Optimal transport-based machine learning with applications to genomics and actuarial science

Thi Thanh Yen Nguyen

Université Paris Cité
ED Sciences Mathématiques de Paris Centre (386)
Laboratoire MAP5, CNRS UMR 8145

THÈSE DE DOCTORAT

Optimal transport-based machine learning with applications to genomics and actuarial science

par

Thi Thanh Yen NGUYEN

Spécialité : Mathématiques appliquées

Directeurs de thèse : Olivier BOUAZIZ *et* Antoine CHAMBAZ

Date de soutenance : 14 Décembre 2023

Composition du Jury :

Olivier BOUAZIZ, MCF-HDR	(Université Paris Cité)	Co-directeur de thèse
Claire BOYER, MCF-HDR	(Sorbonne Université)	Examinatrice
Antoine CHAMBAZ, PR	(Université Paris Cité)	Co-directeur de thèse
Mohamed HEBIRI, MCF-HDR	(Université Gustave Eiffel)	Rapporteur
Estelle KUHN, DR	(INRAE)	Examinatrice
Christian NERI, DR	(Sorbonne Université)	Membre invité
Pierre NEUVIAL, DR	(Université Toulouse III Paul Sabatier)	Rapporteur

MAP5 (UMR CNRS 8145)
Université Paris Cité
Campus Saint-Germain-des-Prés
45, rue des Saints Pères
75270 Paris cedex 06

Abstract

Optimal transport (OT) is a powerful mathematical theory at the interface between the theories of optimization and probability, with many applications in a wide range of fields. This thesis presents the application of OT and statistics to two domains: biology and actuarial sciences.

The first part of the thesis addresses the biological challenge of better understanding micro-RNA (miRNA) regulation in the striatum of Huntington's disease (HD) model mice. To do so, we build several algorithms designed to learn a pattern of correspondence between two data sets in situations where it is desirable to match elements that exhibit a relationship belonging to a known parametric model. The two data sets contain miRNA and messenger-RNA (mRNA) data, respectively, each data point consisting in a multi-dimensional profile. The strong biological hypothesis is that if a miRNA induces the degradation of a target mRNA or blocks its translation into proteins, or both, then the profile of the former, say y, should be similar to minus the profile of the latter, say $-x$. We consider a loosened hypothesis stating that y is then similar to $t(x)$ where t is an affine transformation in a parametric class that includes minus the identity and translates expert knowledge about the experiment that yielded the data. The algorithms unfold in two stages. During the first stage, an OT plan P and an optimal affine transformation are learned, using the Sinkhorn-Knopp algorithm and a mini-batch gradient descent. During the second stage, P is exploited to derive either several co-clusters or several sets of matched elements. A simulation study illustrates how the algorithms work and perform. The real data application further illustrates their applicability and interest.

The second part of thesis addresses an actuarial problem related to drought events in France. Drought events rank as the second most costly natural disasters within the French legal framework of the natural disaster compensation scheme. A critical aspect of the national compensation scheme involves cities submitting requests for the government declaration of natural disaster for a drought event as a key step. We take on the challenge of forecasting which cities will submit such requests. The problem can be tackled as a classification task, leveraging the power of classification algorithms. Taking a slightly different perspective, we introduce an alternative procedure that hinges on OT and iPiano, an inertial proximal algorithm for nonconvex optimization. The optimization problem is designed so as to yield a sparse vector of predictions because it is known that relatively few cities will submit requests. Additionally, we develop a hybrid procedure that synergistically combines and utilizes both types of predictions, resulting in enhanced forecasting accuracy. The real data application is presented and discussed in details. The convergence of the iPiano algorithm is established, using the notion of o-minimal structures.

Keywords: Huntington's disease; matching; natural disasters; omics data; optimal transport; proximal algorithm; Sinkhorn algorithm; Sinkhorn divergence.

Résumé

Le transport optimal (OT) est une théorie mathématique puissante à l'interface de la théorie de l'optimisation et de celle des probabilités, avec de nombreuses applications dans un large éventail de domaines. Cette thèse présente l'application de la théorie du transport optimal et des statistiques dans deux domaines : la biologie et l'actuariat.

La première partie de la thèse aborde le problème biologique consistant à chercher à mieux comprendre la régulation des micro-ARN (miARN) dans le striatum des souris modèles de la maladie de Huntington (HD). Pour ce faire, nous développons plusieurs algorithmes conçus pour apprendre un modèle de correspondance entre deux ensembles de données dans des situations où il est souhaitable de faire correspondre des éléments qui présentent une relation appartenant à un modèle paramétrique connu. Les deux ensembles de données contiennent des informations sur les miARN et les ARN messagers (ARNm), respectivement, chaque point de données consistant en un profil multidimensionnel. L'hypothèse biologique forte est que si un miARN induit la dégradation d'un ARNm cible ou bloque sa traduction en protéines, ou les deux, alors le profil du premier, disons y, devrait être similaire à moins le profil du second, disons $-x$. Nous considérons une hypothèse plus souple selon laquelle y est alors similaire à $t(x)$, où t est une transformation affine dans une classe paramétrique qui inclut moins l'identité et traduit les connaissances d'experts sur l'expérience qui a produit les données. Les algorithmes se déroulent en deux étapes. Au cours de la première étape, un plan de transport optimal P et une transformation affine optimale sont appris, en utilisant l'algorithme de Sinkhorn-Knopp et une descente de gradient par mini-batch. Au cours de la deuxième étape, P est exploité pour obtenir soit plusieurs co-clusters, soit plusieurs ensembles d'éléments appariés. Une étude de simulation illustre la façon dont les algorithmes fonctionnent et performent. L'application aux données réelles illustrent plus avant leur applicabilité et leur intérêt.

La deuxième partie de la thèse traite d'un problème actuariel lié aux événements de sécheresse en France. Les sécheresses sont les deuxièmes catastrophes naturelles les plus coûteuses dans le cadre du régime français d'indemnisation des catastrophes naturelles. Un aspect critique du régime national d'indemnisation implique que les villes soumettent des demandes de déclaration de catastrophe naturelle pour un événement de sécheresse, ce qui constitue une étape clé. Nous relevons le défi de prévoir quelles villes soumettront de telles demandes. Le problème peut être abordé comme une tâche de classification, en tirant partie de la puissance des algorithmes de classification. Dans une perspective légèrement différente, nous introduisons une procédure alternative qui s'appuie sur la théorie OT et sur iPiano, un algorithme proximal inertiel pour l'optimisation non convexe. Le problème d'optimisation est conçu de manière à produire un vecteur parcimonieux de prédictions, car on sait que relativement peu de villes soumettront des demandes. En outre, nous développons une procédure hybride qui combine et utilise de manière synergique les deux types de prédictions, ce qui permet d'améliorer la précision des prévisions. L'application aux données réelles est présentée et discutée en détail. La convergence de l'algorithme iPiano est établie à l'aide de la notion de structures o-minimales.

Mots-Clefs : Algorithme de Sinkhorn ; algorithme proximal ; catastrophes naturelles ; divergence de Sinkhorn ; données omics ; maladie de Huntington ; matching ; transport optimal.

Remerciements

Contents

List of Figures

List of Tables

Notations

- $[\![M]\!]$: the set of integers $\{1, \dots, M\}$.
- $\mathbf{1}_M$: vector of size M with all entries equal to 1.
- $\mathbf{0}_d$: vector of size d with all entries equal to 0.
- $\mathrm{diag}(\rho)$, any $\rho \in \mathbb{R}^M$: the $M \times M$ matrix with diagonal ρ and zero elsewhere.
- $\langle \cdot, \cdot \rangle_F$: Euclidean dot-product between vectors; for two matrices of the same size A and B, $\langle A, B \rangle_F := \mathrm{Tr}\, A^\top B$ is the Frobenius dot-product.
- $a \otimes b := ab^\top \in \mathbb{R}^{M \times N}$, for any $(a, b) \in \mathbb{R}^M \times \mathbb{R}^N$.
- $a \odot b := (a_m b_m) \in \mathbb{R}^M$ for any $(a, b) \in (\mathbb{R}^M)^2$.
- $\mathbf{f} \oplus \mathbf{g} := \mathbf{f}\mathbf{1}_M^\top + \mathbf{1}_N \mathbf{g}^\top \in \mathbb{R}^{M \times N}$, for any $\mathbf{f} \in \mathbb{R}^M$, $\mathbf{g} \in \mathbb{R}^N$
- $\sharp$: the push forward operator.
- $\mathcal{C}(\mathcal{X})$: the set of continuous functions from $\mathcal{X}$ to $\mathbb{R}$.
- ∇: the gradient operator.
- Ω_M: the probability simplex of dimension $(M-1)$, that is, the set of vectors $x \in (\mathbb{R}_+)^M$ such that $\sum_{m=1}^M x_m = 1$.
- $\mathcal{P}(\mathcal{X})$: the set of probability measures on $\mathcal{X}$.
- a and b: elements of Ω_M and Ω_N viewed as histograms/laws.
- α and β: probability measures on spaces $\mathcal{X}$ and $\mathcal{Y}$.
- δ_x: the Dirac measure on $\{x\}$.
- $\Pi(a, b)$: set of couplings between a and b.
- $\Pi(\alpha, \beta)$: set of couplings between α and β.
- $\mu_{\mathbf{x}}^a := \sum_{m \in [\![M]\!]} a_m \delta_{x_m}$ and $\nu_{\mathbf{y}}^b := \sum_{n \in [\![N]\!]} b_n \delta_{y_n}$): weighted empirical measures attached to $\mathbf{x} := \{x_1, \dots, x_M\}$ and $\mathbf{y} := \{y_1, \dots, y_N\}$.
- $(x, y) \mapsto c(x, y)$: cost function, with associated pairwise cost matrix evaluated on $\mathbf{x}$ and $\mathbf{y}$, $C(\mathbf{x}, \mathbf{y})$, such that $(C(\mathbf{x}, \mathbf{y}))_{m,n} = c(x_m, y_n)$ for all $m \in [\![M]\!]$, $n \in [\![N]\!]$.
- $K := \mathrm{e}^{-C/\gamma}$, any cost matrix C and $\gamma > 0$: Gibbs kernel associated to C and γ.
- $\mathcal{W}_p(\alpha, \beta)$: the p-Wasserstein distance between two probability measures α and β.
- $\mathrm{OT}_c(\alpha, \beta)$: optimal transport criterion specific to α, β and c.
- $\mathrm{OT}_{\gamma,c}(\alpha, \beta)$: regularized optimal transport criterion specific to α, β, c and $\gamma > 0$, a parameter controlling the amount of regularization based on the entropy.

- $\mathcal{S}_{\gamma,c}(\alpha, \beta)$: Sinkhorn loss (or divergence) specific to α, β, c and $\gamma > 0$, a parameter controlling the amount of regularization of the related c-specific optimal transport criterion based on the entropy.

Conflicts in notation between chapters

We have tried to use coherent and non-conflicting notation for the mathematical objects defined in this thesis. However, for the sake of consistency with the conventions of the field, we made the choice to keep conventional notations for known quantities.

Theses notational conflicts have been kept to ease the understanding of the manuscript. They occur between different chapters but not inside each chapter. We stress that the potential uncertainty is removed when the context is taken into consideration.

1

Résumé long

La théorie du transport optimal (OT) a trouvé de nombreuses applications dans divers domaines car elle fournit un outil puissant pour comparer les distributions de probabilités, une étape cruciale du "machine learning". Dans cette thèse, nous exploitons la théorie OT et les statistiques pour aborder des problèmes survenant en biologie et en science actuarielle.

Le problème biologique consiste à mettre en relation des expressions de micro-ARNs et d'ARN messagers dans le striatum de souris modèles de la maladie de Huntington. Le problème actuariel est lié à l'anticipation de la déclaration de catastrophe naturelle pour un événement de sécheresse.

Malgré la disparité apparente entre les deux applications, elles trouvent leur unité sous le cadre général que nous appelons "OT-based machine learning". Dans le reste de ce résumé long, nous entrelaçons les deux études. L'alternance répétée entre les problèmes biologiques et actuariels révèle naturellement les caractéristiques partagées des études.

Contextes

À PROPOS DE LA MALADIE DE HUNTINGTON. La maladie de Huntington (HD) est un trouble neurodégénératif progressif autosomique dominant. HD est caractérisée par des mouvements chromatiques involontaires avec des perturbations cognitives et comportementales. Elle est causée par une expansion d'une série répétitive de triplets CAG dans le gène huntingtin (Walker, 2007 ; MacDonald et al., 1993).

