as* ψ^{-1} *is an embedding.* □

Proposition 1 can be generalised for a vector of functions $g = (g_1, \cdots, g_m)^T$ with $m \in [\![1, n-1]\!]$. Assume that $Pg = 0$. Set

$$\begin{aligned}&\forall j \in [|1, m|],\ N_j(X) = (g_j(X_1) - g_j(X_n), \cdots, g_j(X_{n-1}) - g_j(X_n))^T,\\ &\forall j \in [|1, m|],\ g_j(X) = (g_j(X_1), \cdots, g_j(X_n))^T.\end{aligned}$$

Observe that

$$\dim\big(Vect(g(X_1) - g(X_n), \cdots, g(X_{n-1}) - g(X_n))\big) = \dim\big(Vect((N_j(X))_{1\le j\le m})\big). \tag{B.1}$$

We are ready to state the main result of Section B.3.

Proposition 2. *Assume that* 0 *belongs to the convex hull of* $\{g(X_1), ..., g(X_n)\}$ *and the following equality of dimensions,*

$$l := \dim\big(Vect(g_1(X), ..., g_m(X))\big) = \dim\big(Vect(g(X_1) - g(X_n), \cdots, g(X_{n-1}) - g(X_n))\big) \le m.$$

Then $\mathscr{P}^I(\mathscr{Z})$ *is a* $(n-1-l)$*-dimensional submanifold of* $\mathscr{P}(\mathscr{Z})$.

Remark 1. *A sufficient condition to satisfy the equality of dimensions in Proposition 2 is*

$$\dim Vect\left((1,\cdots,1)^T, M_1(X),\cdots,M_m(X)\right) = m+1 \tag{B.2}$$

with for all $k \in [|1,m|]$, $M_k(X) = g_k(X) = \left(g_k(X_1),\cdots,g_k(X_n)\right)^T$.

Proof. *We keep the same notations as in the previous prooof. It remains to prove that* $\mathscr{E} = \left\{q \in \mathscr{S},\ \sum_{i=1}^n q_i g(X_i) = 0\right\}$ *is a* $(n-1-l)$*-dimensional submanifold. For that, we need a technical lemma.*

Lemma 1. *Let* $x^1,...,x^k \in \mathbb{R}^n$ *with* $k \in [\![1,n-1]\!]$ *and* $n \in \mathbb{N}^*$. *Define, for* $j \in [\![1,k]\!]$,

$$\tilde{x}^j = \left(x_1^j - x_n^j,\cdots,x_{n-1}^j - x_n^j\right)^T \in \mathbb{R}^{n-1}.$$

Then

$$\dim\left(Vect(\tilde{x}^1,\cdots,\tilde{x}^k)\right) \leq \dim\left(Vect(x^1,\cdots,x^k)\right).$$

Moreover if $l := \dim\left(Vect(\tilde{x}^1,\cdots,\tilde{x}^k)\right) = \dim\left(Vect(x^1,\cdots,x^k)\right)$
then there exists a subset $J \subset [\![1,k]\!]$ *of size* l *such that*

$$\dim\left(Vect\left((\tilde{x}^j)_{j\in J}\right)\right) = \dim\left(Vect\left((x^j)_{j\in J}\right)\right).$$

Proof. *If* $k = \dim\left(Vect(x^1,\cdots,x^k)\right)$ *then there is nothing to prove. Assume that* $\dim\left(Vect(x^1,\cdots,x^k)\right) < k$. *So there exists* $j \in [\![1,k]\!]$ *such that* $x^j = \sum_{i\neq j}\lambda_i x^i$ *with* $\lambda_1,...,\lambda_k \in \mathbb{R}$. *So for all* $r \in [\![1,n]\!]$

$$x_r^j - x_n^j = \sum_{i\neq j}\lambda_i x_r^i - \sum_{i\neq j}\lambda_i x_n^i = \sum_{i\neq j}\lambda_i (x_r^i - x_n^i).$$

In others words $\tilde{x}^j = \sum_{i\neq j}\lambda_i \tilde{x}^i$. *We can conclude that*

$$\dim\left(Vect(\tilde{x}^1,\cdots,\tilde{x}^k)\right) \leq \dim\left(Vect(x^1,\cdots,x^k)\right).$$

Now, assume that $l := \dim\left(Vect(\tilde{x}^1,\cdots,\tilde{x}^k)\right) = \dim\left(Vect(x^1,\cdots,x^k)\right)$. *So there exists a subset* $J \subset [\![1,k]\!]$ *of size* l *such that*

$$l = \dim\left(Vect\left((\tilde{x}^j)_{j\in J}\right)\right).$$

But if $\dim\left(Vect\left((x^j)_{j\in J}\right)\right) < l$ *then there exists* $r \in J$ *such that* $x^r = \sum_{i\in J,\ i\neq r}\lambda_i x^i$ *with* $\lambda_1,...,\lambda_k \in \mathbb{R}$. *By the above, we deduce that* $\tilde{x}^r = \sum_{i\in J,\ i\neq r}\lambda_i \tilde{x}^i$. *That contradicts the fact that*

$$l = \dim\left(Vect\left((\tilde{x}^j)_{j\in J}\right)\right).$$

□

We can apply Lemma 1 to $g_1(X), \cdots, g_m(X)$. *Recall that, by (B.1),*

$$\dim\big(Vect(g(X_1)-g(X_n), \cdots, g(X_{n-1})-g(X_n))\big) = \dim\big(Vect((N_j(X))_{1\le j\le m})\big)$$

where, for $j \in [\![1, m]\!]$,

$$N_j(X) = (g_j(X_1)-g_j(X_n), \cdots, g_j(X_{n-1})-g_j(X_n))^T.$$

By the assumption of equality of dimensions we have

$$l := \dim\big(Vect(g_1(X), \cdots, g_m(X))\big) = \dim\big(Vect(g(X_1)-g(X_n), \cdots, g(X_{n-1})-g(X_n))\big).$$

Hence by Lemma 1 there exists a subset $J \subset [\![1, m]\!]$ *of size* l *such that*

$$l = \dim\big(Vect\big(g_1(X), ..., g_m(X)\big)\big) = \dim\big(Vect\big((g_j(X))_{j\in J}\big)\big) = \dim\big(Vect\big((N_j(X))_{j\in J}\big)\big)$$

Denote $\widetilde{g} = (g_j)_{j\in J}$ *and notice that*

$$\big\{Q \in \mathscr{P}(\mathscr{Z}),\ \ Qg = 0\big\} = \big\{Q \in \mathscr{P}(\mathscr{Z}),\ \ Q\widetilde{g} = 0\big\}.$$

So, we can select only the constraints $\widetilde{g}$. *For simplicity, in what follows we denote* $g = (g_j)_{j\in J}$. *Define the function*

$$\begin{aligned} f : \mathscr{S} &\to \mathbb{R}^l \\ q &\mapsto \sum_{i=1}^{n} q_i g(X_i). \end{aligned}$$

Let prove that f *is a submersion. Set* $\gamma := f \circ \pi^{-1} : U \to \mathbb{R}^l$ *defined for all* $q \in U$

$$\gamma(q) = \sum_{i=1}^{n-1} q_i g(X_i) + \left(1 - \sum_{i=1}^{n-1} q_i\right) g(X_n).$$

So γ *is differentiable and for all* $q \in U$

$$D\gamma(q) = (g(X_1)-g(X_n), \cdots, g(X_{n-1})-g(X_n)).$$

Since

$$rk\big(D\gamma(q)\big) = \dim\big(Vect(g(X_1)-g(X_n), \cdots, g(X_{n-1})-g(X_n))\big) = l.$$

We deduce that $D\gamma(q)$ *is surjective. Thus* f *is a submersion and* $\mathscr{E}$ *is a* $(n-1-l)$*-dimensional submanifold. We can conclude that* $\mathscr{P}^I(\mathscr{Z}) = \psi^{-1}(\mathscr{E})$ *is* $(n-1-l)$*-dimensional submanifold because* ψ^{-1} *is an embedding.* □

B.3.2 Existence of a projection

Recall the projection theorem in information geometry.

Theorem 1. *Consider a dually flat manifold S and denote D the canonical divergence of S. Let $p \in S$ and M be a submanifold of S which is ∇^*-autoparallel. Then a necessary and sufficient condition for a point $q \in M$ to satisfy*

$$D(p||q) = \min_{r \in M} D(p||r) \tag{B.3}$$

is that the ∇-geodesic connecting p to q is orthogonal to M in q. The point q is called ∇-projection of p on M. Likewise if M is ∇-autoparallel then a necessary and sufficient condition for a point $q \in M$ to satisfy

$$D^*(p||q) = \min_{r \in M} D^*(p||r) \tag{B.4}$$

is that the ∇^*-geodesic connecting p to q is orthogonal to M in q and the point q is called the ∇^*-projection of p on M.

Moreover, it is possible to relax the autoparallel submanifold assumption.

Proposition 3. *Assume that S a dually flat manifold and denote D the canonical divergence of S. Let $p \in S$ and M be a submanifold of S. A necessary and sufficient condition for a point q to be a stationary point of the function $D(p||\cdot) : r \mapsto D(p||r)$ restraint to M (resp. $D^*(p||\cdot) : r \mapsto D^*(p||r)$) is that the ∇-geodesic (resp. ∇^*-geodesic) connecting p to q is orthogonal to M in q.*

We next apply Theorem 1 and Proposition 3 to define a projection $\mathbb{Q}_n^I$ of $\mathbb{P}_n$ on $\mathscr{P}^I(\mathcal{Z})$.

Corollary 1. *Assume that $\mathscr{P}^I(\mathcal{Z})$ is a submanifold of $\mathscr{P}(\mathcal{Z})$. Then*

- *If $\mathscr{P}^I(\mathcal{Z})$ is a submanifold $\nabla^{(-1)}$-autoparallel then $\mathbb{Q}_n^I$ is the $\nabla^{(1)}$-projection of $\mathbb{P}_n$ on $\mathscr{P}^I(\mathcal{Z})$ that is*

$$\mathbb{Q}_n^I \in arg \min_{Q \in \mathscr{P}^I(\mathcal{Z})} KL^*(\mathbb{P}_n||Q) = arg \min_{Q \in \mathscr{P}^I(\mathcal{Z})} KL(Q||\mathbb{P}_n).$$

- *If $\mathscr{P}^I(\mathcal{Z})$ is a submanifold $\nabla^{(1)}$-autoparallel then $\mathbb{Q}_n^I$ is the $\nabla^{(-1)}$-projection of $\mathbb{P}_n$ on $\mathscr{P}^I(\mathcal{Z})$*

$$\mathbb{Q}_n^I \in arg \min_{Q \in \mathscr{P}^I(\mathcal{Z})} KL(\mathbb{P}_n||Q).$$

- *In the case where $\mathscr{P}^I(\mathcal{Z})$ is autoparallel to neither of these two connections then $\mathbb{Q}_n^I$ is a stationary point of one of these two maps*

$$Q \longmapsto KL(\mathbb{P}_n||Q),$$
$$Q \longmapsto KL^*(\mathbb{P}_n||Q).$$

Remark 2. *Corollary 1 does not determine a unique informed empirical measure by projection and moreover it is generally not easy to check if the submanifold $\mathscr{P}^I(\mathcal{Z})$ is autoparallel.*

B.4 Two measure projections

Assume that Pg belongs to the convex hull of $\{g(X_1),\cdots,g(X_n)\}$ and the assumption (B.2) is verified then by Proposition 2 $\mathscr{P}^I(\mathscr{Z})$ is a submanifold. By Theorem 1 we were able to define a projection of $\mathbb{P}_n$ on $\mathscr{P}^I(\mathscr{Z})$. The next step is to study these two optimization problems

$$arg \min_{Q\in\mathscr{P}^I(\mathscr{Z})} KL(\mathbb{P}_n||Q), \tag{B.5}$$

$$arg \min_{Q\in\mathscr{P}^I(\mathscr{Z})} KL(Q||\mathbb{P}_n). \tag{B.6}$$

Remark that

$$KL(\mathbb{P}_n||Q) = \sum_{i=1}^n \frac{1}{n}\log\left(\frac{1}{nq_i}\right) = -\log(n) - \frac{1}{n}\sum_{i=1}^n \log q_i,$$
$$KL(Q||\mathbb{P}_n) = \sum_{i=1}^n q_i \log(nq_i) = \log n + \sum_{i=1}^n q_i \log q_i.$$

Therefore

$$arg \min_{Q\in\mathscr{P}^I} KL(\mathbb{P}_n||Q) = arg \max_{Q\in\mathscr{P}^I(\mathscr{Z})} \sum_{i=1}^n \log q_i = arg \max_{Q\in\mathscr{P}^I(\mathscr{Z})} \prod_{i=1}^n q_i,$$
$$arg \min_{Q\in\mathscr{P}^I(\mathscr{Z})} KL(Q||\mathbb{P}_n) = arg \min_{Q\in\mathscr{P}^I(\mathscr{Z})} \sum_{i=1}^n q_i \log q_i.$$

B.4.1 First optimization problem

The first optimization problem (B.5) is called empirical likelihood and has been studied by A.B. Owen – refer to [31], [32], [34] and [40]. More precisely, the first optimisation problem is the following

$$\begin{cases} \max_{\mathbf{q}} \sum_{i=1}^n \log(q_i) \\ \sum_{i=1}^n q_i = 1, \\ \forall i \in [\![1,n]\!]\ ,\ q_i > 0, \\ n\sum_{i=1}^n q_i(g(X_i) - Pg) = 0. \end{cases}$$

Denote $\mathscr{C}_n$ the constraints set

$$\mathscr{C}_n = \left\{ q \in [0,1]^n,\ \sum_{i=1}^n q_i = 1,\ n\sum_{i=1}^n q_i(g(X_i) - Pg) = 0 \right\}.$$

Moreover denote also

$$M_k(X) = (g_k(X_1),\cdots,g_k(X_n))^T,\ k \in [\![1,m]\!],$$
$$f(q) = \sum_{i=1}^n \log q_i,\ q \in]0,1[^n.$$

The following theorem ensures that there is a unique solution to this problem.

Theorem 2. *Assume that Pg belongs to the convex hull of $\{g(X_1), \cdots, g(X_n)\}$ and*

$$\dim Vect\left((1, \cdots, 1)^T, M_1(X), \cdots, M_m(X)\right) = m+1. \tag{B.7}$$

Then the first optimization problem has a unique solution $q^* = (q_1^*, \cdots, q_n^*) \in \mathscr{C}_n$. Moreover for all $i \in [\![1, n]\!]$

$$\begin{aligned} q_i^* &> 0, \\ q_i^* &= \frac{1}{n} \frac{1}{1 + \langle \lambda^*, g(X_i) - Pg \rangle} \end{aligned}$$

for a unique $\lambda^* \in \bigcap_{i=1}^n \left\{\lambda \in \mathbb{R}^m,\ 1 + \langle \lambda, g(X_i) - Pg \rangle \geq \frac{1}{n}\right\}$.

Remark 3. *The assumption (B.7) is not restrictive since if it is not verified that means that some constraints are redundant.*

Assume that $Pg = 0$ and write $\hat{\lambda}_n = \lambda^*$ to state an asymptotic result.

Theorem 3. *Assume that $\mathbb{E}\|g(X)\|^2 < +\infty$ and $\Sigma = Pgg^T$ is positive definite. Moreover, we suppose that the assumptions of the theorem 2 are satisfied. Set $\Sigma_n = P_n gg^T$. Then we have*

- $\max_{1 \leq i \leq n} | < \hat{\lambda}_n, g(X_i) > | = o_p(1)$.
- $\hat{\lambda}_n = \Sigma_n^{-1} \mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right)$.
- $\sqrt{n}\hat{\lambda}_n \Rightarrow \mathscr{N}(0, \Sigma^{-1})$.

Remark 4. *We can replace Σ_n by the empirical variance $Var_n(g)$ of g since*

$$\Sigma_n = Var_n(g) + (\mathbb{P}_n g)^T \mathbb{P}_n g = Var_n(g) + o_p\left(\frac{1}{\sqrt{n}}\right).$$

The last two theorems can be proved by using a similar approach that in [32].

B.4.2 Second optimization problem

Let consider the second optimization problem (B.6)

$$\begin{cases} \min_{\mathbf{q}} \sum_{i=1}^n q_i \log(q_i) \\ \sum_{i=1}^n q_i = 1, \\ \forall i \in [\![1, n]\!],\ q_i > 0, \\ n \sum_{i=1}^n q_i (g(X_i) - Pg) = 0. \end{cases}$$

Set

$$\varphi(q) = \sum_{i=1}^n q_i \log q_i,\ q \in]0, 1[^n.$$

The following result ensures that there is a unique solution to the second optimization problem.

Theorem 4. *Assume that Pg belongs to the convex hull of $\{g(X_1), \cdots, g(X_n)\}$ and*

$$\dim Vect\left((1, \cdots, 1)^T, M_1(X), \cdots, M_m(X)\right) = m+1. \tag{B.8}$$

Then the second optimization problem has a unique solution $q^* = (q_1^*, \cdots, q_n^*) \in \mathscr{C}_n$. Moreover for all $i \in [|1, n|]$

$$q_i^* = \frac{\exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)}$$

for a unique $\lambda^* \in \mathbb{R}^m$.

Proof. *Existence. Since $\mathscr{C}_n$ is compact and φ is continuous, there exists $q^* \in \mathscr{C}_n$ such that*

$$\sum_{i=1}^n q_i^* \log q_i^* = \min_{q \in \mathscr{C}_n} \sum_{i=1}^n q_i \log q_i.$$

Uniqueness. By the assumption (B.8) and q^* is a global maximum of φ on $\mathscr{C}_n$, we can apply the Lagrange multiplier theorem. So there exists $(\lambda^*, \mu) \in \mathbb{R}^m \times \mathbb{R}$ such that

$$D\varphi(q^*) = \mu(1, \cdots, 1)^T + \sum_{j=1}^m \lambda_j^* \left(g_j(X_1) - Pg_j, \cdots, g_j(X_n) - Pg_j\right)^T.$$

So for all $i \in [\![1, n]\!]$

$$\log q_i^* + 1 = \mu + \langle \lambda^*, g(X_i) - Pg\rangle.$$

Since $\sum_{i=1}^n q_i^* = 1$, we deduce that

$$\mu = 1 - \log\left(\sum_{i=1}^n \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\right)$$
$$q_i^* = \frac{\exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)}.$$

By injecting this expression, we get

$$\begin{aligned}\varphi(q^*) &= \frac{\sum_{i=1}^n \langle \lambda^*, g(X_i) - Pg\rangle \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)} - \log\left(\sum_{i=1}^n \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\right) \\ &= \frac{\langle \lambda^*, \sum_{i=1}^n (g(X_i) - Pg) \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\rangle}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)} - \log\left(\sum_{i=1}^n \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\right).\end{aligned}$$

Since $\sum_{i=1}^n q_i^* \left(g(X_i) - Pg\right) = 0$, we have

$$\frac{\sum_{i=1}^n (g(X_i) - Pg) \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^n \exp\left(\langle \lambda^*, g(X_j) - Pg\rangle\right)} = 0.$$

So

$$\varphi(q^*) = -\log\left(\sum_{i=1}^{n} \exp\left(\langle \lambda^*, g(X_i) - Pg\rangle\right)\right).$$

Set the following function $F : \mathbb{R}^m \to \mathbb{R}$ *defined by*

$$\forall \lambda \in \mathbb{R}^m,\ F(\lambda) = -\log\left(\sum_{i=1}^{n} \exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)\right).$$

By convex duality, this is equivalent to maximize the function F *on* $\mathbb{R}^m$*. This function is differentiable and concave. So* $\nabla F(\lambda) = 0$ *is equivalent to* $\lambda \in arg\max_{x\in\mathbb{R}^m} F(x)$*. In our case*

$$\nabla F(\lambda) = \frac{-\sum_{i=1}^{n}(g(X_i) - Pg)\exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^{n}\exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)}.$$

Thus $\nabla F(\lambda^*) = 0$*. It remains to prove that* F *is strictly concave. For all* $j, k \in [\![1, m]\!]$

$$\begin{aligned}\frac{\partial^2 F}{\partial\lambda_j\partial\lambda_k} &= \frac{-\sum_{i=1}^{n}(g_j(X_i) - Pg_j)(g_k(X_i) - Pg_k)\exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^{n}\exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)} \\ &\quad + \frac{\sum_{i=1}^{n}(g_j(X_i) - Pg_j)\exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^{n}\exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)}\frac{\sum_{i=1}^{n}(g_k(X_i) - Pg_k)\exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^{n}\exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)}.\end{aligned}$$

Denote for all $\lambda \in \mathbb{R}^m$ *and* $i \in [\![1, n]\!]$*,*

$$\begin{aligned} q_i(\lambda) &= \frac{\exp\left(\langle \lambda, g(X_i) - Pg\rangle\right)}{\sum_{j=1}^{n}\exp\left(\langle \lambda, g(X_j) - Pg\rangle\right)}, \\ Q_n(\lambda) &= \sum_{i=1}^{n} q_i(\lambda)\delta_{X_i}. \end{aligned}$$

Observe that for all $\lambda \in \mathbb{R}^m$*,* $Q_n(\lambda)$ *is a probability measure. The Hessian matrix of* F *is*

$$Hess_F = -Var_{Q_n(\lambda)}(g - Pg) = -Var_{Q_n(\lambda)}(g) < 0.$$

By using assumption (B.8), we deduce that the hessian matrix of F *is negative definite. So* F *is strictly concave.* □

Assume that $Pg = 0$ and denote again $\hat{\lambda}_n := \lambda^*$ to derive the following asymptotic result.

Theorem 5. *Assume that* $\mathbb{E}\|g(X)\|^2 < +\infty$ *and* $\Sigma = Pgg^T$ *is positive definite. We suppose that the assumptions of Theorem 4 are satisfied. Set* $\Sigma_n = P_n gg^T$*. Then we have*

- $\max_{1\le i\le n} | < \hat{\lambda}_n, g(X_i) > | = o_p(1)$.
- $\hat{\lambda}_n = -\Sigma_n^{-1}\mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right)$.
- $\sqrt{n}\hat{\lambda}_n \Rightarrow \mathcal{N}(0, \Sigma^{-1})$.

Remark 5. *Likewise we can replace* Σ_n *by the empirical variance of* g*.*

Proof. *Firstly, let us prove that* $\|\hat{\lambda}_n\| = O_p\left(\frac{1}{\sqrt{n}}\right)$. *Denote* $\hat{\lambda}_n = \rho_n\theta_n$ *with* $\rho_n \geq 0$, $\|\theta_n\| = 1$. *Set for all* $\lambda \in \mathbb{R}^m$

$$q_i(\lambda) = \frac{\exp\left(\langle\lambda, g(X_i)\rangle\right)}{\sum_{j=1}^n \exp\left(\langle\lambda, g(X_j)\rangle\right)},$$
$$\varphi(\lambda) = \sum_{i=1}^n q_i(\lambda) g(X_i).$$

By definition of $\hat{\lambda}_n$, *we have* $\varphi(\hat{\lambda}_n) = 0$. *So*

$$0 = \|\phi(\hat{\lambda}_n)\| = \|\phi(\rho_n\theta_n)\| = |\langle\theta_n, \varphi(\rho_n\theta_n)\rangle|.$$

By denoting $S_n(\lambda) = \sum_{j=1}^n \exp\left(\langle\lambda, g(X_j)\rangle\right) > 0$, *we have*

$$\frac{\theta_n^T}{S_n(\hat{\lambda}_n)} \sum_{i=1}^n \exp\left(\rho_n\theta_n^T g(X_i)\right) g(X_i) = 0.$$

By using the Taylor's theorem with Lagrange remainder (first order) we have that for all $i \in [|1, n|]$, *there exists* r_i *between* 0 *and* $\theta_n^T g(X_i)$ *such that*

$$\exp\left(\rho_n\theta_n^T g(X_i)\right) = 1 + \rho_n\theta_n^T g(X_i)\exp(\rho_n r_i).$$