Comme plusieurs maladies neurodégénératives telles que la maladie d'Alzheimer, la maladie de Parkinson et la sclérose latérale amyotrophique, HD est liée à une dérégulation génique. Cela a encouragé de grandes études sur les mécanismes régulateurs des gènes (voir Langfelder et al., 2018, en particulier). L'expression génique est contrôlée en limitant la quantité d'ARN messager (mRNA) produite à partir d'un gène particulier au niveau de la transcription et en régulant la traduction de mRNA en protéines au niveau post-transcriptionnel. Les acteurs les plus importants à ce dernier niveau sont les petits ARNs non codants ap-

pelés micro-ARNs (miRNAs). Ces éléments d'explication justifient pourquoi les chercheurs s'intéressent à l'étude de l'interaction entre les miRNAs et les mRNAs dans HD, afin d'obtenir une compréhension plus approfondie de la maladie et, éventuellement, de développer de nouveaux traitements.

> Le premier problème que nous abordons dans cette thèse concerne HD. Notre objectif est de contribuer à l'étude de l'interaction complexe entre les miRNAs et les mRNAs dans le contexte de cette maladie.

À PROPOS DE L'ANTICIPATION DE LA DÉCLARATION DE CATASTROPHE NATURELLE POUR UN ÉVÉNEMENT DE SÉCHERESSE. Le changement climatique se réfère à des changements à long terme dans les modèles statistiques de la météo sur Terre (Assadollahi, 2019). Alors que le changement climatique s'est produit très lentement tout au long de l'histoire de la Terre en raison de la variabilité naturelle, il se produit de nos jours plus rapidement en raison des activités humaines. Cela a conduit à une large gamme d'impacts dans toutes les régions de la Terre ainsi que dans de nombreux secteurs économiques. Notamment, le changement climatique exacerbe les sécheresses en les rendant plus fréquentes, plus longues et plus intenses. Par exemple, Spinoni et al. (2015, 2017) ont étudié plusieurs indices de sécheresse sur 60 ans pour montrer que de nombreuses régions européennes ont connu des conditions plus sèches au cours des trois dernières décennies.

Dans cette thèse, nous appelons *événement de sécheresse* le phénomène de gonflement de l'argile en conditions humides et de son retrait en conditions sèches. Compte tenu du paragraphe précédent, les événements de sécheresse devraient devenir plus fréquents et plus intenses également. Ceci est très problématique car, en induisant des déplacements de la surface du sol, les événements de sécheresse peuvent entraîner des dommages importants aux bâtiments, tels que des fissures au sol et dans les murs. Pour donner une idée du défi notons que, selon la Caisse Central de Réassurance (CCR, un réassureur du secteur public fournissant aux cédants opérant en France une couverture contre les catastrophes naturelles et les risques non assurables), le coût annuel moyen des événements de sécheresse entre 2016 et 2020 est de 1,1 milliard d'euros, soit une augmentation d'un facteur trois par rapport à la période 2002-2015 (CCR, 2021).

Les événements de sécheresse ont été intégrés en 1989 dans le régime français d'indemnisation des catastrophes naturelles (également connu sous le nom de régime "CatNat"), sept ans après sa création par la loi. Par conséquent, les coûts engendrés par les dommages liés aux événements de sécheresse sont couverts par toutes les polices d'assurance de propriété privée (MTES, 2016). Depuis lors, les événements de sécheresse sont la deuxième catastrophe naturelle la plus coûteuse avec un coût cumulatif de 14 milliards d'euros, et jusqu'à 2 milliards d'euros en 2003. Étant donné que 90% du marché français de l'assurance contre les catastrophes naturelles est réassuré par CCR (CCR, 2022), c'est finalement l'État français qui est exposé.

Quelques mots sur le régime français d'indemnisation des catastrophes naturelles sont nécessaires. Premièrement, comme expliqué dans (Charpentier et al., 2022a ; Heranval et al., 2022), le régime d'indemnisation des catastrophes naturelles est fondé sur le principe de solidarité : le même taux de prime supplémentaire est appliqué à tous les contrats d'assurance de propriété. Deuxièmement, le déclenchement du régime d'indemnisation des catastrophes naturelles repose sur deux conditions essentielles :

- la propriété ayant subi des pertes ou des dommages doit être couverte par une police d'assurance de biens et de responsabilité civile, une exigence privée ;

- un décret gouvernemental déclarant officiellement une catastrophe naturelle doit être publié dans le “Journal Officiel”, une condition publique.

Il est important de noter que la responsabilité d'initier la demande de cette déclaration gouvernementale au sein des municipalités qu'ils supervisent incombe aux maires.

> Le deuxième problème que nous abordons dans cette thèse concerne les événements de sécheresse. Notre objectif est de construire et d'analyser des outils afin de mieux prévoir les demandes de déclaration gouvernementale de catastrophe naturelle pour un événement de sécheresse.

Formalisation

À PROPOS DE LA MALADIE DE HUNTINGTON. Ces dernières années, l'avènement de technologies de séquençage avancées, telles que la RNAseq, a permis aux chercheurs de générer de grands ensembles de données englobant la génomique, la protéomique, la transcriptomique et la métabolomique. Grâce à l'analyse de ces vastes ensembles de données, les chercheurs ont acquis des connaissances sur la génétique, la biologie humaine et la compréhension de diverses maladies. Notamment, des données sur HD sont de plus en plus disponibles, telles que les signatures d'expression de mRNA de HD dans des cerveaux humains post-mortem (Neueder and Bates, 2014 ; Cha, 2007) et les modifications de l'expression des miRNAs observées dans plusieurs modèles de souris (Langfelder et al., 2018). Cependant, malgré cette richesse croissante d'informations, notre connaissance de l'interaction entre mRNA et miRNA dans HD reste plutôt limitée. Plusieurs défis doivent être relevés. Premièrement, le manque de données de haute qualité en séries temporelles de différents types de cellules et de tissus dans des conditions saines et malades constitue un obstacle. Deuxièmement, il est intrinsèquement difficile de modéliser avec précision les interactions entre les miRNAs et les mRNAs. Troisièmement, la réalisation d'expériences pour la détection et la validation des gènes cibles des miRNAs est à la fois coûteuse et chronophage, comme discuté dans (Nazarov and Kreis, 2021).

Encouragés par les découvertes prometteuses de Langfelder et al. (2018) ; Mégret et al. (2020), notre objectif est de mettre en lumière l'interaction entre les mRNA et les miRNAs à partir de données multidimensionnelles disponibles publiquement via le Gene Expression Omnibus (GEO) et le portail HDinHD. Les données sont collectées à trois âges différents (2, 6, 10 mois) dans quatre régions du cerveau et dans le foie d'une série allélique de souris modèles de HD avec des longueurs croissantes de CAG dans le gène Huntingtin endogène (longueurs polyQ : Q20, Q80, Q92, Q111, Q140, Q175) (Langfelder et al., 2016, 2018). Pour chaque combinaison de longueur polyQ et d'âge, l'expression des mRNA et des miRNAs de huit souris, dont quatre femelles et quatre mâles, a été quantifiée. Après prétraitement (Mégret et al., 2020), le jeu de données final se compose de $M = 13,616$ profils de mRNA, $X := \{x_1, x_2, \dots, x_M\} \subset \mathbb{R}^d$, et $N = 1,143$ profils de miRNA, $Y := \{y_1, y_2, \dots, y_N\} \subset \mathbb{R}^d$ avec $d = 15$.

> L'hypothèse biologique au cœur de notre étude postule que lorsqu'un miRNA déclenche la dégradation d'un mRNA cible ou entrave sa traduction en protéines, ou les deux, alors le profil du premier, disons y, et celui du second, disons x, présentent ce que nous appelons une relation de miroir, signifiant de manière approximative que y et $-x$ sont similaires. Notre objectif est d'identifier des groupes de mRNAs et de miRNAs qui interagissent en exploitant l'hypothèse biologique, X et Y. Il s'agit d'une tâche difficile

car les miRNAs et leurs mRNAs régulés s'engagent souvent dans des relations de miroir complexes, à plusieurs niveaux.

À PROPOS DE L'ANTICIPATION DE LA DÉCLARATION DE CATASTROPHE NATURELLE POUR UN ÉVÉNEMENT DE SÉCHERESSE. L'ensemble de données est obtenu en fusionnant plusieurs jeux de données, soit fournis par les cédants de CCR, soit collectés auprès d'autres sources. Les unités expérimentales sont les villes françaises. Chacune d'entre elles peut contribuer une structure de données pour une année donnée t (par convention, $t = 1, 2, 3$ correspond respectivement aux années 2019, 2020 et 2021) et une semaine donnée u (l'entier $u \in \mathcal{U}_t \subset \mathbb{N}^*$ indiquant le nombre de semaines à partir de la première semaine de l'année t, avec $44 \leq u \leq 85$). Une structure de données englobe de multiples aspects du profil d'une ville, visant à fournir une représentation complète de son contexte et des déclencheurs potentiels pour demander la déclaration gouvernementale de catastrophe naturelle pour un événement de sécheresse. Elle se compose des blocs de variables suivants :

Description de la ville Les variables de ce bloc fournissent une compréhension globale des caractéristiques de la ville.

Exposition de la ville aux événements de sécheresse Les variables de ce bloc décrivent l'exposition de la ville aux événements de sécheresse. Elles s'appuient sur le Soil Wetness Index (SWI). Fourni par Météo-France, les données SWI consistent en des séries temporelles de valeurs (une tous les dix jours) variant entre -3,33 (sol très sec) et 2,33 (sol très humide).

Historique des demandes de la ville Ce bloc nous donne un aperçu du processus décisionnel de la ville, de ses intentions et actions concernant la soumission d'une demande de déclaration gouvernementale de catastrophe naturelle pour un événement de sécheresse.

Statut actuel de la demande de la ville Cette variable indique si la ville a soumis ou non une demande de déclaration gouvernementale de catastrophe naturelle pour un événement de sécheresse pour l'année t pendant la semaine u ou avant.

Description du voisinage de la ville Ce bloc se concentre sur les alentours de la ville.

Désignons par $x_m \in \mathcal{X}$ ($m \in [\![M]\!] = \{1, \ldots, M\}$) la description spécifique de la ville pour l'année t et la semaine u, pour laquelle il est connu si la ville a demandé ou non la déclaration de catastrophe naturelle pour un événement de sécheresse pour l'année t d'ici la semaine u, une information notée par $y_m \in \{0, 1\}$ (avec la convention $y_m = 1$ si la ville avec la description au niveau de la ville spécifique à l'année x_m l'a fait). De plus, désignons par $x'_n \in \mathcal{X}$ ($n \in [\![N]\!]$) la description spécifique de la ville pour l'année t et la semaine u, pour laquelle il n'est pas connu si la ville demandera ou non *in fine* une déclaration de catastrophe naturelle pour un événement de sécheresse pour l'année t, une information notée par $y'_n \in \{0, 1\}$.

Notre objectif ultime est d'apprendre à prédire y'_n en fonction de x'_n (pour chaque $n \in [\![N]\!]$) en exploitant $\{(x_m, y_m) : m \in [\![M]\!]\}$. L'objectif est difficile à atteindre pour plusieurs raisons. Premièrement, relativement peu de villes demandent la déclaration de catastrophe naturelle pour un événement de sécheresse. Deuxièmement, nous souhaitons encourager des prédictions prenant exactement la valeur 0. Troisièmement, même en tentant de créer une description de ville complète et générique, il peut encore y avoir

en jeu des facteurs insaisissables qui déclenchent une demande de déclaration gouvernementale de catastrophe naturelle pour un événement de sécheresse. Par exemple, les affiliations et alliances politiques entre les villes peuvent être influentes, mais elles sont assez difficiles à saisir.

Conception et mise en œuvre d'algorithmes

À PROPOS DE LA MALADIE DE HUNTINGTON. De manière informelle, nous recherchons des couples $(m, n) \in [\![M]\!] \times [\![N]\!]$ tels que le $n$ième miRNA induit la dégradation du $m$ième mRNA ou bloque sa traduction en protéines ou les deux. Nous sommes guidés par l'hypothèse biologique forte que, si c'est le cas, alors le profil y_n du premier est similaire à l'opposé du profil x_m du second — c'est-à-dire, x_m et y_n présentent une relation de miroir. Il est à noter qu'un seul miRNA peut cibler plusieurs mRNAs. Les relations de miroir réelles peuvent être plus ou moins aiguës, par exemple à cause d'effets de seuil, ou de plusieurs miRNAs ciblant le même mRNA ou d'un seul miRNA ciblant plusieurs mRNAs. Par conséquent, au lieu d'utiliser rigoureusement des comparaisons entre $-x_m$ et y_n, nos algorithmes apprendront à partir des données une transformation pertinente $\theta \in \Theta$, un ensemble paramétrique de relations de miroir relâchées, et utiliseront des comparaisons entre $\theta(x_m)$ et y_n. À cette fin, nous nous appuyons sur la théorie OT pour apprendre une transformation optimale $\hat{\theta} \in \Theta$ et une matrice de transport P interprétée comme une *matrice de similarité* entre les profils de miRNAs et de mRNAs. Ensuite, nous exploitons P pour identifier les paires pertinentes en utilisant une procédure d'appariement.

Concrètement, nous identifions un θ pertinent dans Θ en résolvant

$$\min_{\omega \in \Omega_M} \min_{\theta \in \Theta} S_{\gamma,c}(\mu^{\omega}_{\theta(X)}, \nu_Y) \tag{1.1}$$

où

- $\theta(X) := \theta(x_1), \dots, \theta(x_M)$ est l'image de X par $\theta \in \Theta$;
- $\Omega_M := \{\omega \in (\mathbb{R}_+)^M : \sum_{m \in [\![M]\!]} \omega_m = 1\}$ est le simplexe de dimension $(M-1)$;
- $\mu^{\omega}_{\theta(X)} := \sum_{m \in [\![M]\!]} \omega_m \delta_{\theta(x_m)}$ est la mesure empirique pondérée par ω attachée à $\theta(X)$;
- $\nu_Y := \sum_{n \in [\![N]\!]} \delta_{y_n}$ est la mesure empirique attachée à Y ;
- $\mathcal{S}_{\gamma,c}$ est la perte de Sinkhorn correspondant à une fonction de coût c définie sur $\mathbb{R}^d \times \mathbb{R}^d$ et un paramètre de régularisation $\gamma > 0$ (Genevay et al., 2018).

Nous décidons d'optimiser par rapport à $\omega \in \Omega_M$ car nous ne nous attendons pas à associer finalement un y_n à chaque x_m. Pour résoudre (1.1), nous mettons à jour itérativement ω puis θ, en utilisant l'algorithme Sinkhorn-Knopp et une descente de gradient à mini-batches.

Notre code est écrit en `python`. Nous adaptons l'algorithme Sinkhorn implémenté par Aude Genevay et disponible ici. Les descentes de gradient stochastiques s'appuient sur le cadre d'apprentissage automatique `pytorch`.

Finalement, nous sommes intéressés par le minimiseur $(\hat{\omega}, \hat{\theta})$ et par le plan de transport optimal $\hat{P}$ caché dans la définition de $S_{\gamma,c}(\mu^{\omega}_{\theta(X)}, \nu_Y)$. Une fois le plan $\hat{P}$ obtenu, nous nous appuyons sur une procédure d'appariement pour trouver les paires pertinentes. Enfin, une analyse biologique supplémentaire est menée pour identifier les paires les plus fiables.

À PROPOS DE L'ANTICIPATION DE LA DÉCLARATION DE CATASTROPHE NATURELLE POUR UN ÉVÉNEMENT DE SÉCHERESSE. Rappelons que $\{(x_m, y_m) : m \in [\![M]\!]\} \subset \mathcal{X} \times \{0,1\}$ et $\{(x'_n, y'_n) : n \in [\![N]\!]\} \subset \mathcal{X} \times \{0,1\}$ sont deux collections de couples pour lesquels on souhaite prédire y'_n en fonction de x'_n, pour chaque $n \in [\![N]\!]$, en utilisant les observations passées (x_1, y_1), …, (x_M, y_M). Pour ce faire, nous proposons de résoudre le problème d'optimisation suivant :

$$\arg\min_{\theta \in \mathbb{R}^N} \{\mathcal{S}\gamma(\mathbf{z}, \mathbf{z}'(\theta)) + g_\tau(\theta)\} \tag{1.2}$$

où

- pour tout $\theta \in \mathbb{R}^N$,

$$\mathbf{z} := ((x_1, y_1), \ldots, (x_M, y_M)), \quad \mathbf{z}'(\theta) := ((x'_1, \theta_1), \ldots, (x'_N, \theta_N));$$

- la fonction de coût $c : (\mathcal{X} \times \mathbb{R}) \times (\mathcal{X} \times \mathbb{R}) \to \mathbb{R}_+$ est donnée par

$$c((x, y), (x', \theta)) := \mathrm{dis}(x, x')^2 + (y - \theta)^2$$

  pour une distance ou dissimilarité dis sur $\mathcal{X}$;

- g_τ ($\tau > 0$) est une fonction convexe donnée soit par $g_\tau(\theta) := \tau\|\theta\|_1 + \mathbf{I}\{\theta \in [0,1]^N\}$, avec $\|\theta\|_1 := \sum_{n \in [\![N]\!]} |\theta_n|$, soit par $g_\tau(\theta) := \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$, où $\mathbf{I}\{A\}$ vaut 0 si A est vrai et $+\infty$ sinon ;

- $\mathcal{S}\gamma(\mathbf{z}, \mathbf{z}'(\theta))$ est la perte de Sinkhorn entre $\mathbf{z}$ et $\mathbf{z}'(\theta)$ correspondant à la fonction de coût ci-dessus c et au paramètre de régularisation $\gamma > 0$.

Résoudre (1.2) n'est pas simple, en partie parce que le critère à minimiser est la somme de la fonction non convexe différentiable $f : \theta \mapsto \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta))$ et de la fonction convexe non différentiable g_τ. Heureusement, nous pouvons nous appuyer sur l'algorithme iPiano (Ochs et al., 2015), qui a été développé précisément pour traiter ce type de problèmes d'optimisation.

> La convergence de l'algorithme iPiano est établie, en utilisant la notion de structures o-minimales.
>
> Notre code est écrit en `python` et utilise `pytorch`. Une fois de plus, nous adaptons l'algorithme de Sinkhorn implémenté par Aude Genevay et disponible ici. De plus, nous nous appuyons sur un algorithme efficace disponible pour implémenter la projection mentionnée sur la boule ℓ^1 (Duchi et al., 2008). Une procédure mini-batches permet de faire face à des situations où M et N sont grands.

De plus, nous nous appuyons notamment sur HYPERBAND (Li et al., 2018), une approche fondée sur les bandits pour l'optimisation des hyperparamètres, afin de définir la cruciale fonction de coût, et sur une recherche gourmande afin d'affiner ensuite les autres hyperparamètres. Nous comparons les résultats obtenus en agrégeant les prédictions acquises à partir d'algorithmes de classification avec ceux obtenus par la procédure OT. De plus, nous introduisons la procédure hybride qui combine et utilise de manière synergique les deux types de prédictions.

Résultats

À PROPOS DE LA MALADIE DE HUNTINGTON. Nous appliquons notre algorithme d'appariement pour découvrir des motifs cachés dans les données de séquençage d'ARN obtenues dans le

striatum de souris modèles de HD afin de trouver les appariements potentiels. Pour garantir la pertinence biologique des appariements, nous ne retenons que ceux connus pour exhiber des preuves de liaison, comme indiqué dans les bases de données TargetScan, MicroCosm et miRDB. Spécifiquement, une paire (x, y) est retenue si et seulement si le mRNA dont le profil est x et le miRNA dont le profil est y sont tous les deux parmi les 27,355 mRNA et 1,478 miRNAs apparaissant dans les bases de données TargetScan, MicroCosm, et miRDB.

> Les 1,247 appariements retenus sur 7,521 produits par l'algorithme d'appariement sont tous présentés sur cette page du site web compagnon.

Nous évaluons et comparons la signification biologique des mRNA retenus par les algorithmes WGCNA (Langfelder et al., 2018), MiRAMINT (Mégret et al., 2020) et notre algorithme d'appariement.

> L'analyse d'enrichissement révèle que les appariements mRNA-miRNA produits par notre algorithme d'appariement sont principalement annotés pour *"extracellular matrix organization"* (qui se rapporte à l'identité cellulaire) et secondairement annotés pour *"mitigation of host antiviral defense response"*, et pour *"conventional motile cilium"*.

Au contraire, les appariements produits par l'algorithme MiRAMINT sont principalement annotés pour *"regulation of defense response to virus by host"*, ce qui est lié à la réponse au stress et à l'immunité innée. De plus, les appariements produits par l'algorithme WGCNA sont principalement annotés pour *"axonogenesis"*, ce qui est lié à la dynamique du cytosquelette et à la morphologie cellulaire.

À PROPOS DE L'ANTICIPATION DE LA DÉCLARATION DE CATASTROPHE NATURELLE POUR UN ÉVÉNEMENT DE SÉCHERESSE.

> Nous appliquons le super learner (un algorithme d'apprentissage automatique), notre "procédure OT" et une procédure hybride qui tire parti des deux approches précédentes pour prédire les probabilités de soumettre une demande relative à l'année 2021 pour chaque semaine u et toutes les villes qui n'ont pas encore soumis de demande d'ici la semaine u. Les prédictions hybrides semblent trouver un juste équilibre entre les prédictions produites par le super learner et par la procédure OT.
>
> Lors de l'évaluation des trois approches en utilisant l'erreur quadratique moyenne comme critère, la procédure hybride surpasse la procédure OT qui, à son tour, performe mieux que le super learner. De plus, la procédure hybride surpasse l'algorithme actuellement utilisé chez CCR.

2

Introduction

Optimal transport (OT) theory has found many applications in diverse fields as it provides a powerful tool for comparing probability distributions, one crucial step in machine learning. In this thesis, we leverage OT theory and statistics to deal with problems arising in biology and actuarial science.

The biological problem is to assess the possible relationships between microRNA and mRNA expression in the striatum of Huntington's disease model mice. The actuarial problem relates to the anticipation of the declaration of natural disaster for a drought event.

Despite the apparent disparity between the two applications, they find unity under the overarching framework that we call "OT-based machine learning". In the rest of this introductory chapter, we intertwine the two studies. The repeated alternation between the biological and the actuarial problems naturally reveals the studies' shared features.

In Section 2.1, we provide the applications' contexts. In Section 2.2, we introduce elements of formalization and clarify what are our objectives. In Section 2.3, we discuss the states of the art upon which our work builds. In Section 2.4, we concisely expose the algorithms that we designed in order to fulfill our objectives. This includes some details about the algorithms' implementation. In Section 2.5, we briefly describe the results of our studies. Finally, in Section 2.6, we outline the structure of the rest of the document – where we present the full-fledged studies.

2.1 What is this thesis about?

About Huntington's disease. Huntington's disease (HD), an autosomal-dominant, progressive neurodegenerative disorder, is characterized by involuntary chromatic movements with cognitive and behavioral disturbances. It is caused by an expansion of a repeating CAG triplet series in the huntingtin gene (Walker, 2007; MacDonald et al., 1993). In normal individuals the CAG repeat length ranges from 10 to 35, while in HD individuals, it ranges from 36 to more than 120 (see Figure 2.1). In detail, HD patients with 36-40 CAG

repeats may have late onset or may not develop symptoms; HD patients with 41-59 CAG repeats may have symptoms onset in their fourth decade; and CAG repeats greater than 60 in length lead to juvenile onset (Andrew et al., 1993). There are currently no treatments to prevent the onset or to slow the progression of HD.

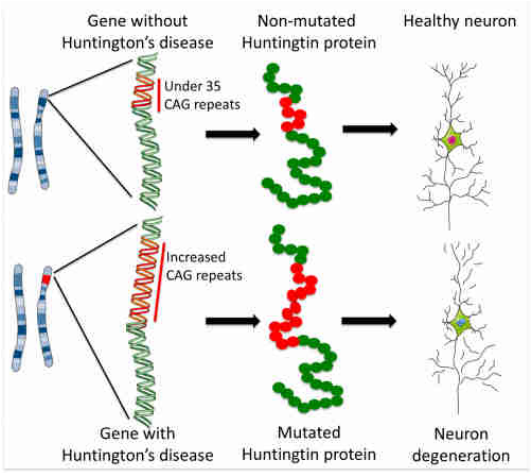

Figure 2.1 – CAG repeat expansions in HD. Source: California's Stem Cell Agency.