Denote S^{m-1} *the unit sphere of* $\mathbb{R}^m$. *So*

$$\begin{aligned}
\frac{\theta_n^T}{S_n(\hat{\lambda}_n)} \sum_{i=1}^n \exp\left(\rho_n\theta_n^T g(X_i)\right) g(X_i) &= \frac{n\theta_n^T}{S_n(\hat{\lambda}_n)}\mathbb{P}_n g + \frac{n\rho_n}{S_n(\hat{\lambda}_n)}\frac{1}{n}\sum_{i=1}^n (\theta_n^T g(X_i))^2 \exp(\rho_n r_i) \\
&\geq \frac{n\theta_n^T}{S_n(\hat{\lambda}_n)}\mathbb{P}_n g + \frac{n\rho_n}{S_n(\hat{\lambda}_n)}\frac{1}{n}\sum_{i=1}^n (\theta_n^T g(X_i))^2 1_{\theta_n^T g(X_i)\geq 0} \\
&\geq \frac{n\theta_n^T}{S_n(\hat{\lambda}_n)}\mathbb{P}_n g + \frac{n\rho_n}{S_n(\hat{\lambda}_n)}\inf_{\theta\in S^{m-1}}\frac{1}{n}\sum_{i=1}^n (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0}.
\end{aligned}$$

since for all $i \in [\![1, n]\!]$ *if* $\theta_n^T g(X_i) \geq 0$ *then* $r_i \geq 0$. *Thus*

$$\frac{n\theta_n^T}{S_n(\hat{\lambda}_n)}\mathbb{P}_n g + \frac{n\rho_n}{S_n(\hat{\lambda}_n)}\inf_{\theta\in S^{m-1}}\frac{1}{n}\sum_{i=1}^n (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0} \leq 0.$$

By multiplying by $\frac{S_n(\hat{\lambda}_n)}{n}$ *and since* $\|\theta_n\| = 1$, *we have*

$$\rho_n \inf_{\theta\in S^{m-1}}\frac{1}{n}\sum_{i=1}^n (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0} \leq \|\mathbb{P}_n g\|.$$

It is enough to prove that $\inf_{\theta\in S^{m-1}}\frac{1}{n}\sum_{i=1}^n (\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0}$ *converges in probability to a strictly positive constant in order to deduce that* $\rho_n = O_p\left(\frac{1}{\sqrt{n}}\right)$. *For that, we need a technical lemma.*

Lemma 2. *We have*

$$\inf_{\theta\in S^{m-1}} \frac{1}{n}\sum_{i=1}^{n}(\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0} \xrightarrow[n\to\infty]{a.s.} \inf_{\theta\in S^{m-1}} P(\theta^T g)^2 1_{\theta^T g\geq 0} > 0.$$

Proof. *Set for all $\theta \in S^{m-1}$*

$$\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0}.$$

Firstly, let us prove that γ_n is continuous. Endow S^{m-1} with the subspace topology of $\mathbb{R}^m$. Define for all $i \in [\![1,n]\!]$

$$\begin{array}{rccc} h_i & : & S^{m-1} & \to & \mathbb{R} \\ & & \theta & \mapsto & \theta^T g(X_i). \end{array}$$

Set the following function $l : \mathbb{R} \to \mathbb{R}$ defined by $l(x) = x^2 1_{x\geq 0}$. Remark that for all $i \in [\![1,n]\!]$ the functions h_i and l are continuous and $\gamma_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l\circ h_i(\theta)$. So for all $n \in \mathbb{N}^*$, γ_n is continuous on S^{m-1}. Define $\psi : C^0(S^{m-1}) \to \mathbb{R}$ by $\psi(f) = \inf_{\|\theta\|=1} f(\theta)$. Let us prove that ψ is continuous. Indeed for all $f_1, f_2 \in C^0(S^{m-1})$

$$|\,\psi(f_1) - \psi(f_2)\,| = |\inf f_1 - \inf f_2| \leq \|f_1 - f_2\|_\infty.$$

So ψ is 1-Lipschitz. Notice that

$$\psi(\gamma_n) = \inf_{\theta\in S^{m-1}} \frac{1}{n}\sum_{i=1}^{n}(\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0}.$$

Prove that γ_n converge uniformly almost surely to γ where for all $\theta \in S^{m-1}$ $\gamma(\theta) = P(\theta^T g)^2 1_{\theta^T g\geq 0}$. For that, remark that for all $\theta \in S^{m-1}$

$$|\gamma_n(\theta) - \gamma(\theta)| = |(\mathbb{P}_n - P)(\theta^T g)^2 1_{\theta^T g\geq 0}|.$$

Define the following class of functions

$$\mathscr{F} = \left\{h_\theta : x \mapsto \theta^T g(x))^2 1_{\theta^T g(x)\geq 0}\,,\ \theta \in S^{m-1}\right\}.$$

Since the sphere S^{m-1} is compact, for all $x \in \mathscr{X}$ the map $\theta \mapsto h_\theta(x)$ is continuous and

$$\sup_{\theta\in S^{m-1}} |h_\theta| \leq \|g\|^2 \in L^1(P).$$

We deduce that the class $\mathscr{F}$ is P-Glivenko-Cantelli. So

$$\gamma_n \xrightarrow[n\to\infty]{\|.\|_\infty,\ a.s.} \gamma.$$

By continuity of ψ, we have $\psi(\gamma_n) \xrightarrow{a.s} \psi(\gamma)$. We conclude that

$$\inf_{\theta\in S^{m-1}} \frac{1}{n}\sum_{i=1}^{n}(\theta^T g(X_i))^2 1_{\theta^T g(X_i)\geq 0} \xrightarrow[n\to\infty]{a.s} \inf_{\theta\in S^{m-1}} P(\theta^T g)^2 1_{\theta^T g\geq 0}.$$

Finally we show that $\inf_{\theta\in S^{m-1}} P(\theta^T g)^2 1_{\theta^T g\geq 0} > 0$. *Since* $P\|g\|^2 < +\infty$, *the map* $\theta \mapsto P(\theta^T g)^2 1_{\theta^T g\geq 0}$ *is continuous. By compactness of the sphere, there exists* $\theta_* \in S^{m-1}$ *such that*

$$\inf_{\theta\in S^{m-1}} P(\theta^T g)^2 1_{\theta^T g\geq 0} = P(\theta_*^T g)^2 1_{\theta_*^T g\geq 0}.$$

If $P(\theta_*^T g)^2 1_{\theta_*^T g\geq 0} = 0$ *then* P*-a.s* x, $\theta_*^T g(x) \leq 0$. *In others words* $\mathbb{P}\left(\theta_*^T g(X) > 0\right) = 0$. *Remark that*

$$0 = \theta_*^T P g = P\theta_*^T g = P(\theta_*^T g) 1_{\theta_*^T g\leq 0}.$$

So P*-a.s* x, $\theta_*^T g(x) = 0$. *Since* Σ *is a matrix definite positive, we obtain a contradiction*

$$0 = P(\theta_*^T g)^2 = \theta_*^T P g g^T \theta_* = \theta_*^T \Sigma \theta_* > 0.$$

□

Since $\rho_n = \|\hat{\lambda}_n\|$, *we deduce by Lemma 2*

$$\|\hat{\lambda}_n\| = O_p\left(\frac{1}{\sqrt{n}}\right),$$
$$\max_{1\leq i\leq n} |\langle \hat{\lambda}_n, g(X_i)\rangle| \leq \|\hat{\lambda}_n\| \max_{1\leq i\leq n} \|g(X_i)\|.$$

Prove that $\max_{1\leq i\leq n} \|g(X_i)\| = o_p(\sqrt{n})$. *For that, we will use Owen's lemma (1990, [31]).*

Lemma 3. *Let* $(Y_n)_{n\in\mathbb{N}^*}$ *be a sequence i.i.d. of positive random variables and* $Z_n = \max_{1\leq i\leq n} Y_i$. *If* $\mathbb{E}Y^2 < +\infty$ *then*

$$Z_n = o_p(\sqrt{n}),$$
$$\frac{1}{n}\sum_{i=1}^n Y_i^3 = o_p(\sqrt{n}).$$

Since $\mathbb{E}\|g(X)\|^2 < +\infty$, *we deduce that by Owen's lemma that* $\max_{1\leq i\leq n} \|g(X_i)\| = o_p(\sqrt{n})$. *So we get the first assertion*

$$\max_{1\leq i\leq n} |\langle \hat{\lambda}_n, g(X_i)\rangle| = o_p(1).$$

About the second assertion, apply the Taylor's theorem with Lagrange remainder (second order). More precisely for all $i \in [\![1, n]\!]$, *there exists* s_i *between* 0 *and* $\theta_n^T g(X_i)$ *such that*

$$\begin{aligned}\exp\left(\rho_n \theta_n^T g(X_i)\right) &= 1 + \rho_n \theta_n^T g(X_i) + \rho_n^2 \left(\theta_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right)\\ &= 1 + \hat{\lambda}_n^T g(X_i) + \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right).\end{aligned}$$

Thus

$$\begin{aligned}0 &= \frac{1}{S_n(\hat{\lambda}_n)} \sum_{i=1}^n \exp\left(\hat{\lambda}_n^T g(X_i)\right) g(X_i)\\ &= \frac{1}{S_n(\hat{\lambda}_n)} \sum_{i=1}^n g(X_i) + \frac{1}{S_n(\hat{\lambda}_n)} \left(\sum_{i=1}^n g(X_i) g(X_i)^T\right) \hat{\lambda}_n + \frac{1}{S_n(\hat{\lambda}_n)} \sum_{i=1}^n \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i).\end{aligned}$$

By multiplying by $\frac{S_n(\hat{\lambda}_n)}{n}$*, we obtain*

$$0 = \mathbb{P}_n g + \Sigma_n \hat{\lambda}_n + \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i).$$

This is equivalent to

$$-\Sigma_n \hat{\lambda}_n = \mathbb{P}_n g + \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i).$$

Finally we prove that

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i) \right\| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

By Owen's lemma and the first assertion, we have for all $i \in [\![1, n]\!]$

$$\exp\left(\rho_n s_i\right) \leq \exp\left(\rho_n |\theta_n^T g(X_i)|\right) \leq \exp\left(\max_{1\leq i\leq n} |\hat{\lambda}_n^T g(X_i)|\right) = 1 + o_p(1), \tag{B.9}$$

$$\frac{1}{n}\sum_{i=1}^{n} \|g(X_i)\|^3 = o_p(\sqrt{n}), \tag{B.10}$$

$$\|\hat{\lambda}_n\|^2 = O_p\left(\frac{1}{n}\right). \tag{B.11}$$

Hence

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\lambda}_n^T g(X_i)\right)^2 \exp\left(\rho_n s_i\right) g(X_i) \right\| &\leq \exp\left(\max_{1\leq i\leq n} |\hat{\lambda}_n^T g(X_i)|\right) \|\hat{\lambda}_n\|^2 \frac{1}{n}\sum_{i=1}^{n} \|g(X_i)\|^3 \\ &\leq (1 + o_p(1)) O_p\left(\frac{1}{n}\right) o_p(\sqrt{n}) = o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Therefore

$$\hat{\lambda}_n = -\Sigma_n^{-1} \mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right).$$

We conclude that $\sqrt{n}\hat{\lambda}_n \Rightarrow \mathcal{N}(0, \Sigma^{-1})$. □

B.5 Towards a definition of the informed empirical measure

B.5.1 Equivalence of projections and the informed empirical measure

Assume that $Pg = 0$. In the previous section, we have studied these two optimization problems

$$arg \min_{Q \in \mathscr{P}^I(\mathcal{Z})} KL(\mathbb{P}_n || Q), \tag{B.12}$$

$$arg \min_{Q \in \mathscr{P}^I(\mathcal{Z})} KL(Q || \mathbb{P}_n). \tag{B.13}$$

In this section, we assume that the conditions of existence and uniqueness of the solution of (B.12) and (B.13) are satisfied – see Theorem 3 and 5. So by Theorem 2, the solution of the first optimization problem is given by

$$q_i^{(1)} = \frac{1}{n}\frac{1}{1+\langle \hat{\lambda}_n^{(1)}, g(X_i)\rangle},\ i \in [\![1,n]\!].$$

By Theorem 4, the solution of the second optimization problem is given by

$$q_i^{(2)} = \frac{\exp\left(\langle \hat{\lambda}_n^{(2)}, g(X_i)\rangle\right)}{S_n(\hat{\lambda}_n^{(2)})},\ i \in [\![1,n]\!],$$
$$S_n(\hat{\lambda}_n^{(2)}) = \sum_{k=1}^{n} \exp\left(\langle \hat{\lambda}_n^{(2)}, g(X_k)\rangle\right).$$

Denote these two probability measures, respectively,

$$\mathbb{P}_n^{(1)} = \sum_{i=1}^{n} q_i^{(1)} \delta_{X_i}, \tag{B.14}$$
$$\mathbb{P}_n^{(2)} = \sum_{i=1}^{n} q_i^{(2)} \delta_{X_i}. \tag{B.15}$$

Since these weights $q^{(1)}$ and $q^{(2)}$ are not explicit and we don't know if the submanifold $\mathscr{P}^I(\mathscr{Z})$ is autoparallel – see Theorem 1 – it is necessary to find an explicit approximation of these solutions.