Like several neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease and amyotrophic lateral sclerosis, HD relates to gene deregulation. This has encouraged large studies of gene regulatory mechanisms (see Langfelder et al., 2018, in particular). Gene expression is controlled by limiting the amount of messenger-RNA (mRNA) produced from a particular gene at the transcription level and by regulating the translation of mRNA into proteins at the post-transcriptional level. The most important instruments in the latter level are small non-coding RNAs called micro-RNAs (miRNAs). It binds to a complementary sequence in the 3'UTR of the target mRNA resulting in a rapid degradation of the mRNA or less frequently in an inhibition of its translation into protein (see Pasquinelli, 2012, and Figure 2.2). These basic facts explain why researchers are interested in studying the interaction between miRNAs and mRNAs in HD to gain a deeper understanding on the disease and, eventually, to develop new therapeutics.

> The first problem that we tackle in this thesis pertains to HD. Our aim is to make a substantial and noteworthy contribution to the study of the intricate interplay between miRNAs and mRNAs in the context of HD.

ABOUT THE ANTICIPATION OF THE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT. Climate change refers to long-term shifts in the statistical patterns of weather on Earth (Assadollahi, 2019). While climate change has occurred very slowly throughout Earth's history due to natural variability, it is nowadays happening more rapidly due to human activities. This has led to a wide range of impacts across every region of Earth as well as many economic sectors. Notably, climate change exacerbates droughts by making them more frequent, longer, and more intense. For instance, Spinoni et al. (2015, 2017) studied multiple drought indexes over 60 years to show that many European regions has experienced drier conditions in the last 3 decades.

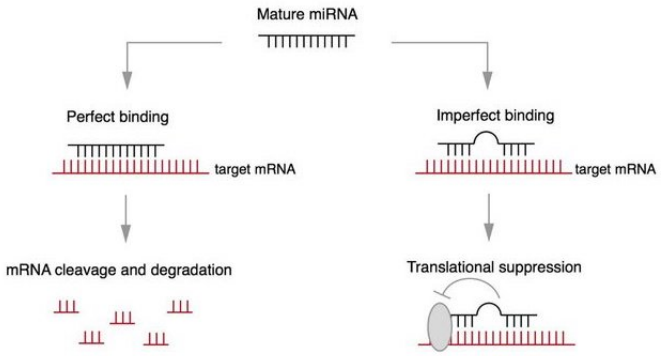

Figure 2.2 – Mechanism of miRNA action. MiRNA can bind to specific regions of target mRNA transcripts and destabilizes the target transcript and/or blocks its translation. Source: (Teixeira et al., 2014).

Figure 2.3 – Left: the clay shrinking-swelling phenomenon. Right: an example of crack due to the clay shrinking and swelling phenomenon.

In this dissertation we call *drought event* the phenomenon of clay swelling in humid conditions and shrinking in dry ones. In view of the previous paragraphs, drought events are expected to become more frequent and more intense too. This is very problematic because, by inducing displacements of the ground surface, drought events can lead to significant damages to buildings, such as cracks on the floor and in the walls, see Figure 2.3. To give a sense of the challenge, note that, according to Caisse Central de Réassurance (CCR, a public-sector reinsurer providing cedents operating in France with coverage against natural catastrophes and uninsurable risks), the average annual cost of drought events between 2016 and 2020 is 1.1 billion euros, a threefold increase relative to the 2002-2015 period (CCR, 2021).

Drought events have been integrated in 1989 into the French natural disaster compensation scheme (also known as the "Cat Nat" scheme), 7 years after its creation by law. Consequently, the costs incurred by damages related to drought events are covered by all private property insurance policies (MTES, 2016). Since then, drought events are the second costliest natural disaster with a cumulative cost of 14 billions euros, and as much as 2 billion euros in 2003. Because 90% of the French natural disasters insurance market is reinsured by CCR (CCR, 2022), it is the French State that is exposed eventually.

A few words on the French natural disaster compensation scheme are in order. First, as

explained in (Charpentier et al., 2022a; Heranval et al., 2022), the natural disaster compensation scheme is based on the principle of solidarity: the same additional-premium insurance rate is applied to all property insurance contracts. Second, the initiation of the natural disaster compensation scheme hinges on two essential prerequisites:

- the property that has incurred losses or damages must fall under the coverage of a property and casualty insurance policy, a private requirement;
- a government decree officially declaring a natural disaster must be published in the "Journal Officiel", a public condition.

Importantly, the responsibility for initiating the request for this government declaration within the municipalities they oversee lies with the mayors.

> The second problem that we address in this thesis pertains to drought events. Our objective is to build and analyze tools to better forecast the requests of the government declaration of natural disaster for a drought event.

2.2 Formalization

ABOUT HUNTINGTON'S DISEASE. In recent years, the advent of advanced sequencing technologies, such as RNAseq, has allowed researchers to generate large datasets encompassing genomics, proteomics, transcriptomics and metabolomics. Through the analysis of these extensive datasets, the researchers have gained insights into genetics, human biology and the understanding of various diseases. Notably, data on HD are increasingly available, such as mRNA expression signatures of HD in post-mortem human brains (Neueder and Bates, 2014; Cha, 2007) and the alterations in miRNA expression observed across multiple mouse models (Langfelder et al., 2018). However, despite this growing wealth of information, our knowledge of the interaction between mRNA and miRNA in HD remains rather limited. Several challenges must be dealt with. Firstly, the lack of high quality, time-series data from different cell types and tissues from healthy and diseased conditions poses a hurdle. Secondly, it is inherently difficult to model accurately the interactions between miRNAs and mRNAs. Thirdly, conducting experiments for the detection and validation of miRNA target genes is both costly and time-consuming, as discussed in (Nazarov and Kreis, 2021).

Encouraged by the promising findings of Langfelder et al. (2018); Mégret et al. (2020), our goal is to shed light on the interaction between mRNAs and miRNAs based on multidimensional data which are publicly available through Gene Expression Omnibus (GEO) and the HDinHD portal. The data are collected at three different ages (2, 6, 10 months) in four brain regions and liver from an allelic series of HD model knock-in mice with increasing CAG lengths in the endogenous Huntingtin gene (poly Q lengths: Q20, Q80, Q92, Q111, Q140, Q175) (Langfelder et al., 2016, 2018). For each combination of poly Q length and age, miRNA and mRNA expression of 8 mice including 4 females and 4 males have been quantified. After preprocessing (Mégret et al., 2020), the final dataset consists of $M = 13,616$ mRNA profiles, $X := \{x_1, x_2, \dots, x_M\} \subset \mathbb{R}^d$, and $N = 1,143$ miRNA profiles, $Y := \{y_1, y_2, \dots, y_N\} \subset \mathbb{R}^d$ with $d = 15$.

> The biological hypothesis at the core of our study posits that when a miRNA triggers the degradation of a target mRNA or hinders its translation into proteins, or both, then the profile of the former, say y, and the one of the latter, say x, exhibit what we call a mirroring relationship, meaning loosely that y and $-x$ are similar. Our aim is

to identify groups of mRNAs and miRNAs that interact by leveraging the biological hypothesis, X and Y. This is a challenging task because miRNA and their regulated mRNAs often engage in complex, many-to-many mirroring relationships.

ABOUT THE ANTICIPATION OF THE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT. The data set is obtained by merging several data sets, either provided by CCR's cedents or collected from other sources, namely the National Institute for Statistical and Economic Studies (Insee), Geographic National Institute (IGN), French Geological Survey (BRGM) and Météo-France. The experimental units are the French cities. Each of them can contribute a data structure for a given year t (by convention, $t = 1, 2, 3$ respectively correspond to years 2019, 2020 and 2021) and a given week u (the integer $u \in \mathcal{U}_t \subset \mathbb{N}^*$ being the number of weeks starting from the first week of year t, with $44 \leq u \leq 85$). A data structure encompasses multiple aspects of a city's profile, aiming to provide a comprehensive representation of its context and potential triggers for requesting the government declaration of natural disaster for a drought event. It consists of the following blocks of variables:

City description The variables within this block provide a global understanding of the city's characteristics.

City exposure to drought events The variables within this block outline the city's exposure to drought events. They build upon the Soil Wetness Index (SWI). Provided by Météo-France, the SWI data consist of time series of values (one every ten-day period) ranging between -3.33 (very dry soil) and 2.33 (very wet soil).

City history of requests This block gives us insight into the city's decision-making process, intentions and actions regarding the submission of a request for the government declaration of natural disaster for a drought event.

City current request status This variable indicates whether or not the city submitted a request for the government declaration of natural disaster for a drought event for year t during week u or before.

City's vicinity description This block focuses on the city's surroundings.

Let us denote by $x_m \in \mathcal{X}$ ($m \in [\![M]\!] = \{1, \ldots, M\}$) the mth year-t and week-u specific description of a city for which it is known whether or not the city requested the declaration of natural disaster for a drought event for year t by week u, a piece of information denoted by $y_m \in \{0, 1\}$ (with convention $y_m = 1$ if the city with year-specific city-level description x_m did). Moreover, let us denote by $x'_n \in \mathcal{X}$ ($n \in [\![N]\!]$) the nth year-t and week-u specific description of a city for which it is not known whether of not the city will request during week u a declaration of natural disaster for a drought event for year t, a piece of information denoted by $y'_n \in \{0, 1\}$.

Our ultimate objective is to learn to predict y'_n based on x'_n (for every $n \in [\![N]\!]$) by leveraging $\{(x_m, y_m) : m \in [\![M]\!]\}$. The objective is challenging for several reasons. First, relatively few cities do request the declaration of natural disaster for a drought event. Secondly, we aim to encourage predictions that lean towards 0. Thirdly, even when attempting to create a comprehensive and generic city description, there may still be elusive factors at play that trigger a request for the government declaration of

natural disaster for a drought event. For instance, political affiliations and alliances among cities may be influential, but they are quite difficult to capture.

2.3 State of the art

About Huntington's disease. Since the first discovery of miRNAs in 1993 (Lee et al., 1993), numerous methods of computational prediction, experimental detection and validation of miRNA target genes have been developed to understand how miRNA function and to identify their role in varied biological processes. As reviewed in (Huang et al., 2010; Nazarov and Kreis, 2021), the traditional *experimental methods* of miRNA target gene interaction include

- mutation studies,
- gene-silencing techniques,
- classic genetic studies.

On the other hand, the current experimental methods include

- reporter gene assay,
- protein level analyses,
- crosslinking followed by immunoprecipitation of RICS complexes (CLIP and CLASH methods),
- other biochemical approaches.

Although experimental methods can provide direct evidence between miRNAs and their targets, they are time-consuming and expensive, especially when multiple miRNAs are of interest. Moreover, false positive results may arise in some experimental methods, for example when analyzing data from RNASeq experiments following over-expression of miRNAs.

Computational methods for miRNA target prediction leverage databases that have been published over the past 10 years, including but not limited to TargetScan (Lewis et al., 2005), MicroCosm (Betel et al., 2010) and miRDB (Ding et al., 2016). Citing Nazarov and Kreis (2021), these *computational methods* for miRNA target prediction rely upon several criteria:

- degree of Watson-Crick pairing between miRNA seed region and target site,
- evolutionary conservation across species,
- thermodynamic properties,
- accessibility of target sites,
- sequence composition in the vicinity of seeds and target sites.

For instance, the analysis of TargetScan conducted by Huang et al. (2010) integrates thermodynamics-based modeling of miRNA-mRNA interactions and sequence alignment analysis to predict conserved miRNA binding sites among different species.

Today, with the development of high-throughout technologies, the datasets of mRNA and miRNA profiles across many samples and conditions are increasingly available for data integration. Nazarov and Kreis (2021) enumerate three main approaches to integrate different datasets:

- *(i)* data-driven methods based on similarities,
- *(ii)* data-driven methods based on matrix factorization,
- *(iii)* so-called hybrid methods.

When adopting approach *(i)*, one typically chooses one of two methods: the first consists in defining a similarity matrix based on classical similarity measures (such as the Pearson and Spearman correlation coefficients; the cosine similarity; the mutual information); the second consists in building on canonical correlation analysis to establish linear relations between two datasets. When adopting approach *(ii)*, the expression matrices A and B of mRNA and miRNA data measured in m samples (a column for each sample) are approximated by products of lower rank matrices, $A \approx \widetilde{A}\Omega_a$ and $B \approx \widetilde{B}\Omega_b$. The matrices $\widetilde{A}$ and $\widetilde{B}$ can be interpreted as expression matrices of "meta-mRNA" and "meta-miRNA", Ω_a and Ω_b as weight-matrices. Integration can be performed by correlating the weight profiles over the samples, yielding a network of linked components. Unfortunately, neither approach *(i)* or *(ii)* can discriminate between true interactions and fake interactions originating from hidden regulators such as transcription factors. Nazarov and Kreis (2021) argue that the hybrid approach, by combining information about miRNA targets and experimental observations, is best qualified to identify the highest potential interactions.

Two previous analyses of miRNA regulation have been performed using the same datasets as us of mRNA and miRNA profiles in the striatum of HD knock-in mice. The first analysis relies on the WGCNA algorithm, a weighted gene co-expression network analysis which yields clusters of genes whose expression profiles are correlated (Langfelder et al., 2018). The second analysis relies on the MiRAMINT algorithm (Mégret et al., 2020). MiRAMINT is a pipeline whose main steps consist in *(a)* carrying out a weighted gene co-expression network analysis, *(b)* using random forests to select candidate matchings, and *(c)* using Spearman's correlation test and a multiple testing procedure to identify the more reliable matchings. The two analyses suffer from little congruence with only one mRNA-miRNA pair output by both the MiRAMINT and WGCNA algorithms: Mir132-Pafah121. As a matter of fact, this is a common issue of all approaches for miRNA target prediction.

> In view of the two previous paragraph, the algorithms that we develop in this thesis to study how miRNAs "come together to regulate the expression of a gene or a group of genes" (an expression borrowed from Nazarov and Kreis (2021)) in HD are based on similarities. We do not rely on classical similarity measures but instead exploit tools from OT theory. Naturally, we compare our findings to those of Langfelder et al. (2018); Mégret et al. (2020).

ABOUT THE ANTICIPATION OF THE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT. Paraphrasing (Logar and van den Bergh, 2011, page 4, first paragraph), "[t]he existing literature on the costs of drought [events] is scarce, fragmented and heterogeneous, and there is a need for comprehensive costs estimations to help designing effective policy responses." To the best of our knowledge, only five recent works (Chatelain and Loisel, 2021; Charpentier et al., 2022b; Heranval et al., 2022; Ecoto et al., 2021; Ecoto and Chambaz, 2022) are publicly available about the cost prediction of a drought event (all in France), while the studies conducted by insurance companies are confidential.

The problem can be separated into two sub-problems (Ecoto et al., 2021; Ecoto and Chambaz, 2022): sub-problem 1 consists in predicting which cities will make a request for the government's declaration of natural disaster for a drought event; sub-problem 2 consists

in predicting the cost of a drought event for those cities that obtained the government declaration of natural disaster for a drought event. (Chatelain and Loisel, 2021) takes on both sub-problems simultaneously. On the other hand, (Charpentier et al., 2022b; Heranval et al., 2022) predict which cities will experience claims (a proxy for sub-problem 1) and subsequently estimate the cost for these cities.

In the work of Charpentier et al. (2022b), the authors employed a combination of Generalized Linear Models (GLM) and tree-based models, which are variants of the random forest algorithm, to predict both the average cost per claim and the number of houses experiencing losses in each city. Subsequently, they calculated a city-specific predicted cost by multiplying these two values. To obtain the overall cost, they summed up all the city-specific costs. As for Heranval et al. (2022); Chatelain and Loisel (2021), they utilized penalized GLM and machine learning algorithms, including random forest and extreme gradient boosting, for the same purpose. For a given drought event, Heranval et al. (2022) predicted city-specific costs by considering the number of houses and employing a common linear regression model. In contrast, Chatelain and Loisel (2021) predicted costs at the house level, using geolocated data and applying several GLMs. In both cases, the overall cost was eventually estimated by summing up either the city-specific or house-specific costs.

In (Ecoto et al., 2021; Ecoto and Chambaz, 2022), the authors develop and apply a new methodology to predict the cost of a drought event. The methodology hinges on Super Learning, a general aggregation strategy to learn a feature of the law of the data identified through an ad hoc risk function by relying on a library of algorithms. Theoretical guarantees reveal that it is possible to learn from a short time series (thirty years of data, one data-structure per year) of many slightly dependent data (each data-structure gathers data across all French cities).

> In this thesis we focus on sub-problem 1. We also rely on machine learning algorithms, but our main contribution lies in the use of tools from OT theory to try and obtain better performances.

2.4 Design and implementation of algorithms

ABOUT HUNGTINTON'S DISEASE. Informally, we look for couples $(m, n) \in [\![M]\!] \times [\![N]\!]$ such that the nth miRNA induces the degradation of the mth mRNA or blocks its translation into proteins or both. We are guided by the strong biological hypothesis that, if that is the case, then the profile y_n of the former is similar to minus the profile x_m of the latter — that is, x_m and y_n exhibit a mirroring relationship. Of note, it it expected that a single miRNA can target several mRNAs. The actual mirroring relationships can be more or less acute, for instance because of threshold effects, or of multiple miRNAs targeting the same mRNA or of a single miRNA targeting several mRNAs. Therefore, instead of rigidly using comparisons between $-x_m$ and y_n, our algorithms will learn from the data a relevant transformation $\theta \in \Theta$, a parametric set of loose mirroring relationships, and use comparisons between $\theta(x_m)$ and y_n. To this end, we rely on OT theory to learn an optimal transformation $\hat{\theta} \in \Theta$ and an OT matrix P interpreted as a *similarity matrix* between miRNA and mRNA profiles. Then, we exploit P to derive the relevant pairs using a matching procedure. As such, our algorithms belong to the family of data-driven methods based on similarities as described by (Nazarov and Kreis, 2021).

Concretely, we identify a relevant $\theta \in \Theta$ by solving

$$\min_{\omega \in \Omega_M} \min_{\theta \in \Theta} S_{\gamma,c}(\mu^{\omega}_{\theta(X)}, \nu_Y) \tag{2.1}$$

where

- $\theta(X) := \{\theta(x_1), \dots, \theta(x_M)\}$ is the image of X by $\theta \in \Theta$;
- $\mu^{\omega}_{\theta(X)} := \sum_{m \in [\![M]\!]} \omega_m \delta_{\theta(x_m)}$ is the ω-weighted empirical measure attached to $\theta(X)$;
- $\nu_Y := \sum_{n \in [\![N]\!]} \delta_{y_n}$ is the empirical measure attached to Y;
- $\mathcal{S}_{\gamma,c}$ is the Sinkhorn loss corresponding to a cost function c defined on $\mathbb{R}^d \times \mathbb{R}^d$ and a regularization parameter $\gamma > 0$ (Genevay et al., 2018).

We decide to optimize with respect to $\omega \in \Omega_M$ because we do not expect to associate a y_n to every x_m eventually. To solve (2.1), we iteratively update ω then θ, using the Sinkhorn-Knopp algorithm and a mini-batch gradient descent.

> Our code is written in `python`. We adapt the Sinkhorn algorithm implemented by Aude Genevay and available here. The stochastic gradient descents relies on the machine learning framework `pytorch`.

Eventually, we are interested in the minimizer $(\hat{\omega}, \hat{\theta})$ and in the OT plan $\hat{P}$ hidden in the definition of $S_{\gamma,c}(\mu^{\omega}_{\theta(X)}, \nu_Y)$. Once the OT plan is derived, we rely on a matching procedure to find the relevant pairs. Finally, a further biological analysis is conducted to identify the more reliable pairs.

ABOUT THE ANTICIPATION OF THE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT. Recall that $\{(x_m, y_m) : m \in [\![M]\!]\} \subset \mathcal{X} \times \{0,1\}$ and $\{(x'_n, y'_n) : n \in [\![N]\!]\} \subset \mathcal{X} \times \{0,1\}$ are two collections of couples for which it is desired to predict y'_n based on x'_n, for every $n \in [\![N]\!]$, using past observations (x_1, y_1), ..., (x_M, y_M). To do so, we propose to solve the following optimization problem:

$$\underset{\theta \in \mathbb{R}^N}{\arg\min} \left\{ \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta)) + g_\tau(\theta) \right\} \tag{2.2}$$

where

- for all $\theta \in \mathbb{R}^N$,

$$\mathbf{z} := ((x_1, y_1), \dots, (x_M, y_M)), \quad \mathbf{z}'(\theta) := ((x'_1, \theta_1), \dots, (x'_N, \theta_N));$$

- the cost function $c : (\mathcal{X} \times \mathbb{R}) \times (\mathcal{X} \times \mathbb{R}) \to \mathbb{R}_+$ is given by

$$c((x, y), (x', \theta)) := \mathrm{dis}(x, x')^2 + (y - \theta)^2$$

  for a distance or dissimilarity dis on $\mathcal{X}$;

- g_τ ($\tau > 0$) is a convex function given by either $g_\tau(\theta) := \tau \|\theta\|_1 + \mathbf{I}\{\theta \in [0,1]^N\}$, with $\|\theta\|_1 := \sum_{n \in [\![N]\!]} |\theta_n|$, or $g_\tau(\theta) := \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$, where $\mathbf{I}\{A\}$ equals 0 if A is true and $+\infty$ otherwise;
- $\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta))$ is the Sinkhorn loss between $\mathbf{z}$ and $\mathbf{z}'(\theta)$ corresponding to the above cost function c and the regularization parameter $\gamma > 0$.

Solving (2.2) is not straightforward, in part because the criterion to minimize is the sum of the non-convex differentiable function $f : \theta \mapsto \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta))$ and of the convex non-differentiable function g_τ. Luckily, we can rely on the so-called iPiano algorithm (Ochs et al., 2015) which was developed precisely to deal with such optimization problems.

The iPiano algorithm starts from an initial $\theta^{-1} = \theta^0 \in]0,1[^N$ and the update scheme informally writes as (below, α, β are positive constants)

$$\theta^{k+1} = \mathrm{Prox}_{\alpha g_\tau}\left(\theta^k - \alpha \nabla f(\theta^k) + \beta(\theta^k - \theta^{k-1})\right), \tag{2.3}$$

where the proximal map $\mathrm{Prox}_{\alpha g_\tau}$ is defined by

$$\mathrm{Prox}_{\alpha g_\tau}(t) := \underset{\theta \in \mathbb{R}^N}{\arg\min} \left\{ \tfrac{1}{2}\|\theta - t\|_2^2 + \alpha g_\tau(\theta) \right\}. \tag{2.4}$$

On the one hand, if $g_\tau(\theta) = \tau\|\theta\|_1 + \mathbf{I}\{\theta \in [0,1]^N\}$ then (2.4) is simply given by

$$\left(\mathrm{Prox}_{\alpha g_\tau}(t)\right)_n = \min\{(|t_n| - \alpha\tau)_+, 1\}.$$

In particular, if $t \in [0,1]^N$ then $(\mathrm{Prox}_{\alpha g_\tau}(t))_n = (t_n - \alpha\tau)_+$ for every $n \in [\![N]\!]$. On the other hand, if $g_\tau(\theta) = \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$ then the proximal map is the Euclidean projection onto the ℓ^1-ball centered at 0 and with radius τ.

> The convergence of the iPiano algorithm is established, using the notion of o-minimal structures.
>
> Our code is written in `python` and uses the machine learning framework `pytorch`. Once again we adapt the Sinkhorn algorithm implemented by Aude Genevay and available here. Moreover, we rely on an efficient algorithm available to implement the aforementioned projection on the ℓ^1-ball (Duchi et al., 2008). A mini-batch procedure allows to cope with situations where M and N are large.

Furthermore, we notably rely on HYPERBAND (Li et al., 2018), a bandit-based approach to hyperparameter optimization, to define the pivotal cost function, and on a simple grid search to then fine-tune the other hyperparameters. We compare the results obtained by aggregating the predictions acquired from classification algorithms with those achieved through the OT-procedure. Moreover, we introduce the hybrid procedure which synergistically combines and utilizes the two types of predictions.

2.5 Results

ABOUT HUNGTINTON'S DISEASE. We apply our matching algorithm to discover patterns hidden in RNA-seq data obtained in the striatum of HD model mice to find the potential matching. In an effort to guarantee biological relevance to the matchings, we only retain those showing evidence for binding sites as indicated in the databases TargetScan, MicroCosm and miRDB. Specifically, a pair (x, y) is retained if and only if the mRNA whose profile is x and the miRNA whose profile is y are both among the 27,355 mRNAs and 1,478 miRNAs appearing in TargetScan, MicroCosm, and miRDB databases.

> The 1,247 matchings retained out of 7,521 output by the matching algorithm are all presented on this page of the companion website.