Proposition 4. *Assume that* 0 *belongs to the convex hull of* $\{g(X_1), \cdots, g(X_n)\}$ *and that assumption (B.2) is satisfied. Moreover, suppose that* $\Sigma = Var_P g$ *is invertible. Denote* Σ_n *the empirical variance of* g*. Then for all* $i \in [\![1,n]\!]$

$$q_i^{(1)} = p_i + \varepsilon_{i,n}^{(1)},$$
$$q_i^{(2)} = p_i + \varepsilon_{i,n}^{(2)},$$

with $\varepsilon_n^{(1)}$ *and* $\varepsilon_n^{(2)}$ *independent of* i *and such that*

$$\max_{1\le i\le n} |\varepsilon_{i,n}^{(j)}| = o_p\left(\frac{1}{n}\right),\ j \in [\![1,2]\!],$$
$$p_i = \frac{1}{n}\left(1 - g(X_i)^T \Sigma_n^{-1} \mathbb{P}_n g + (\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g\right),\ i \in [\![1,n]\!].$$

Moreover it holds

$$\sum_{i=1}^{n} p_i = 1,$$
$$\sum_{i=1}^{n} p_i g(X_i) = 0.$$

Proof. *First assume that $j = 1$. Then for all $i \in [\![1, n]\!]$*

$$q_i^{(1)} = \frac{1}{n} \frac{1}{1 + \langle \hat{\lambda}_n^{(1)}, g(X_i) \rangle}$$

and, by Theorem 3 we have

$$\max_{1 \le k \le n} |\langle \hat{\lambda}_n^{(1)}, g(X_k) \rangle| = o_p(1),$$
$$\hat{\lambda}_n^{(1)} = \Sigma_n^{-1} \mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Thus

$$q_i^{(1)} = \frac{1}{n}\left(1 - \langle \hat{\lambda}_n^{(1)}, g(X_i) \rangle + o_p(1)\right) = \frac{1}{n}\left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle - \langle o_p\left(\frac{1}{\sqrt{n}}\right), g(X_i) \rangle + o_p(1)\right).$$

Since $\max_{1 \le k \le n} \|g(X_k)\| = o_p(\sqrt{n})$ *– see Owen's lemma in [31] – we deduce that*

$$q_i^{(1)} = \frac{1}{n}\left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle + o_p(1)\right) = \frac{1}{n}\left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle\right) + o_p\left(\frac{1}{n}\right).$$

Moreover since $\frac{1}{n}(\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g = o_p\left(\frac{1}{n}\right)$ *we get*

$$q_i^{(1)} = p_i + o_p\left(\frac{1}{n}\right).$$

Now assume that $j = 2$. Then for all $i \in [|1, n|]$,

$$q_i^{(2)} = \frac{\exp\left(\langle \hat{\lambda}_n^{(2)}, g(X_i) \rangle\right)}{S_n(\hat{\lambda}_n^{(2)})},$$
$$\text{with } S_n(\hat{\lambda}_n^{(2)}) = \sum_{k=1}^{n} \exp\left(\langle \hat{\lambda}_n^{(2)}, g(X_k) \rangle\right).$$

Hence Theorem 5 implies

$$\max_{1 \le k \le n} |\langle \hat{\lambda}_n^{(2)}, g(X_k) \rangle| = o_p(1),$$
$$\hat{\lambda}_n^{(2)} = -\Sigma_n^{-1} \mathbb{P}_n g + o_p\left(\frac{1}{\sqrt{n}}\right),$$

and

$$\begin{aligned} q_i^{(2)} &= \frac{1}{n} \frac{n}{S_n(\hat{\lambda}_n^{(2)})} \left(1 + \langle \hat{\lambda}_n^{(2)}, g(X_i) \rangle + o_p(1)\right) \\ &= \frac{1}{n} \frac{n}{S_n(\hat{\lambda}_n^{(2)})} \left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle + \langle o_p\left(\frac{1}{\sqrt{n}}\right), g(X_i) \rangle + o_p(1)\right). \end{aligned}$$

We deduce that $q_i^{(2)} = \frac{1}{n} \frac{n}{S_n(\hat{\lambda}_n^{(2)})} \left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle + o_p(1)\right)$.

Let us prove that $\frac{n}{S_n(\hat{\lambda}_n^{(2)})} = 1 + o_p(1)$. *Since* $e^x \geq 1 + x$ *for all* $x \in \mathbb{R}$ *we have*

$$1 + \langle \hat{\lambda}_n^{(2)}, \mathbb{P}_n g \rangle \leq \frac{S_n(\hat{\lambda}_n^{(2)})}{n} \leq \exp\left(\max_{1 \leq k \leq n} |\langle \hat{\lambda}_n^{(2)}, g(X_k) \rangle| \right).$$

It ensues that $\frac{n}{S_n(\hat{\lambda}_n^{(2)})} = 1 + o_p(1)$ *and*

$$q_i^{(2)} = \frac{1}{n}(1 + o_p(1)) \left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle + o_p(1)\right) = \frac{1}{n}\left(1 - \langle \Sigma_n^{-1} \mathbb{P}_n g, g(X_i) \rangle + o_p(1)\right).$$

Likewise since $\frac{1}{n}(\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g = o_p\left(\frac{1}{n}\right)$ *it holds* $q_i^{(2)} = p_i + o_p\left(\frac{1}{n}\right)$.

Finally remark that

$$\sum_{i=1}^n p_i = 1 - (\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g + (\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g = 1,$$

$$\sum_{i=1}^n p_i g(X_i) = \mathbb{P}_n g - \mathbb{P}_n g g^T \Sigma_n^{-1} \mathbb{P}_n g + \mathbb{P}_n g (\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g = \mathbb{P}_n g - \mathbb{P}_n g = 0,$$

by using $\Sigma_n = \mathbb{P}_n g g^T - \mathbb{P}_n g (\mathbb{P}_n g)^T$. □

Proposition 4 allows to define the informed empirical measure as follows.

Definition 1. *Assume that* $\Sigma = Var_P g$ *is invertible. The informed empirical measure is defined to be*

$$\mathbb{P}_n^I := \sum_{i=1}^n p_i \delta_{X_i}$$

where, for all $i \in [\![1, n]\!]$

$$p_i = \frac{1}{n}\left(1 - g(X_i)^T \Sigma_n^{-1} \mathbb{P}_n g + (\mathbb{P}_n g)^T \Sigma_n^{-1} \mathbb{P}_n g\right).$$

Remark 6. $\mathbb{P}_n^I$ *is always defined as soon as* $\Sigma_n = Var_n g$ *is invertible.*

Observe that for any measurable function $f : \mathscr{X} \to \mathbb{R}$ it holds

$$\mathbb{P}_n^I f = \mathbb{P}_n f - cov_n(g, f)^T \Sigma_n^{-1} \mathbb{P}_n g.$$

It turns out that this measure coincides with the adaptive estimator of the measure with auxiliary information studied by M. Albertus [4] which is a particular case of the general principle of S. Tarima and D. Pavlov [37].

B.5.2 Weights of the informed empirical measure

Let illustrate the difference of these four measures $\mathbb{P}_n^{(1)}$, $\mathbb{P}_n^{(2)}$, $\mathbb{P}_n^I$ and $\mathbb{P}_n$ by comparing the distribution of weights between them – see Figure 1. To this aim we simulate $n = 500$ *i.i.d.* random variables with distribution $P = \mathcal{N}(0, 1)$ and we incorporate the auxiliary information I given by Pg with for all $x \in \mathbb{R}$, $g(x) = (x, x^2)^T$. Observe that the distribution of weights between these three measures $\mathbb{P}_n^{(1)}$, $\mathbb{P}_n^{(2)}$, $\mathbb{P}_n^I$ are very similar.

The following proposition states that under a moment condition, with probability one $\mathbb{P}_n^I$ is a probability measure for n sufficiently large.

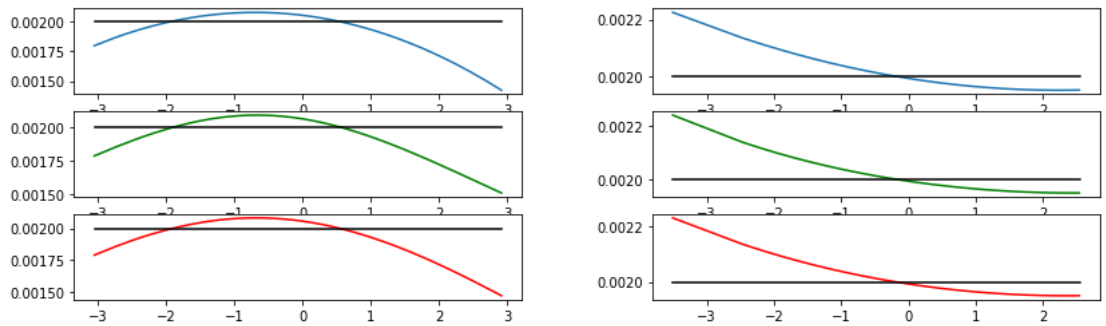

FIGURE B.1 – Comparison between $\mathbb{P}_n^I$ (blue), $\mathbb{P}_n^{(1)}$ (green), $\mathbb{P}_n^{(2)}$ (red) and $\mathbb{P}_n$ (black).

Proposition 5. *Assume that there exists $\varepsilon > 0$ such that $P\|g\|^{4+\varepsilon} < +\infty$. Then almost surely for n sufficiently large*

$$\mathbb{P}_n^I = \sum_{i=1}^n p_i \delta_{X_i}$$

is a probability measure.

Proof. *By Proposition 4, it is enough to prove that almost surely for all n sufficiently large it holds $\min_{1\le i\le n} p_i > 0$. For that, we need the following technical lemma that extends Owen's lemma 3.*

Lemma 4. *Let $(Y_n)_{n\in\mathbb{N}^*}$ be i.i.d. positive random variables and $Z_n = \max_{1\le i\le n} Y_i$. If $\mathbb{E}Y^s < +\infty$ with $s > 0$ then*

$$Z_n = o_{a.s.}(n^{\frac{1}{s}}),$$
$$\frac{1}{n}\sum_{i=1}^n Y_i^{s+1} = o_{a.s.}(n^{\frac{1}{s}}).$$

Proof. *Since $\mathbb{E}Y^s < +\infty$, we have*

$$\sum_{n\ge1} \mathbb{P}(Y_n > n^{1/s}) = \sum_{n\ge1} \mathbb{P}(Y_1^s > n) < +\infty$$

and, by Borel-Cantelli lemma

$Y_n > n^{\frac{1}{s}}$ is a.s. satisfied for a finite set in $\mathbb{N}^*$. Therefore, for all $A > 0$, $Z_n > An^{\frac{1}{s}}$ is a.s. satisfied for a finite set in $\mathbb{N}^*$.

Since $\mathbb{N}^*$ is countable, we then a.s. have, for all $m \in \mathbb{N}^*$,

$$0 \le \liminf_n \frac{Z_n}{n^{\frac{1}{s}}} \le \limsup_n \frac{Z_n}{n^{\frac{1}{s}}} \le \frac{1}{m}$$

hence $Z_n = o_{a.s.}(n^{\frac{1}{s}})$.

Concerning the second assertion, observe that

$$0 \le \frac{1}{n}\sum_{i=1}^n Y_i^{s+1} \le \frac{Z_n}{n}\sum_{i=1}^n Y_i^s = o_{a.s.}(n^{\frac{1}{s}})$$

by the strong law of large numbers. □

Since $P\|g\|^{4+\varepsilon} < +\infty$ by assumption, the previous lemma yields

$$\frac{\max_{1\le i\le n}\|g(X_i)\|}{n^{\frac{1}{4+\varepsilon}}} \xrightarrow[n\to\infty]{a.s.} 0.$$

Remark that

$$\begin{aligned}\min_{1\le i\le n} np_i &= 1 - \max_{1\le i\le n} g(X_i)^T\Sigma_n^{-1}\mathbb{P}_n g + (\mathbb{P}_n g)^T\Sigma_n^{-1}\mathbb{P}_n g,\\ |\max_{1\le i\le n} g(X_i)^T\Sigma_n^{-1}\mathbb{P}_n g| &\le \max_{1\le i\le n} |g(X_i)^T\Sigma_n^{-1}\mathbb{P}_n g|\\ &\le \max_{1\le i\le n}\|g(X_i)\|\|\Sigma_n^{-1}\|\|\mathbb{P}_n g\|.\end{aligned}$$

It suffices to prove that

$$n^{\frac{1}{4+\varepsilon}}\|\mathbb{P}_n g\| \xrightarrow[n\to\infty]{a.s.} 0.$$

For that, we need the following Theorem of Wellner and Van der Vaart in chapter 2.5 [38].

Lemma 5. *Let $\mathscr{F}$ be a P-measurable class with envelope function F. Then for all $p \ge 1$, there exists $c_p > 0$ such that,*

$$\|\|\alpha_n\|^*_{\mathscr{F}}\|_{L^p(P)} \le c_p J(1,\mathscr{F})\|F\|_{L^{2\vee p}(P)}$$

where

$$J(1,\mathscr{F}) = \sup_Q \int_0^1 \sqrt{1+\log N\left(\epsilon\|F\|_{L^2(Q)},\mathscr{F},L^2(Q)\right)}\, d\epsilon.$$

and $N\left(\epsilon\|F\|_{L^2(Q)},\mathscr{F},L^2(Q)\right)$ is the minimal numbers of balls of radius $\epsilon\|F\|_{Q,2}$ needed to cover $\mathscr{F}$ in $L^2(Q)$. Here the supremum is taken over all finitely discrete probability measure on $(\mathscr{X},\mathscr{A})$.

Set for all $j \in [\![1,m]\!]$, $\mathscr{F}_j = \{g_j\}$ and denote $s = 4+\varepsilon$. Remark that $J(1,\mathscr{F}_j) = 1$. Then

$$\mathbb{P}(n^{\frac{1}{s}}\|\mathbb{P}_n g\| > \epsilon) \le \sum_{j=1}^m \mathbb{P}(n^{\frac{1}{s}} \mid \mathbb{P}_n g_j \mid > \epsilon) = \sum_{j=1}^m \mathbb{P}\left(\mid \alpha_n(g_j)\mid^s > \left(\frac{\sqrt{n}}{n^{\frac{1}{s}}}\epsilon\right)^s\right).$$

So

$$\mathbb{P}(n^{\frac{1}{s}}\|\mathbb{P}_n g\| > \epsilon) \le \sum_{j=1}^m \frac{\mathbb{E}(\mid \alpha_n(g_j)\mid^s)}{n^{\alpha(1/2-1/s)}\epsilon^s} \le c_s \sum_{j=1}^m \frac{\|g_j\|^s_{L^s(P)}}{n^{s\frac{s-2}{2s}}\epsilon^s}.$$

Since $s > 4$, we get $s\frac{s-2}{2s} > 1$ and, by Borel-Cantelli's lemma, $n^{\frac{1}{s}}\|\mathbb{P}_n g\| \xrightarrow[n\to\infty]{a.s.} 0$. We conclude that

$$n^{\frac{1}{4+\varepsilon}}\|\mathbb{P}_n g\| \xrightarrow[n\to\infty]{a.s.} 0.$$

□

The weights $(p_i)_{1\le i\le n}$ are given for all $i \in [\![1,n]\!]$

$$p_i = \frac{1}{n}\left(1 - g(X_i)^T \Sigma_n^{-1}\mathbb{P}_n g + (\mathbb{P}_n g)^T \Sigma_n^{-1}\mathbb{P}_n g\right).$$

Set $A_n = -\frac{1}{n}\Sigma_n^{-1}\mathbb{P}_n g$ and $B_n = \frac{1}{n}\left(1 + \langle \Sigma_n^{-1}\mathbb{P}_n g, \mathbb{P}_n g\rangle\right)$. Define for all $(x,y) \in \mathscr{X} \times \mathbb{R}^m$

$$\psi_n(y) = \langle A_n, y\rangle + B_n,$$
$$\varphi_n(x) = \langle A_n, g(x)\rangle + B_n = \psi_n \circ g(x).$$

Observe that for all $i \in [\![1,n]\!]$, $p_i = \varphi_n(X_i)$ and φ_n is an affine transformation of g. Suppose that $m = 1$ and g is monotone. The weights associated to the ordered sample are also ordered. Illustrate this by a simulation. We generate $n = 500$ random variables *i.i.d.* with distribution $P = \mathcal{N}(0,1)$. In the left hand (resp. right hand) graph, we plot φ_n with I given by $g(x) = x$ (resp. $g(x) = x^2$) for all $x \in \mathbb{R}$.

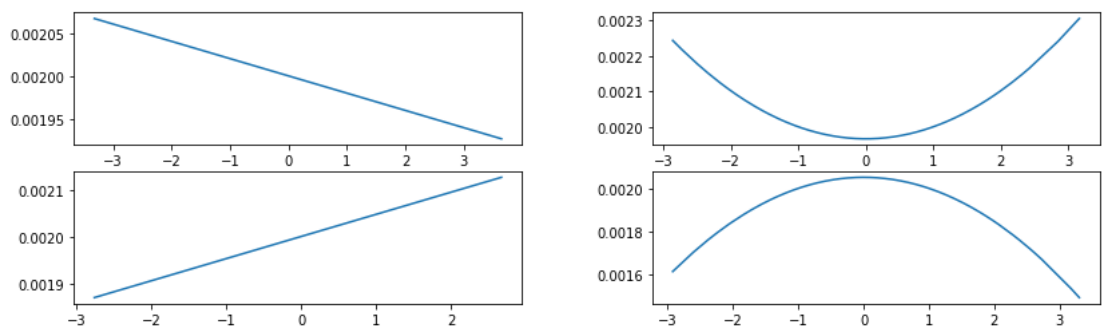

FIGURE B.2 – The sign of A_n changes randomly.

B.6 Asymptotic results and concentration

B.6.1 P-Glivenko-Cantelli and P-Donsker properties under minimal assumptions

Remind that $\mathbb{P}_n^I$ is the informed empirical measure of Definition 1 and write $\alpha_n^I = \sqrt{n}\left(\mathbb{P}_n^I - P\right)$ the informed empirical process. In this section we derive asymptotic results for $\mathbb{P}_n^I$ under minimal assumptions.

Given a class of functions $\mathscr{F}$, if $\|\mathbb{P}_n^I - P\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} |\mathbb{P}_n^I f - Pf|$ is not measurable, its minimal measurable majorant $\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^*$ is used – as well as the outer probability $\mathbb{P}^*$, see Chapter 1.2 in [38]. Denote $\|\cdot\|$ the euclidean norm on $\mathbb{R}^m$. Let $A \in \mathscr{M}_m(\mathbb{R})$ be a matrix, we denote $\|A\|$ the operator norm with respect to the euclidean norm.

The following theorem states that a P-Glivenko Cantelli class for $\mathbb{P}_n$ is also a P-Glivenko Cantelli for $\mathbb{P}_n^I$ as soon as the envelope function $F \in L^2(P)$.

Theorem 6. *Assume that $\Sigma = Var_P g$ is invertible. Let $\mathscr{F}$ be a P-Glivenko-Cantelli class with measurable envelope function $F \in L^2(P)$. Then we have*

$$\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^* \xrightarrow[n\to\infty]{a.s.} 0.$$

Proof. *Let $f \in \mathscr{F}$. Remark that*

$$\|cov_n(g,f)^T\Sigma_n^{-1}\| \le \sqrt{m}\sqrt{Var_n f}\|\Sigma_n^{-1}\| \max_{1\le i\le m}\sqrt{Var_n g_i} \le \sqrt{m}\|F\|_{L^2(\mathbb{P}_n)}\|\Sigma_n^{-1}\| \max_{1\le i\le m}\sqrt{Var_n g_i}.$$

Hence

$$\|\mathbb{P}_n^I - P\|_{\mathscr{F}}^* \le \|\mathbb{P}_n - P\|_{\mathscr{F}}^* + \|\mathbb{P}_n g\| \sup_{f\in\mathscr{F}} \|cov_n(g,f)^T\Sigma_n^{-1}\| \xrightarrow[n\to\infty]{a.s.} 0.$$

□

We have a similar result for the P-Donsker classes.

Theorem 7. *Assume that $\Sigma = Var_P g$ is invertible. Let $\mathscr{F}$ be a P-Donsker class with measurable envelope function $F \in L^2(P)$. Then*

$$\alpha_n^I \Rightarrow G_I \ \ in \ \ l^\infty(\mathscr{F})$$

where G is a P-Brownian bridge having almost surely continuous sample paths with respect to the semimetric $\rho^2(h_1,h_2) = Var_P(h_1 - h_2)$ for $h_1, h_2 \in L^2(P)$ and, for all $f \in \mathscr{F}$

$$G_I(f) = G(f) - cov_P(g,f)^T\Sigma^{-1}G(g) \tag{B.16}$$

with $G(g) = (G(g_1),\cdots,G(g_m))^T$.

Moreover for all $f \in \mathscr{F}$, $Var(G_I(f)) = Var_P f - cov_P(g,f)^T\Sigma^{-1}cov_P(g,f)$ and

$$Var(G_I(f)) \le Var(G(f)).$$

Remark 7. *Observe that G_I is a mean-zero Gaussian process such that for all $f \in \mathscr{F}$, $Var(G_I(f)) < Var(G(f))$ provided that $cov_P(g,f) \neq 0$. In other words, any function linearly correlated to g benefits of the information I.*

Proof. *First remark that*

$$\alpha_n^I(f) = \alpha_n(f) - cov_P(g,f)^T\Sigma^{-1}\alpha_n(g) + o_p(1).$$

Indeed

$$\begin{aligned}(cov_n(g,f)^T\Sigma_n^{-1} - cov(g,f)^T\Sigma^{-1})\alpha_n(g) &= (cov_n(g,f)^T - cov_P(g,f)^T)\Sigma_n^{-1}\alpha_n(g)\\ &\quad + cov_P(g,f)^T(\Sigma_n^{-1} - \Sigma^{-1})\alpha_n(g).\end{aligned}$$

The second term tends to 0 in probability,

$$\begin{aligned}|cov_P(g,f)^T(\Sigma_n^{-1} - \Sigma^{-1})\alpha_n(g)| &\le \sqrt{m}\sqrt{Var_P f}\max_{1\le i\le m}\sqrt{Var_P g_i}\|\Sigma_n^{-1} - \Sigma^{-1}\|\|\alpha_n(g)\|\\ &\le \sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\|F\|_{L^2(P)}\|\Sigma_n^{-1} - \Sigma^{-1}\|\|\alpha_n(g)\| = o_p(1).\end{aligned}$$

For the first term observe that

$$\begin{aligned}|(cov_n(g,f)^T - cov_P(g,f)^T)\Sigma_n^{-1}\alpha_n(g)| &\leq \|\mathbb{P}_n fg - Pfg - \mathbb{P}_n f\mathbb{P}_n g\|\|\Sigma_n^{-1}\|\|\alpha_n(g)\| \\ &\leq \sqrt{m}\max_{1\leq j\leq m}\|\mathbb{P}_n - P\|^*_{\mathscr{F}_j}\|\Sigma_n^{-1}\|\|\alpha_n(g)\| \\ &+ |\mathbb{P}_n F|\|\mathbb{P}_n g\|\|\Sigma_n^{-1}\|\|\alpha_n(g)\|\end{aligned}$$

where for all $j \in [\![1,m]\!]$, $\mathscr{F}_j = \{g_j f, f \in \mathscr{F}\}$.