We assess and compare the biological significance of the mRNAs retained by the WGCNA, MiRAMINT and our matching algorithms.

The enrichment analysis reveals that the mRNA-miRNA matchings output by our matching algorithm are primarily annotated for *extracellular matrix organization* (which relates to cell identity)[a] and secondarily annotated for *mitigation of host antiviral defense response*[b], and for *conventional motile cilium*[c].

[a]GO:0030198, a process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of an extracellular matrix.

[b]GO:0050690, evasion by virus of host immune response.

[c]GO:0097729, a motile cilium where the axoneme has a ring of 9 outer microtubules doublets plus 2 central micro tubules.

On the contrary, the matchings output by the MiRAMINT algorithm are primarily annotated for *regulation of defense response to virus by host**, which relates to stress response and innate immunity. Furthermore, the matchings output by the WGCNA algorithm are primarily annotated for *axonogenesis*†, which relates to cytoskeleton dynamics and cell morphology.

ABOUT THE ANTICIPATION OF THE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT.

We apply the super learner (a machine learning algorithm), our procedure based on OT (say the OT-procedure) and a hybrid procedure that leverages both previous approaches to predict the probabilities of submitting a request relative to year 2021 for every week u and all cities which did not submit a request yet by week u. The hybrid predictions seem to strike a fine balance between the predictions output by the super learner and the OT-procedure.

When evaluating the three approaches using mean squared error as the criterion, the hybrid procedure outperforms the OT-procedure which, in turn, performs better than the super learner. Moreover, the hybrid procedure outperforms the algorithm currently in use at CCR.

2.6 Overview of this thesis

This thesis covers the author's work as part of the Ph.D. requirements. The layout of the thesis is intended to facilitate independent reading of chapters by minimizing dependencies between them.

Chapter 3 presents a modicum of OT theory. While this chapter contains crucial notions that will be referred to throughout the thesis, a reader familiar with basic concepts of OT theory can safely skip it.

Chapter 4 addresses the problem of learning a pattern of correspondence between two data sets in situation where it is desirable to match elements that exhibit a relationship belonging to a known parametric model. Our ultimate objective is to shed light on the interaction between mRNAs and miRNAs based on data collected in the striatum (a brain region) of HD model mice. The main part concerns the optimization program at the core of the study and several algorithms to solve it. We also present and comment upon the real data application.

*GO:0050691, any host process that modulates the frequency, rate or extent of the antiviral response of a host cell or organism.

†GO:0007409, de novo generation of a long process of a neuron, including the terminal branched region. Refers to the morphogenesis or creation of shape or form of the developing axon, which carries efferent (outgoing) action potential from the cell body towards target cells.

Chapter 5 deals with the actuarial problem consisting in predicting which cities will submit a request for the government declaration of natural disaster for a drought event in France. We present there the so called OT-procedure that we developed to make sparse predictions. We also discuss how to solve the nonconvex optimization task that sits at its core using the algorithm iPiano (Ochs et al., 2015), from both theoretical and computational perspectives. Additionally, we developed a hybrid procedure that synergistically combines and utilizes both types of predictions, derived from classification algorithms and the OT-procedure. We describes the full-fledged application to the challenge of forecasting which cities will submit a request for the government declaration of natural disaster for a drought event. In the last part, we gather the proofs of the convergence of of the iPiano algorithm using a theorem proven in (Ochs et al., 2015). The Kurdyka-Lojasiewicz property (Attouch et al., 2010) and notion of o-minimal structures (Wilkie, 1996) play a central role.

Chapter 6 brings the thesis to a close. It discusses the implications of the contributions of the results of the preceding chapters under a unified view, elaborates on connections between them, and proposes various avenues of future work.

3

Elements of optimal transport

This chapter is dedicated to a concise but self-contained introduction to optimal transport (OT). It largely builds upon the monograph (Peyré and Cuturi, 2019).

We describe the basics of OT by introducing the related assignment and Monge problems along with their generalization, the Kantorovich problem. After that, we consider regularized OT and discuss its advantage in practice. Theoretical and numerical results for regularized OT are presented. We finally describe a family of divergences, the so-called Sinkhorn divergences, interpolating between regularized OT and Maximum Mean Discrepancy (MMD) losses.

3.1 The assignment and Monge problems

OPTIMAL ASSIGNMENT PROBLEM. Fix two integers $M, N \geq 1$ and denote two datasets by $\mathbf{x} := \{x_1, \dots, x_M\} \subset \mathcal{X}$ and $\mathbf{y} := \{y_1, \dots, y_N\} \subset \mathcal{Y}$ where $\mathcal{X}, \mathcal{Y}$ are two metric spaces. Let $[\![M]\!] := \{1, \dots, M\}$ be the set of all positive integers up to M, and consider a cost matrix $C(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+^{M \times N}$ where $(C(\mathbf{x}, \mathbf{y}))_{m,n}$ represents the cost of moving a unit of mass from x_m to y_n. Assuming $M = N$, the optimal assignment problem consists of finding a bijective function $\sigma : [\![M]\!] \to [\![M]\!]$ such that the total cost $\sum_{m \in [\![M]\!]} (C(\mathbf{x}, \mathbf{y}))_{m,\sigma(m)}$ is minimized (see Figure 3.1). A naive solution is to evaluate the total cost of $M!$ permutations of M elements. However, $M!$ is huge even for small M so this may be very inefficient. Although we can use either the techniques of Linear Programming or the transportation method to solve the assignment problem, the Hungarian method developed by Kuhn (1955) is much faster and efficient with complexity $O(M^3)$ in the worst case.

MONGE PROBLEM. A generalization of optimal assignment problem, known as the Monge problem, was introduced by the French mathematician Monge (1781) as follows: a worker must find the "best" way to transport a certain quantity of soil from the ground to places where it should be used in a construction. Assume that the source and target places are

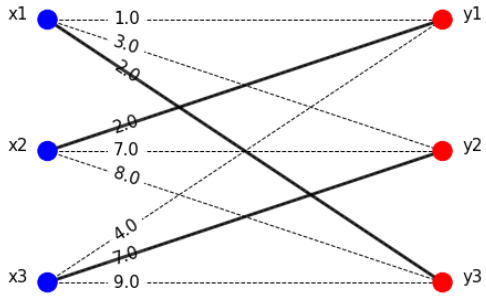

Figure 3.1 – An assignment of x_1, x_2, x_3 to y_1, y_2, y_3 with the cost matrix $C = \begin{pmatrix} 1 & 3 & 2 \\ 2 & 7 & 8 \\ 4 & 7 & 9 \end{pmatrix}$ and the permutation $\sigma : 1 \to 3, 2 \to 1, 3 \to 2$ represented by the solid lines.

known and the transportation cost to move a unit of mass between two points is known as well. The goal is to determine the destination to which a source point should be transported so that the total cost is minimal. This problem can be stated equivalently as follows. Denote $\Omega_d := \{a \in (\mathbb{R}_+)^d | \sum_{i \in [\![d]\!]} a_i = 1\}$ the $(d-1)-$dimensional simplex. For any $(a, b) \in \Omega_M \times \Omega_N$, let $\alpha := \sum_{m \in [\![M]\!]} a_m \delta_{x_m}$, $\beta := \sum_{n \in [\![N]\!]} b_n \delta_{y_n}$ be two weighted empirical measure attached to $\mathbf{x}$ and $\mathbf{y}$. Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ define the transportation cost to move a unit of mass from x_m to y_n, the Monge problem consists in solving

$$\min_{T \in \mathcal{T}} \sum_{m \in [\![M]\!]} c(x_m, T(x_m)), \tag{3.1}$$

where $\mathcal{T} := \{T : \mathbf{x} \to \mathbf{y} | b_n = \sum_{m : T(x_m) = y_n} a_m\}$, the so-called feasible set, is the set of all mappings that associates each point x_m to a single point y_n and such that the mass conservation constraints are met. Note that the mapping T between two finite sets can be represented in a straightforward way by an assignment $\sigma : [\![M]\!] \to [\![N]\!]$ where $\sigma(m) = n$ if and only if $T(x_m) = y_n$ and the constraints are equivalent to $\sum_{m \in \sigma^{-1}(n)} a_m = b_n$. When $M = N$ and the two measures are uniform, i.e. $\alpha := \frac{1}{M} \sum_{m \in [\![M]\!]} \delta_{x_m}, \beta := \frac{1}{M} \sum_{n \in [\![M]\!]} \delta_{y_n}$, then the conservation constraints induce that T is a bijection, such that $T(x_m) = y_{\sigma(n)}$ and the Monge problem corresponds to the optimal assignment problem with the cost matrix $(C(\mathbf{x}, \mathbf{y}))_{m,n} = c(x_m, y_n)$. Note that the set $\mathcal{T}$ may be empty if the two measures α and β are incompatible, for example if $M < N$ or $\sum_{m \in [\![M]\!]} a_m \neq \sum_{n \in [\![N]\!]} b_n$ so that the Monge problem may not have a solution. In case a solution exists, it is very difficult and costly to solve this problem.

3.2 The Kantorovich relaxation

We have shown that the assignment problem is a special case of the Monge problem when the two measures are attached to two sets of the same size and are uniform. Also, the Monge problem allows to consider two arbitrary measures and to assign several source points to a target point. However, both problems are hard to solve in practice.
Much later after its introduction, the Monge problem was rediscovered by a Russian mathematician, Loenid Vitaliyevich Kantorovich, motivated by an economic problem (see Kantorovich, 1942). He proposed an ingenious idea that allows to split the mass of each source point and move them to several target points. Therefore, Kantorovich formulation consists

in solving, in place of a map T, a probabilistic matrix P where P_{mn} describes the amount of mass moved from x_m to y_n. This coupling matrix satisfies the mass conservation constraints, i.e., the sums of rows and columns should be equal to a and b, respectively. Formally, the set of admissible couplings is defined by

$$\Pi(a,b) := \{P \in (\mathbb{R}_+)^{M\times N} | P\mathbf{1}_N = a, P^\top \mathbf{1}_M = b\}.$$

In fact, $\Pi(a,b)$ can be expressed as the set of the joint probability matrix over $(\mathbf{x},\mathbf{y})$ with marginal distributions w and w', respectively. Obviously, this set contains $a \times b$ so it is nonempty. Another benefit is the symmetric property in the sense that P is an element of $\Pi(a,b)$ if and only if $P^\top$ is an element of $\Pi(b,a)$ as well. Given a cost matrix $C(\mathbf{x},\mathbf{y}) \in (\mathbb{R}_+)^{M\times N}$, where $(C(\mathbf{x},\mathbf{y}))_{mn} = c(x_m, y_n)$, Kantorovich's formulation consists in solving

$$\mathrm{OT}_c(\alpha,\beta) := \min_{P\in\Pi(a,b)} \langle C(\mathbf{x},\mathbf{y}), P\rangle_F, \tag{3.2}$$

where $\langle C(\mathbf{x},\mathbf{y}), P\rangle_F := \sum_{(m,n)\in[\![M]\!]\times[\![N]\!]} (C(\mathbf{x},\mathbf{y}))_{mn} P_{mn}$ is the P-specific expected cost of transport from $\mathbf{x}$ to $\mathbf{y}$. In many cases, the notation $\mathrm{OT}_c(\alpha,\beta)$ is useful to indicate explicitly the dependence on the cost function c to define the cost matrix $C(\mathbf{x},\mathbf{y})$.

We generalize the definition (3.2) of OT_c to the case of arbitrary measures by first introducing some useful notations of functions and probability measures. Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures over $\mathcal{X}$. If $f : \mathcal{X} \to \mathcal{Y}$ is a continuous function, we define its associated push-forward operator $f_\sharp : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$, i.e., the push-forward measure $\beta = f_\sharp(\alpha)$ of $\alpha \in \mathcal{P}(\mathcal{X})$ that satisfies

$$\int_{\mathcal{Y}} h(y)\mathrm{d}\beta(y) = \int_{\mathcal{X}} h(f(x))\mathrm{d}\alpha(x), \quad \forall h \in \mathcal{C}(\mathcal{Y}),$$

where $\mathcal{C}(\mathcal{Y})$ is the space of continuous functions over $\mathcal{Y}$. In the general case, we consider, in place of coupling matrices, joint distributions over the product space $\mathcal{X}\times\mathcal{Y}$ that must satisfy the mass conservation constraints. Therefore, the set of admissible couplings can be defined as

$$\Pi(\alpha,\beta) := \{P \in \mathcal{P}(\mathcal{X},\mathcal{Y}) | \pi_{X\sharp}(P) = \alpha, \pi_{Y\sharp}(P) = \beta\},$$

where $\pi_{\mathcal{X}\sharp}$ and $\pi_{\mathcal{Y}\sharp}$ are the push-forward operators of the projections $\pi_{\mathcal{X}}(x,y) = x$ and $\pi_{\mathcal{Y}}(x,y) = y$, respectively. The Kantorovich problem now reads

$$\mathrm{OT}_c(\alpha,\beta) := \min_{P\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)dP(x,y). \tag{3.3}$$

This infinite-dimensional linear optimization over a space of measures have a solution under mild assumptions, for example that $(\mathcal{X},\mathcal{Y})$ are compact spaces and the cost function c is continuous. Furthermore, the OT loss can be rewritten as the expectation of $c(X,Y)$

$$\mathrm{OT}_c(\alpha,\beta) = \min_{(X,Y)} \{\mathbb{E}_{X,Y}(c(X,Y)) : X \sim \alpha, Y \sim \beta\}, \tag{3.4}$$

where (X,Y) is a couple of random variables with the joint law $P \in \Pi(\alpha,\beta)$ and fixed marginals α and β, respectively.

OPTIMAL TRANSPORT LOSS AS THE DISTANCE. One advantage of OT theory is that OT loss can be seen as a distance between probability measures if the cost function satisfies certain suitable properties. Specifically, when $\mathcal{X} = \mathcal{Y}$ is equipped with a metric d and $c = d^p$ with

$p \geq 1$, we define the p-Wasserstein distance $\mathcal{W}_p$ between two measures $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ by $\mathcal{W}_p(\alpha, \beta) := (\mathrm{OT}_{d^p}(\alpha, \beta))^{1/p}$. The distance $\mathcal{W}_1$ is also called the Kantorovich-Rubinstein distance in statistics or the Earth Mover's Distance in computer vision. To prove that the p-Wasserstein distance is a metric on a space of probability measures, we rely on the following classical result.

Lemma 3.1 (Gluing lemma (Berkes and Philipp, 1977))**.** *Let $(\mathcal{X}_i, \alpha_i), i = 1, 2, 3$, be Polish probability spaces. If (X_1, X_2) is a coupling of (α_1, α_2) and (Y_2, Y_3) is a coupling of (α_2, α_3), then one can construct a triple of random variables (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same law as (X_1, X_2) and (Z_2, Z_3) has the same law as (Y_2, Y_3).*

This lemma allows us to "glue together" two couplings having a common marginal: if $P_{1,2}$ stands for the law of (X_1, X_2) on $\mathcal{X}_1 \times \mathcal{X}_2$ and $P_{2,3}$ stands for the law of (X_2, X_3) on $\mathcal{X}_2 \times \mathcal{X}_3$ then one can "glue" $P_{1,2}$ and $P_{2,3}$ along their common marginal to obtain the joint law $P_{1,2,3}$. Using this lemma, we will prove the triangle inequality of $\mathcal{W}_p$.

Proposition 3.1 (adapted from Theorem 7.3 in (Villani and Society, 2003))**.** *The quantity $\mathcal{W}_p$ is a distance over $\mathcal{P}(\mathcal{X})$.*

Proof. Of course $\mathcal{W}_p$ is symmetric by symmetry of d^p. It it clear that $\mathcal{W}(\alpha, \alpha) = 0$ and $\mathcal{W}(\alpha, \beta) \geq 0, \forall \alpha, \beta \in \mathcal{P}(\mathcal{X})$. On the other hand, since d is a metric, it must satisfy $d(x, y) = 0$ iff $x = y$. Therefore, if $\mathcal{W}_p(\alpha, \beta) = 0$ it can only be that there exists a transportation plan entirely concentrated on the diagonal $(y = x)$ in $\mathcal{X} \times \mathcal{X}$, so that $\beta = id_\sharp \alpha = \alpha$.

All that remains to be proved is the triangle inequality. Let $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{P}(\mathcal{X})$ and let (X_1, X_2) be an optimal coupling of α_1 and α_2 and analogously for (Y_2, Y_3) with respect to α_2 and α_3. By the Gluing Lemma, there exists random variables (Z_1, Z_2, Z_3) such that $(Z_1, Z_2) \overset{\mathrm{d}}{=} (X_1, X_2)$ and $(Z_2, Z_3) \overset{\mathrm{d}}{=} (Y_2, Y_3)$. Clearly, (Z_1, Z_3) is a coupling of α_1 and α_3. Moreover, using in turn the optimality of $\mathcal{W}_p$, the triangle inequality of distance d then Minkowski's inequality, we obtain

$$\begin{aligned}
\mathcal{W}_p(\alpha_1, \alpha_3) &\leq (\mathbb{E}[d(Z_1, Z_3)^p])^{1/p} \\
&\leq (\mathbb{E}[(\mathrm{d}(Z_1, Z_2) + \mathrm{d}(Z_2, Z_3))^p])^{1/p} \\
&\leq (\mathbb{E}[\mathrm{d}(Z_1, Z_2)^p])^{1/p} + (\mathbb{E}[\mathrm{d}(Z_2, Z_3)^p])^{1/p} \\
&= \mathcal{W}_p(\alpha_1, \alpha_2) + \mathcal{W}_p(\alpha_2, \alpha_3).
\end{aligned}$$

So $\mathcal{W}_p$ satisfies the triangle inequality. This concludes the proof. □

The p-Wasserstein distance is an effective tool to compare measures because of its ability to capture their underlying geometry by relying on the cost function that encodes the metric of the space $\mathcal{X}$. Besides, the coupling matrix P provides a mapping from one measure to the other which can be of interest in domain adaptation (see Courty et al., 2017). However, solving the Kantorovich problem is not an easy task. In the discrete case, the Kantorovich problem can be solved either by using Orlin's program or by interior point methods both of which run in $O(M^3 \ln(M))$ operations (see Pele and Werman, 2009). Furthermore, OT suffers from the curse of dimensionality. These limitations have led to the neglect of OT in machine learning applications for a long time.

3.3 Entropic regularization

We introduce a family of numerical schemes to reduce the high computational complexity of Kantorovich's formalization of optimal transport. The idea (see Cuturi, 2013a) is to add an

entropy regularization term in Equation (3.2). We focus on the case of discrete measures. The discrete entropy of a coupling matrix is defined by

$$E(P) := -\sum_{(m,n)\in[\![M]\!]\times[\![N]\!]} P_{mn}(\log P_{mn} - 1),$$

with the convention $E(P) = -\infty$ if one of the elements P_{mn} is 0 or negative. By penalizing the entropy of the original problem, we obtain a regularized version of problem (3.2)

$$\mathrm{OT}_{\gamma,c}(\alpha,\beta) = \min_{P\in\Pi(a,b)} \left\{\langle C(\mathbf{x},\mathbf{y}), P\rangle_F - \gamma E(P)\right\}. \tag{3.5}$$

Since $P \mapsto E(P)$ is a 1-strongly concave function and $P \mapsto \langle C(\mathbf{x},\mathbf{y}), P\rangle_F$ is a linear function on domain $\Pi(a,b)$, the function $P \mapsto \langle C(\mathbf{x},\mathbf{y}), P\rangle_F - \gamma E(P)$ is γ-strongly convex. Therefore, Problem (3.5) has a unique optimal solution. Furthermore, the following proposition proves the convergence of the solution of that regularized optimal transport.

Proposition 3.2 (adapted from Proposition 4.1 in (Peyré and Cuturi, 2019))**.** *Let P_γ be the unique solution of* (3.5) *for $\gamma > 0$. Then P_γ converges to the solution with maximal entropy of* (3.2) *as γ tends to zero, namely*

$$P_\gamma \xrightarrow{\gamma\to 0} \arg\min_P \{-E(P) : P \in \Pi(a,b), \langle C(\boldsymbol{x},\boldsymbol{y}), P\rangle_F = \mathrm{OT}_c(a,b)\}. \tag{3.6}$$

So that, in particular

$$\mathrm{OT}_{\gamma,c}(a,b) \xrightarrow{\gamma\to 0} \mathrm{OT}_c(a,b). \tag{3.7}$$

Moreover, P_γ converges to the coupling with maximal entropy between two marginals a and b as γ tends to infinity, namely

$$P_\gamma \xrightarrow{\gamma\to\infty} a\otimes b = ab^\top = (a_m b_n)_{m,n}. \tag{3.8}$$

Proof. Let $(\gamma_\ell)_{\ell\geq 0}$ be a strictly positive sequence converging to zero. We denote P_ℓ the solution of (3.5) for $\gamma = \gamma_\ell$. The set $\Pi(a,b)$ is compact because it is closed and bounded. So there exists a subsequence (P_{ℓ_k}) and $P^\star \in \Pi(a,b)$ such that P_{ℓ_k} converges to $P^\star$. Let $\hat{P}$ be a solution of (3.2). By the definitions of $\hat{P}$ and P_{ℓ_k}, we get

$$0 \leq \langle C, P_{\ell_k}\rangle - \langle C, \hat{P}\rangle \leq \gamma_{\ell_k}\left(E(P_{\ell_k}) - E(\hat{P})\right). \tag{3.9}$$

Since E is continuous, taking the limit $\ell_k \to +\infty$ in this expression show that $\langle C, P^\star\rangle = \langle C, \hat{P}\rangle$. Therefore, $P^\star$ is also a solution of (3.2). Moreover, dividing by γ_{ℓ_k} in (3.9) and taking the limit, we obtain $E(\hat{P}) \leq E(P^\star)$, which means that $P^\star$ is a solution of (3.6). Because of the strict convexity of $-E$, the problem

$$\min_P \{-E(P) : P \in \Pi(a,b), \langle C(\mathbf{x},\mathbf{y}), P\rangle_F = \mathrm{OT}_c(a,b)\}$$

has a unique solution. This shows (3.6) and (3.7).

Similarly, let $(\gamma_\ell)_{\ell\geq 0}$ be a strictly positive sequence converging to infinity. We denote P_ℓ the solution of (3.6) for $\gamma = \gamma_\ell$. Then there exists a subsequence (P_{ℓ_k}) and $P_\infty \in \Pi(a,b)$ such that $P_{\ell_k} \to P_\infty$. It is straightforward to show that the problem

$$\min_{P\in\Pi(a,b)} -E(P)$$

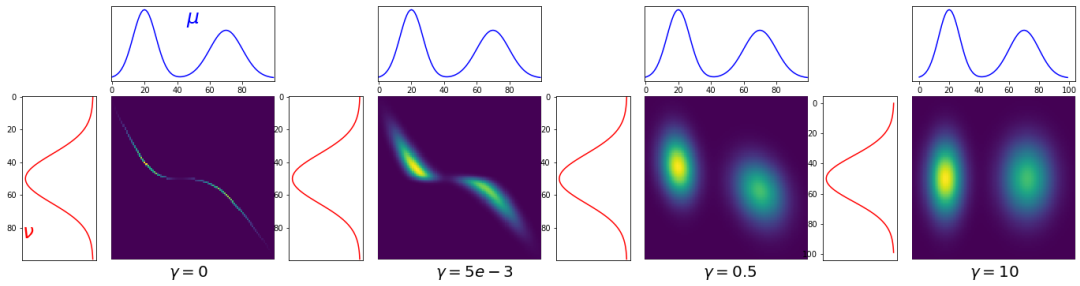

Figure 3.2 – Effect of the entropic regularization parameter γ on the optimal coupling P between two $1D$ probability distributions. As γ increases the coupling tends to blur and converges to the marginals' product coupling.

has the unique solution $\bar{P} = a \otimes b$. By the definitions of $\bar{P}$ and P_{ℓ_k}, we get

$$0 \leq E(\bar{P}) - E(P_{\ell_k}) \leq \frac{1}{\gamma_{\ell_k}} \left(\langle C, \bar{P} \rangle - \langle C, P_{\ell_k} \rangle \right).$$

Taking the limit $k \to +\infty$ in this expression shows that $E(\bar{P}) = E(P_\infty)$, which means that $P_\infty = a \otimes b$. This completes the proof. □

As stated in formulas (3.6) and (3.8), the convergence of optimal transport matrix depends on the regularization parameter, which is illustrated in Figures 3.2. When γ is small, the OT matrix becomes more sparse, in the sense of having few entries larger than a threshold and many zero entries. In contrast, when γ is very large, the OT matrix becomes blurry.