Since g is integrable with respect to P then for all $j \in [\![1,m]\!]$, $\{g_j\}$ is P-Glivenko-Cantelli. Moreover $\mathscr{F}$ is P-Glivenko-Cantelli since $\mathscr{F}$ is P-Donsker. Set for all $(x,y) \in \mathbb{R}^2$, $\phi(x,y) = xy$. Remark that $\mathscr{F}_j = \phi(\{g_j\},\mathscr{F})$. Since ϕ is continuous, we deduce that $j \in [\![1,m]\!]$, $\mathscr{F}_j$ is a P-Glivenko-Cantelli Class. So

$$\max_{1\leq j\leq m}\|\mathbb{P}_n - P\|^*_{\mathscr{F}_j} = o_{a.s.}(1).$$

Thus

$$\sup_{f\in\mathscr{F}}|(cov_n(g,f)^T\Sigma_n^{-1} - cov_P(g,f)^T\Sigma^{-1})\alpha_n(g)| = o_p(1).$$

Denote for all $f \in \mathscr{F}$

$$W_n(f) = \alpha_n(f) - cov_P(g,f)^T\Sigma^{-1}\alpha_n(g).$$

It remains to prove that W_n converges in distribution. Remark that by the central limit theorem (CLT)

$$\begin{pmatrix}\alpha_n(f)\\ \alpha_n(g)\end{pmatrix} \Rightarrow \begin{pmatrix}G(f)\\ G(g)\end{pmatrix},\ f \in \mathscr{F}$$

where G is a P-Brownian bridge. Define, for $(f,a,b) \in \mathscr{F}\times\mathbb{R}^2$, $\psi_f(a,b) = (a, cov_P(g,f)^T\Sigma^{-1}b)$. Since the map ψ_f is continuous on $\mathbb{R}^2$, by the continuous mapping theorem we have

$$\widetilde{\mathbb{Y}}_n(f) = \begin{pmatrix}\alpha_n(f)\\ cov_P(g,f)^T\Sigma^{-1}\alpha_n(g)\end{pmatrix} \Rightarrow \widetilde{\mathbb{Y}}(f) = \begin{pmatrix}G(f)\\ cov_P(g,f)^T\Sigma^{-1}G(g)\end{pmatrix},\ f \in \mathscr{F}.$$

Likewise for all $f = (f_1,\cdots,f_k) \in \mathscr{F}^k$, $\left(\widetilde{\mathbb{Y}}_n(f_1),\cdots,\widetilde{\mathbb{Y}}_n(f_k)\right)$ converges in distribution to $\left(\widetilde{\mathbb{Y}}(f_1),\cdots,\widetilde{\mathbb{Y}}(f_k)\right)$. Indeed it suffices to apply the CLT and to consider the following continuous map

$$\tilde{\psi}_f(a_1,...,a_{k+1}) = \left(\psi_{f_1}(a_1,a_{k+1}),...,\psi_{f_k}(a_k,a_{k+1})\right),\ (a_1,...,a_{k+1}) \in \mathbb{R}^{k+1}.$$

Applying once again the continuous mapping theorem we deduce

$$\left(W_n(f_1),\cdots,W_n(f_k)\right)^T \Rightarrow \left(G(f_1) - cov_P(g,f_1)^T\Sigma^{-1}G(g),\cdots,G(f_k) - cov_P(g,f_k)^T\Sigma^{-1}G(g)\right)^T.$$

Let us prove that $W_n \Rightarrow G_I$ in $l^\infty(\mathscr{F})$ with G_I defined at (B.16). To this aim, we need the following theorem – see chapter 1.5 in [38].

Lemma 6. *Let $X_n : \Omega_n \to l^\infty(T)$ a sequence of maps. Then these two assertions are equivalents*

- *We have*
  1. *For all $(t_1,...,t_k) \in T^k$, $(X_n(t_1),...,X_n(t_k))$ converges weakly to a $\mathbb{R}^k$ valued random vector, for all $k \in \mathbb{N}^*$,*

2. *There exists a semimetric ρ such that (T,ρ) is totally bounded and for all $\epsilon > 0$*

$$\lim_{\delta\to 0}\limsup_{n} \mathbb{P}^*\left(\sup_{\rho(s,t)<\delta} |X_n(s)-X_n(t)| > \epsilon\right) = 0.$$

- *There exists $X:\Omega \to l^\infty(T)$ a measurable and tight process such that*

$$X_n \Rightarrow X \ \ in\ l^\infty(T).$$

First verify that a.s $W_n \in l^\infty(\mathscr{F})$

$$\begin{aligned}\|W_n\|_{\mathscr{F}} &\le 2\max\left(\|\alpha_n\|_{\mathscr{F}}, \sup_{f\in\mathscr{F}} |(Pfg)^T\Sigma^{-1}\alpha_n(g)|\right)\\ &\le 2\max\left(\|\alpha_n\|_{\mathscr{F}}, \sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\|F\|_{L^2(P)}\|\Sigma^{-1}\|\|\alpha_n(g)\|\right) < +\infty.\end{aligned}$$

In order to check the second point let introduce the usual semimetric in $L^2(P)$ defined by $\rho^2(f_1,f_2) = Var_P\left(f_1-f_2\right)$ for all $f_1,f_2\in L^2(P)$. Since $\mathscr{F}$ is P-Donsker then $(\mathscr{F},\rho)$ is totally bounded. For $\varepsilon>0$ and $\delta>0$, we have

$$\begin{aligned}&\limsup_{n} \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\,W_n(f_1)-W_n(f_2)\,| > 2\epsilon\right)\\ &\le \limsup_{n} \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} \max\left(|\alpha_n(f_1)-\alpha_n(f_2)|, |(P(f_1-f_2)g)^T\Sigma^{-1}\alpha_n(g))|\right) > \epsilon\right)\\ &\le \limsup_{n} \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1)-\alpha_n(f_2)| > \epsilon\right)\\ &+\limsup_{n} \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} \sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\,\rho(f_1,f_2)\|\Sigma^{-1}\|\|\alpha_n(g)\| > \epsilon\right)\\ &\le \limsup_{n} \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1)-\alpha_n(f_2)| > \epsilon\right)\\ &+\limsup_{n} \mathbb{P}^*\left(\delta\sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|\alpha_n(g)\| > \epsilon\right).\end{aligned}$$

Since $\mathscr{F}$ is P-Donsker, we have

$$\lim_{\delta\to 0}\limsup_{n} \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta} |\alpha_n(f_1)-\alpha_n(f_2)| > \epsilon\right) = 0.$$

By using the fact that

$$\|\alpha_n(g)\| \Rightarrow \|G(g)\|,$$

we have by the porte-manteau lemma

$$\begin{aligned}
&\limsup_n \mathbb{P}^*\left(\delta\sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|\alpha_n(g)>\epsilon\right)\\
&\le \limsup_n \mathbb{P}^*\left(\delta\sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|\alpha_n(g)\ge\epsilon\right)\\
&\le \mathbb{P}^*\left(\delta\sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|G(g)\|\ge\epsilon\right).
\end{aligned}$$

Since $\delta\sqrt{m}\max_{1\le i\le m}\sqrt{Var_P g_i}\|\Sigma^{-1}\|\|G(g)\| = o_p(1)$, *we get*

$$\lim_{\delta\to 0}\limsup_n \mathbb{P}^*\left(\sup_{\rho(f_1,f_2)<\delta}|\,W_n(f_1)-W_n(f_2)\,|>2\epsilon\right)=0.$$

This establishes that $W_n \Rightarrow G_I$ *in* $l^\infty(\mathscr{F})$. *Thus*

$$\alpha_n^I \Rightarrow G_I \;\; in \;\; l^\infty(\mathscr{F}).$$

Finally we compute the variance of $G_I(f)$ *for* $f\in\mathscr{F}$,

$$\begin{aligned}
Var\left(G_I(f)\right) &= Var\left(G(f)-cov_P(g,f)^T\Sigma^{-1}G(g)\right)\\
&= Var\left(G(f)\right)+Var\left(cov_P(g,f)^T\Sigma^{-1}G(g)\right)-2cov\left(G(f),cov_P(g,f)^T\Sigma^{-1}G(g)\right)\\
&= Var_P f+cov_P(g,f)^T\Sigma^{-1}\left(Var_P g\right)\Sigma^{-1}cov_P(g,f)-2cov_P(g,f)^T\Sigma^{-1}cov_P(g,f)\\
&= Var_P f-cov_P(g,f)^T\Sigma^{-1}cov_P(g,f).
\end{aligned}$$

Since $\Sigma^{-1}>0$, *we deduce that* $cov_P(g,f)^T\Sigma^{-1}cov_P(g,f)\ge 0$. *Thus*

$$Var(G_I(f))\le Var(G(f)).$$

□

B.6.2 Concentration of the informed empirical process

Next we show that the informed empirical process is more concentrated than the classical empirical process for all n sufficiently large. Moreover, we prove that the supremum of limit process G_I of Theorem 7 on a P-Donsker class is more concentrated than the supremum of G. By Theorem 7, we have for all $f\in L^2(P)$

$$Var\left(G_I(f)\right)\le Var\left(G(f)\right).$$

A first consequence is that for all $f\in L^2(P)$, the informed empirical process $\alpha_n^I(f)$ is more concentrated than $\alpha_n(f)$.

Proposition 6. *Assume that* $\Sigma = Var_P g$ *is invertible and let* $f\in L^2(P)$. *Then for any* $\lambda>0$

$$\mathbb{P}\left(|G_I(f)|>\lambda\right)\le\mathbb{P}\left(|G(f)|>\lambda\right),$$

and, if moreover $cov_P(g,f) \neq 0$ *then*

$$\mathbb{P}\left(|G_I(f)| > \lambda\right) < \mathbb{P}\left(|G(f)| > \lambda\right).$$

In addition there exists $N > 0$ *such that for all* $n \geq N$ *it holds*

$$\mathbb{P}\left(|\alpha_n^I(f)| > \lambda\right) < \mathbb{P}\left(|\alpha_n(f)| > \lambda\right).$$

Proof. *Let* $\lambda > 0$. *By Theorem 7 and Central Limit Theorem (CLT) we have*

$$\mathbb{P}\left(|\alpha_n^I(f)| > \lambda\right) \xrightarrow[n\to\infty]{} \mathbb{P}\left(|G_I(f)| > \lambda\right),$$
$$\mathbb{P}\left(|\alpha_n(f)| > \lambda\right) \xrightarrow[n\to\infty]{} \mathbb{P}\left(|G(f)| > \lambda\right).$$

Denote $\sigma_1 := \sqrt{VarG_I(f)} < \sqrt{VarG(f)} =: \sigma_2$. *Observe that*

$$\mathbb{P}\left(|G_I(f)| > \lambda\right) = \mathbb{P}\left(|\mathcal{N}(0,1)| > \frac{\lambda}{\sigma_1}\right) < \mathbb{P}\left(|\mathcal{N}(0,1)| > \frac{\lambda}{\sigma_2}\right) = \mathbb{P}\left(|G(f)| > \lambda\right).$$

So there exists $N > 0$ *such that for all* $n \geq N$

$$\mathbb{P}\left(|\alpha_n^I(f)| > \lambda\right) < \mathbb{P}\left(|\alpha_n(f)| > \lambda\right).$$

□

The next step is to extend this result to a P-Donsker class $\mathcal{F}$. Recall that $\mathcal{F}$ is a pointwise separable class – see [38] for more details – if there exists a countable subset $\mathcal{G} \subset \mathcal{F}$ such that for each $n \in \mathbb{N}^*$ there is a P^n-null set $N_n \subset \mathcal{X}^n$ such that for all $(x_1, \cdots, x_n) \notin N_n$ and $f \in \mathcal{F}$, there exists a sequence $(h_k)_{k\in\mathbb{N}^*} \subset \mathcal{G}$ such that $h_k \xrightarrow[k\to\infty]{} f$ in $L^2(P)$ and $(h_k(x_1), \cdots, h_k(x_n)) \xrightarrow[k\to\infty]{} (f(x_1), \cdots, f(x_n))$.

Proposition 7. *Assume that* $\Sigma = Var_P g$ *is invertible and let* $\mathcal{F}$ *be a* P*-Donsker class. Suppose that* $\mathcal{F}$ *is pointwise separable. Then for all* $\lambda > 0$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |G_I(f)| > \lambda\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} |G(f)| > \lambda\right).$$

Moreover if there exists $\lambda > 0$ *such that* $\mathbb{P}\left(\sup_{f\in\mathcal{F}} |G_I(f)| > \lambda\right) < \mathbb{P}\left(\sup_{f\in\mathcal{F}} |G(f)| > \lambda\right)$ *then there exists* $N > 0$ *such that for all* $n \geq N$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\alpha_{n,I}(f)| > \lambda\right) < \mathbb{P}\left(\sup_{f\in\mathcal{F}} |\alpha_n(f)| > \lambda\right).$$

Proof. *To prove this proposition, we need the following result.*

Lemma 7. *(Slepian, Fernique, Marcus, Shepp)*

Let X *and* Y *be separable, mean-zero Gaussian processes indexed by a common index set* T *such that*

$$\mathbb{E}(X_s - X_t)^2 \leq \mathbb{E}(Y_s - Y_t)^2 \quad \textit{for all } s, t \in T.$$

Then for all $\lambda > 0$

$$\mathbb{P}\left(\sup_{t\in T} X_t > \lambda\right) \leq \mathbb{P}\left(\sup_{t\in T} Y_t > \lambda\right).$$

Notice that G and G_I are two mean-zero Gaussian processes and we can take a separable version of these processes. Set $\widetilde{\mathscr{F}} = \mathscr{F} \cup (-\mathscr{F})$ and observe that

$$\sup_{f\in\mathscr{F}} |G_I(f)| = \sup_{f\in\mathscr{F}} \max\big(G_I(f), -G_I(f)\big) = \sup_{f\in\mathscr{F}} \max\big(G_I(f), G_I(-f)\big) = \sup_{f\in\widetilde{\mathscr{F}}} G_I(f).$$

Similarly $\sup_{f\in\mathscr{F}} |G(f)| = \sup_{f\in\widetilde{\mathscr{F}}} G(f)$. *Remark that for all $f, h \in \widetilde{\mathscr{F}}$*

$$\mathbb{E}\big(G_I(f) - G_I(h))\big)^2 = \mathbb{E}\big(G_I(f-h))\big)^2 = Var\big(G_I(f-h)\big) \le Var\big(G(f-h)\big) = \mathbb{E}\big(G(f) - G(h))\big)^2.$$

By Lemma 7

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} |G_I(f)| > \lambda\right) = \mathbb{P}\left(\sup_{f\in\widetilde{\mathscr{F}}} G_I(f) > \lambda\right) \le \mathbb{P}\left(\sup_{f\in\widetilde{\mathscr{F}}} G(f) > \lambda\right) = \mathbb{P}\left(\sup_{f\in\mathscr{F}} |G(f)| > \lambda\right).$$

Assume that there exists $\lambda > 0$ such that

$$\mathbb{P}\left(\sup_{f\in\widetilde{\mathscr{F}}} G_I(f) > \lambda\right) < \mathbb{P}\left(\sup_{f\in\widetilde{\mathscr{F}}} G(f) > \lambda\right).$$

Observe that the map $X \mapsto \|X\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} |X(f)|$ for every $X \in l^\infty(\mathscr{F})$ is continuous. Since $\mathscr{F}$ is pointwise separable, there is no problem with measurability. By Theorem 7, we deduce that there exists $N > 0$ such that for all $n \ge N$

$$\mathbb{P}\left(\|\alpha_{n,I}\|_{\mathscr{F}} > \lambda\right) = \mathbb{P}\left(\sup_{f\in\mathscr{F}} |\alpha_{n,I}(f)| > \lambda\right) < \mathbb{P}\left(\sup_{f\in\mathscr{F}} |\alpha_n(f)| > \lambda\right) = \mathbb{P}\left(\|\alpha_n\|_{\mathscr{F}} > \lambda\right).$$

□

As a corollary we immediately get that the quantiles of $\sup_{f\in\mathscr{F}} |G(f)|$ and $\sup_{f\in\mathscr{F}} |G_I(f)|$ are also ordered. Denote F_1 (resp. F_2) the cumulative distribution function of $\sup_{f\in\mathscr{F}} |G(f)|$ (resp. $\sup_{f\in\mathscr{F}} |G_I(f)|$) and, for $i \in [\![1,2]\!]$, $F_i^{-1}(\alpha) = \inf\{t \in \mathbb{R}_+,\ F_i(t) \ge \alpha\}$.

Corollary 2. *Assume that $\Sigma = Var_P g$ is invertible and let $\mathscr{F}$ be a P-Donsker class. Suppose that $\mathscr{F}$ is pointwise separable. Then, for all $\alpha \in [0,1[$ it holds $F_2^{-1}(\alpha) \le F_1^{-1}(\alpha)$.*

This corollary has many interesting applications. For instance, it can be used in order to improve Kolmogorov-Smirnov test – see [4] at the page 34 for an auxiliary information given by a partition of $\mathscr{X}$.

B.7 Informed empirical quantiles

As an illustration let consider the informed estimator of a single quantile built from $\mathbb{P}_n^I$ in the case P is a real probability measure. Standard empirical process methods could be applied to extend this estimation uniformly on compact sets of $]0,1[$ – and on $]0,1[$ under additional assumptions on the regularity and rate of decay of the density in tails. We would obtain similar results and same

limiting process as in [42] where the quantile process based on the probability measure $\mathbb{P}_n^{(1)}$ of (B.14) is shown to converge in the appropriated topology. Proposition 4 suggests that similar results are also valid for the quantiles derived from the unusual $\mathbb{P}_n^{(2)}$.

Denote F the cumulative distribution function of P and $\mathbb{F}_n$ the empirical distribution function. Assume that $Pg = 0$. We can define for all $t \in \mathbb{R}$ the following function

$$\mathbb{F}_{n,I}(t) = \mathbb{P}_n^I 1_{\cdot \leq t} = \sum_{i=1}^{n} p_i 1_{X_i \leq t}.$$

This function is called the informed empirical distribution function since $\mathbb{P}_n^I$ is a probability measure almost surely for n sufficiently large – see Proposition 5.

Let us first compare $\mathbb{F}_{n,I}$, $\mathbb{F}_n$ and F through a simulation. For that, we generate $n = 100$ random variables *i.i.d.* with distribution $P = \mathcal{N}(0,1)$ and we incorporate the auxiliary information I given by the function $g(x) = (x, x^2)$ defined for all $x \in \mathbb{R}$. We remark that $\mathbb{F}_{n,I}$ is closer of F than $\mathbb{F}_n$.

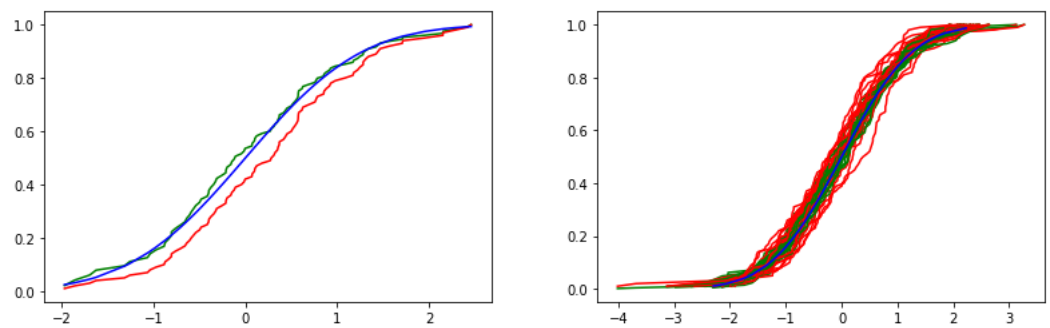

FIGURE B.3 – Comparison between $\mathbb{F}_{n,I}$ (green), $\mathbb{F}_n$ (red) and F (blue). The number of experiments is 1 in the left hand graph and 20 in the right hand graph.

Given $\alpha \in]0,1[$, we estimate the α-quantile $q_\alpha = F^{-1}(\alpha) = \inf\{t \in \mathbb{R},\ F(t) \geq \alpha\}$ of P by

$$q_{n,\alpha}^I = \mathbb{F}_{n,I}^{-1}(\alpha) = \inf\left\{t \in \mathbb{R},\ \mathbb{F}_{n,I}(t) \geq \alpha\right\}.$$

Clearly $q_{n,\alpha}^I$ is well defined since for all $t \in]-\infty, \min_{1\leq i\leq n} X_i[$, $F_{n,I}(t) = 0$ and for all $t \in [\max_{1\leq i\leq n} X_i, +\infty[$, $F_{n,I}(t) = 1$. The estimator $q_{n,\alpha}^I$ is called the informed empirical α-quantile.

In order to asymptotically control $q_{n,\alpha}^I$ let apply the classical delta method.

Lemma 8. *Let $\mathcal{S} = D(\mathbb{R}) \cap l^\infty(\mathbb{R})$ be the set of bounded càdlàg function defined on $\mathbb{R}$ equipped with the uniform norm. Let $\alpha \in]0,1[$ and $\phi_\alpha : D_\alpha \subset \mathcal{S} \to \mathbb{R}$ be the map defined by $\phi_\alpha(F) = F^{-1}(\alpha)$ for all $F \in D_\alpha$ with D_α a domain of ϕ_α which contains the set of all cumulative distribution function on $\mathbb{R}$ and $\mathbb{F}_{n,I}$ for all $n \in \mathbb{N}^*$, all $(X_n(\omega))_{n\in\mathbb{N}^*}$ and all $\omega \in \Omega$. Let F be a cumulative distribution function.*

- *If F is strictly increasing at $F^{-1}(\alpha)$ then ϕ_α is continuous at F.*
- *Assume that F is differentiable at $F^{-1}(\alpha)$ and $F'(F^{-1}(\alpha)) = f(F^{-1}(\alpha)) > 0$. Then ϕ_α is Hadamard-differentiable in F tangentially to*

$$D_0 = \{h \in \mathcal{S},\ h \text{ continue en } F^{-1}(\alpha)\}.$$

Moreover

$$\phi'_\alpha(h) = -\frac{h(\phi_\alpha(F))}{F'(\phi(F))} = -\frac{h(F^{-1}(\alpha))}{f(F^{-1}(\alpha))}.$$

Remark 8. *Remark that we can take* $D_\alpha = \{F \in \mathscr{S},\ \lim_{t\to-\infty} F(t) = 0,\ \lim_{t\to+\infty} F(t) = 1\}$.

Proof. *The prooof of the second assertion can be found in [38] for instance. About the first assertion, let* $\varepsilon > 0$. *Since* F *is strictly increasing at* $F^{-1}(\alpha)$ *we have*

$$F(F^{-1}(\alpha) - \varepsilon) < \alpha < F(F^{-1}(\alpha) + \varepsilon).$$

Let $(F_n)_{n\in\mathbb{N}} \subset D_\alpha$ *such that* $\|F_n - F\|_\infty \xrightarrow[n\to\infty]{} 0$. *Then there exists* $N > 0$ *such that for all* $n \geq N$

$$F_n(F^{-1}(\alpha) - \varepsilon) < \alpha < F_n(F^{-1}(\alpha) + \varepsilon)$$

which implies that

$$F^{-1}(\alpha) - \varepsilon \leq F_n^{-1}(\alpha) \leq F^{-1}(\alpha) + \varepsilon.$$

Thus $|F_n^{-1}(\alpha) - F^{-1}(\alpha)| \leq \varepsilon$ *and* ϕ_α *is continuous at* F. □

The following results shows that $q^I_{n,\alpha}$ has an asymptotic variance strictly less than the uninformed empirical quantile whenever the vector $cov_P(g, 1_{]-\infty,F^{-1}(\alpha)]}) \neq 0$.

Theorem 8. *Let* F *be the cumulative distribution function generating the sample and* q_α *the* α*-quantile of* F, $\alpha \in]0,1[$.

- *If* F *is strictly increasing at* $F^{-1}(\alpha)$ *then*

$$q^I_{n,\alpha} \xrightarrow[n\to\infty]{a.s.} q_\alpha.$$

- *If* F *is differentiable at* $F^{-1}(\alpha)$ *and* $F'(F^{-1}(\alpha)) = f(F^{-1}(\alpha)) > 0$ *then*

$$\sqrt{n}(q^I_{n,\alpha} - q_\alpha) \Rightarrow \mathscr{N}\left(0, \frac{\alpha(1-\alpha) - \tilde{I}}{\left(f(F^{-1}(\alpha))\right)^2}\right)$$

where $\tilde{I} = cov_P(g, 1_{]-\infty,F^{-1}(\alpha)]})^T \Sigma^{-1} cov_P(g, 1_{]-\infty,F^{-1}(\alpha)]}) \geq 0$.

Proof. *Remind that the class of function* $\mathscr{F} = \{f_s = 1_{]-\infty,s]},\ s \in \mathbb{R}\}$ *is* P*-Donsker and so* P*-Glivenko-Cantelli. By Theorem 6 and 7, we have*

$$\|\mathbb{F}_{n,I} - F\|_\infty \xrightarrow[n\to\infty]{a.s.} 0,$$
$$\sqrt{n}(\mathbb{F}_{n,I} - F) \Rightarrow G_I \ in\ l^\infty(\mathbb{R}),$$

where for all $s \in \mathbb{R}$, $G_I(s) := G_I(f_s) = G(f_s) - cov_P(g, f_s)^T \Sigma^{-1} G(g)$. *Observe that* $G_I \in l^\infty(\mathbb{R})$. *To get the first assertion, apply Lemma 8 to obtain*

$$\phi_\alpha(\mathbb{F}_{n,I}) \xrightarrow[n\to\infty]{a.s.} \phi_\alpha(F).$$

To derive the second assertion, let prove that G_I is càdlàg on $\mathbb{R}$ and continuous at $F^{-1}(\alpha)$. For that it is enough to prove that these following real-valued maps are càdlàg on $\mathbb{R}$ and continuous at $F^{-1}(\alpha)$

$$\psi_1 : s \mapsto G(f_s),$$
$$\psi_2 : s \mapsto cov_P(g, 1_{]-\infty,s]}) = Pg1_{]-\infty,s]} - PgF(s).$$

Concerning ψ_1, recall that G is a P-Brownian bridge having almost surely continuous sample paths with respect to the semimetric $\rho^2(h_1,h_2) = Var_P(h_1 - h_2)$ for all $h_1, h_2 \in L^2(P)$ – see Theorem 7. The map $\varphi : (\mathbb{R}, |\cdot|) \to \left(\mathscr{F}, \rho\right)$ defined by $\varphi(s) = f_s$ for all $s \in \mathbb{R}$ is càdlàg and continuous at $F^{-1}(\alpha)$. As a matter of fact, F is continuous at $F^{-1}(\alpha)$ and for all s, s'

$$\rho(f_s, f_{s'}) = \sqrt{Var_P\left(f_s - f_{s'}\right)} \leq \|f_s - f_{s'}\|_{L^2(P)} = \sqrt{F(s) + F(s') - 2F(\min(s,s'))}$$

hence ψ_1 is càdlàg on $\mathbb{R}$ and continuous at $F^{-1}(\alpha)$. Concerning ψ_2, let show that $s \mapsto P\varphi_g(P)1_{]-\infty,s]}$ is càdlàg. Let $(s_n)_{n\in\mathbb{N}^*} \subset \mathbb{R}$ be a decreasing sequence tending to s when $n \to +\infty$. Since $g \in L^1(P)$, we deduce by dominated convergence theorem,

$$Pg1_{]-\infty,s_n]} \xrightarrow[n\to\infty]{} Pg1_{]-\infty,s]}.$$

Likewise, by dominated convergence theorem,

$$Pg1_{]-\infty,s']} \xrightarrow[s'\to s,\ s'<s]{} Pg1_{]-\infty,s[}.$$

Since F is continuous at $F^{-1}(\alpha)$, we deduce that ψ_2 is càdlàg and continuous at $F^{-1}(\alpha)$. Finally the functional delta method and Lemma 8 readily imply

$$\sqrt{n}(q^I_{n,\alpha} - q_\alpha) = \sqrt{n}(\phi_\alpha(\mathbb{F}_{n,I}) - \phi_\alpha(F)) \Rightarrow \phi'_\alpha(G_I) = -\frac{G_I(F^{-1}(\alpha))}{f(F^{-1}(\alpha))}.$$

Since $F(F^{-1}(\alpha)) = \alpha$, we have $Var_P(1_{]-\infty,F^{-1}(\alpha)]}) = \alpha(1-\alpha)$. □

To conclude, besides the asymptotics of Theorem 8 let show that the information I impacts also small samples. Let use $n = 210$ *i.i.d.* random variables with distribution $P = \mathscr{N}(0,1)$ to estimate sequentially the median of P – that is 0. It is assumed that the auxiliary information I is given by Pg with $g(x) = (x, x^2)$ for all $x \in \mathbb{R}$. We draw at figure 4 $q_{n,\alpha}$ the sequences $q^I_{n,\alpha}$ for $\alpha = \frac{1}{2}$ for every $n \in [\![2, 210]\!]$ – with the same sample.

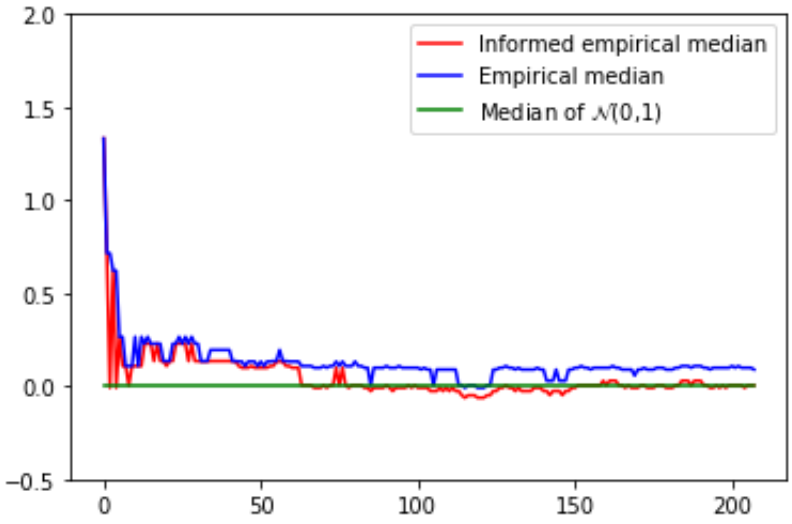

FIGURE B.4 – Medians

Bibliographie

[1] Albertus, M. (2019a). Exponential increase of test power for z-test and chi-square test with auxiliary information. *Prepint, arXiv :2003.02941*.

[2] Albertus, M. (2019b). Raking-ratio empirical process with auxiliary information learning. *Prepint, arXiv :1901.08519*.

[3] Albertus, M. (2020a). Exponential increase of the power of the independence and homogeneity chi-square tests with auxiliary information. *Prepint, arXiv :2005.02952*.

[4] Albertus, M. (2020b). *Processus empirique avec informations auxiliaires*. PhD thesis, Université Paul Sabatier.

[5] Albertus, M. and Berthet, P. (2019). Auxiliary information : the raking-ratio empirical process. *Electron. J. Stat.*, 13(1) :120–165.

[6] Amari, S.-i. and Nagaoka, H. (2007). *Methods of Information Geometry*. American Mathematical Society.

[7] Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning : Theoretical Foundations*. Cambridge University Press.

[8] Arradi-Alaoui, S. (01/07/2021). Optimal use of auxiliary information : information geometry and empirical process. *Prepint, arXiv :2107.00563*.

[9] Berthet, P. and Mason, D. (2006). Revisiting two strong approximation results of dudley and philipp. *High Dimensional Probability, p 155-172*.

[10] Blanchard, G., Delattre, S., and Roquain, E. (2014). Testing over a continuum of null hypothseses with false discovery rate control. *Bernoulli, p304-333, vol 20*.

[11] Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for fdr control. *Electron. J. Stat, p 963-992*.

[12] Bru, B. (1988). Estimations laplaciennes. un exemple : la recherche de la population d'un grand empire 1785-1812. *Journal de la société française de statistique, p 6-45.*

[13] Cadre, B. and Vial, C. (2012). *Statistique Mathématique Cours & Exercices Corrigés.* ELLIPSES.

[14] Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, p 208-214.*

[15] Cleveland, W. and Devlin, S. (1998). Locally-weighted regression : An approach to regression analysis by local fitting. *Journal of the American Statistical Association 83 numéro 403, 596–610.*

[16] Deville, J.-C. and Särndal, C.-E. (2013). Calibration estimators in survey sampling. *Journal of the American Statistical Association, p 376–382.*

[17] Feelders, A. J. (1999). *Handling Missing Data in Trees : Surrogate Splits or Statistical Imputation.* Lecture Notes in Computer Science, t. 1704, Springer, p. 329–334.

[18] Fischer, A. (2011). *Apprentissage statistique non supervisé : grande dimension et courbes principales.* PhD thesis, université Pierre et Marie Curie - UPMC (Paris 6).

[19] Fischer, A. (2014). Deux méthodes d'apprentissage non supervisé : synthèse sur la méthode des centres mobiles et présentation des courbes principales. *Journal de la Société Française de Statistique, 2-35*, 155.

[20] Giné, E. and Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probab. Theory Relat. Fields, 333-387.*

[21] Gudmundsson, S. (2019). Lecture notes : An introduction to riemannian geometry.

[22] Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika, p 179-188*, 55.

[23] Knottnerus, P. (2003). *Sample Survey Theory.* Springer.

[24] Laplace, P. (1783). Sur les naissances, les mariages et les morts. *Histoire de l'Académie Royale des sciences.*

[25] Lauritzen, S. (1987). *Statistical Manifolds*, volume X. Differential Geometry in Statistical Inference. p 165-216. Institute of Mathematical Statistics. IMS Monograph Hayward.

[26] Lehmann, E., Romano, J., and Shaffer, J. (2005). On optimality of stepdown and stepup multiple test procedures. *Ann. Statist., p 1084 - 1108.*

[27] Lesage, E. (2013). *Use of auxiliary information in survey sampling at the sampling stage and the estimation stage.* PhD thesis, Université de Rennes 1.

[28] Matsumoto, K. (1993). Any statistical manifold has a contrast function on the c^3-functions taking the minimum at the diagonal of the product manifold. *Hiroshima mathematical journal, p 327 - 332.*

[29] M.Evans and T.Swartz (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods.* OUP Oxford.

[30] Nielsen, F. and Nock, R. (2018). On the geometry of mixtures of prescribed distributions. *IEEE ICASSP*.

[31] Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist., p 90-120*.

[32] Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall/CRC.

[33] Pollard, D. (1982). Quantization and the method of k-means. *IEEE Transactions on Information Theory, p 199 - 205*.

[34] Qin, J. and J.Lawless (1994). Empirical likelihood and general estimating equations. *Ann. Statist., p 300-325*.

[35] Roquain, E. (2011). Type i error rate control for testing many hypotheses : a survey with proofs. *Journal de la Société Française de la Statistique, p 3-38*.

[36] Sarkar, J. (1991). One-armed bandit problems with covariates. *Ann. Statist, p 1978 - 2002*.

[37] Tarima, S. and Pavlov, D. (2006). Using auxiliary information in statistical function estimation. *ESAIM Probab. Stat.*, 10 :11–23.

[38] Van der Vaart, A. and Wellner, A. (1996). *Weak convergence and Empirical processes.* Springer.

[39] Yuan, A., He, W., Wang, B., and Qin, G. (2012). U-statistic with side information. *Journal of Multivariate Analysis, p 20-38*.

[40] Zhang, B. (1997a). Empirical likelihood confidence intervals for m-functionals in the presence of auxiliary information. *Statistics* & *Probability Letters, p 87-97*.

[41] Zhang, B. (1997b). Estimating a distribution function in the presence of auxiliary information. *Metrika, p 221–244*.

[42] Zhang, B. (1997c). Quantile processes in the the presence of auxiliary information. *Annals of the Institute of Statistical Mathematics, p 35-55*.

[43] Zhang, J. (2007). A note on curvature of α-connections of a statistical manifold. *Annals of the Institute of Statistical Mathematics volume, p 161–170*.