Using the Lagrangian duality, we show that the solution of (3.5) has a specific form, which can be parameterized using only $N + M$ variables.

Proposition 3.3 (adapted from Proposition 4.3 in (Peyré and Cuturi, 2019))**.** *The solution to* (3.5) *is unique and has the form*

$$P^\star = \mathrm{diag}(u) K \, \mathrm{diag}(v) \tag{3.10}$$

where $K = \mathrm{e}^{-\frac{C}{\gamma}}$ *is the Gibbs kernel associated to the cost matrix* C *and* $u \in (\mathbb{R}_+^*)^M, v \in (\mathbb{R}_+^*)^N$ *are two (unknown) scaling variables.*

Proof. The Lagrangian with respect to (3.5) is

$$\mathcal{L}(P, \mathbf{f}, \mathbf{g}) := \langle P, C \rangle - \gamma E(P) - \langle \mathbf{f}, P\mathbf{1}_N - a \rangle - \langle \mathbf{g}, P^\top \mathbf{1}_M - b \rangle,$$

where $\mathbf{f} \in \mathbb{R}_+^M$ and $\mathbf{g} \in \mathbb{R}_+^N$. Now, let us calculate the gradient

$$\frac{\partial \mathcal{L}(P, \mathbf{f}, \mathbf{g})}{\partial P_{m,n}} = C_{m,n} + \gamma \log(P_{m,n}) - (\mathbf{f}_m + \mathbf{g}_n),$$

and set it equal to 0, which implies that $P_{m,n} = \mathrm{e}^{\mathbf{f}_m/\gamma} \mathrm{e}^{-C_{m,n}/\gamma} \mathrm{e}^{\mathbf{g}_n/\gamma}$. Therefore, we obtain the optimal solution as (3.10) by using the notation $u = (\mathrm{e}^{\mathbf{f}_m/\gamma})_{m \in [\![M]\!]}$ and $v = (\mathrm{e}^{\mathbf{g}_n}/\gamma)_{n \in [\![N]\!]}$. □

The factorization of the OT matrix $P^\star$ allows us to easily solve that problem by finding two nonnegative vectors (u, v). The two conservation constraints can be expressed as the following equations

$$\operatorname{diag}(u) K \operatorname{diag}(v)\, \mathbf{1}_N = a \quad \text{and} \quad \operatorname{diag}(v) K^\top \operatorname{diag}(v)\, \mathbf{1}_M = b.$$

Since $\operatorname{diag}(v)\, \mathbf{1}_M = v$ and $\operatorname{diag}(u)\, \mathbf{1}_N = u$, we simplify those equations into an equivalent form

$$u \odot (Kv) = a \quad \text{and } v \odot K^\top u = b, \tag{3.11}$$

where $\odot$ denotes the component-wise multiplication of vectors. This problem, the so-called classical matrix scaling problem, can be solved through an iterative method which alternately normalizes u and v to satisfy the left and right-hand sides of Equation (3.11). More specifically, initialized with any positive vector $v^{(0)} = \mathbf{1}_N$, we implement two updates in each iteration of the procedure known as the Sinkhorn's algorithm

$$u^{(\ell+1)} := \frac{a}{Kv^{(\ell)}} \quad \text{and } v^{(\ell+1)} := \frac{b}{K^\top u^{(\ell+1)}} \tag{3.12}$$

where the division operator between two vectors is to be understood element-wise. Now we present an elementary proof of linear convergence of the iterations by using the Hilbert projective metric on $(\mathbb{R}_+^*)^d$.

Definition 3.1. *The Hilbert projective metric on $(\mathbb{R}_+^*)^d$ is defined by*

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') := \log \max \left\{ \tfrac{x_i x'_j}{x'_i x_j} : i, j \in [\![d]\!] \right\}.$$

We will use the following properties (see Birkhoff, 1957):

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = \|\log(x) - \log(x')\|_{\text{var}}; \tag{3.13}$$

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = d_{\mathcal{H}}(x/x', \mathbf{1}_d) = d_{\mathcal{H}}(\mathbf{1}_d/x', \mathbf{1}_d/x); \tag{3.14}$$

$$\forall K \in (\mathbb{R}_+^*)^{d \times d'}, \forall x, x' \in (\mathbb{R}_+^*)^{d'}, d_{\mathcal{H}}(Kx, Kx') \le \lambda(K) d_{\mathcal{H}}(x, x'), \tag{3.15}$$

where $\|x\|_{\text{var}} := \max\{x_i : i \in [\![d]\!]\} - \min\{x_i : i \in [\![d]\!]\}$ is the variation seminorm and $\lambda(K) := \frac{\sqrt{\eta(K)}-1}{\sqrt{\eta(K)}+1} < 1$ with $\eta(K) := \max\left\{ \frac{K_{i,k} K_{j,\ell}}{K_{j,k} K_{i,\ell}} : i, j \in [\![d]\!], k, \ell \in [\![d']\!] \right\}$. We have the following convergence theorem.

Theorem 3.1 (adapted from Theorem 4.2 in (Peyré and Cuturi, 2019))**.** *We have $(u^{(\ell)}, v^{(\ell)}) \to (u^\star, v^\star)$ and*

$$d_{\mathcal{H}}(u^{(\ell)}, u^\star) = O\big(\lambda(K)^{2\ell}\big), \quad d_{\mathcal{H}}(v^{(\ell)}, v^\star) = O\big(\lambda(K)^{2\ell}\big), \tag{3.16}$$

where $u^\star, v^\star$ are the optimal solutions. Furthermore,

$$d_{\mathcal{H}}(u^{(\ell)}, u^\star) \le \frac{d_{\mathcal{H}}\big(P^{(\ell)} \mathbf{1}_M, a\big)}{1 - \lambda(K)^2}, \tag{3.17}$$

$$d_{\mathcal{H}}(v^{(\ell)}, v^\star) \le \frac{d_{\mathcal{H}}\big((P^{(\ell)})^\top \mathbf{1}_N, b\big)}{1 - \lambda(K)^2}, \tag{3.18}$$

where $P^{(\ell)} := \operatorname{diag}(u^{(\ell)}) K \operatorname{diag}(v^{(\ell)})$. Last, we have

$$||\log(P^{(\ell)}) - \log(P^\star)||_{\max} \le d_{\mathcal{H}}(u^{(\ell)}, u^\star) + d_{\mathcal{H}}(v^{(\ell)}, v^\star), \tag{3.19}$$

where $P^\star$ is the unique solution of (3.5) *and $||P||_{\max} := \max\{P_{m,n} : m \in [\![M]\!], n \in [\![N]\!]\}$.*

Proof. Using (3.14) and (3.15), we get

$$\begin{aligned} d_{\mathcal{H}}(u^{(\ell+1)}, u^\star) &= d_{\mathcal{H}}\left(\frac{a}{Kv^{(\ell)}}, \frac{a}{Kv^\star}\right) \\ &= d_{\mathcal{H}}(Kv^{(\ell)}, Kv^\star) \leq \lambda(K) d_{\mathcal{H}}(v^{(\ell)}, v^\star). \end{aligned} \tag{3.20}$$

Using the fact that $\lambda(K^\top) = \lambda(K)$, we get in the same manner

$$\begin{aligned} d_{\mathcal{H}}(v^{(\ell)}, v^\star) &= d_{\mathcal{H}}\left(\frac{b}{K^\top u^{(\ell)}}, \frac{b}{K^\top u^\star}\right) \\ &= d_{\mathcal{H}}(K^\top u^{(\ell)}, K^\top u^\star) \\ &\leq \lambda(K^\top) d_{\mathcal{H}}(u^{(\ell)}, u^\star) = \lambda(K) d_{\mathcal{H}}(u^{(\ell)}, u^\star). \end{aligned} \tag{3.21}$$

The inequalities (3.20) and (3.21) imply that

$$d_{\mathcal{H}}(u^{(\ell+1)}, u^\star) \leq (\lambda(K))^2 d_{\mathcal{H}}(u^{(\ell)}, u^\star).$$

That is equivalent to the left-hand side of equation (3.16). We obtain the right-hand side of equation (3.16) in the same manner. Now, by invoking the triangle inequality and both equations (3.14) and (3.15), we get

$$\begin{aligned} d_{\mathcal{H}}(u^{(\ell)}, u^\star) &\leq d_{\mathcal{H}}(u^{(\ell+1)}, u^{(\ell)}) + d_{\mathcal{H}}(u^{(\ell+1)}, u^\star) \\ &\leq d_{\mathcal{H}}\left(\frac{a}{Kv^{(\ell)}}, u^{(\ell)}\right) + \lambda(K)^2 d_{\mathcal{H}}(u^{(\ell)}, u^\star) \\ &= d_{\mathcal{H}}\left(a, u^\ell \odot (Kv^{(\ell)})\right) + \lambda(K)^2 d_{\mathcal{H}}(u^{(\ell)}, u^\star). \end{aligned}$$

The above inequality and the fact that $u^\ell \odot (Kv^{(\ell)}) = P^{(\ell)}\mathbf{1}_M$ imply (3.17). Equation (3.18) can be proved in an analogous way. (3.19) follows from (Franklin and Lorenz, 1989, Lemma 3). □

The bound (3.17) and (3.18) suggest that we can implement the stopping criteria based on the marginal constraint violation, for instance $||P^{(\ell)}\mathbf{1}_M - a||_1$ and $||P^{(\ell)^\top}\mathbf{1}_N - b||_1$, to monitor the convergence. Furthermore, formula (3.16) states that the variable $(u^{(\ell)}, v^{(\ell)})$ converges linearly for the Hilbert metric. By (3.13), the dual variable $(\mathbf{f}^{(\ell)}, \mathbf{g}^{(\ell)}) = (\gamma \ln(u^{(\ell)}), \gamma \ln(v^{(\ell)}))$ converges linearly for the variation seminorm $||\cdot||_{\text{var}}$. Therefore, the convergence of Sinkhorn's algorithm deteriorates as γ tends to zero.

In practice, the Sinkhorn's algorithm will fail for small values of γ because of the division by zero in (3.12). In fact, the elements of the kernel $K = e^{-C/\gamma}$ vanish rapidly and become null in memory by overflow error. To address this problem, we will use a log-sum-exp stabilization trick (see Schmitzer, 2019) which allows to numerically run the algorithm at small regularizations and reduces the number of required iterations.

3.4 Sinkhorn loss

Due to the addition of entropy term, $\text{OT}_\gamma(\alpha, \alpha)$ is no longer zero and OT_γ suffers from biased sample gradients (see Bellemare et al., 2017). Following (Genevay et al., 2018), the problem is solved by considering instead the Sinkhorn divergence

$$\mathcal{S}_{\gamma,c}(\alpha, \beta) := 2\text{OT}_{\gamma,c}(\alpha, \beta) - \text{OT}_{\gamma,c}(\alpha, \alpha) - \text{OT}_{\gamma,c}(\beta, \beta).$$

The Sinkhorn divergence has many appealing properties that make it a useful tool in machine learning. It is positive, symmetric, convex and metrizes convergence of measures (see Feydy et al., 2019b). Furthermore, Sinkhorn divergences, based on regularized OT, interpolate between OT and MMD. This allows to leverage the geometry of OT on the one hand and the properties of MMD (favorable high-dimensional sample complexity and sensitivity to differences in both location and shape of distributions that makes MMD a versatile tool for detecting various types of distributional discrepancies) one the other hand, which comes with unbiased gradient estimates (see Genevay et al., 2018).

4

Optimal transport-based machine learning to match specific patterns: application to the detection of molecular regulation patterns in omics data

In this chapter, we present several algorithms designed to learn a pattern of correspondence between two data sets in situations where it is desirable to match elements that exhibit a relationship belonging to a known parametric model. In the motivating case study, the challenge is to better understand micro-RNA (miRNA) regulation in the striatum of Huntington's disease (HD) model mice. The two data sets contain miRNA and messenger-RNA (mRNA) data, respectively, each data point consisting in a multi-dimensional profile. The strong biological hypothesis is that if a miRNA induces the degradation of a target mRNA or blocks its translation into proteins, or both, then the profile of the former, say y, should be similar to minus the profile of the latter, say $-x$. We consider a loosened hypothesis stating that y is then similar to $t(x)$ where t is an affine transformation in a parametric class that includes minus the identity and translates expert knowledge about the experiment that yielded the data.

The algorithms unfold in two stages. During the first stage, an optimal transport plan P and an optimal affine transformation are learned, using the Sinkhorn-Knopp algorithm and a mini-batch gradient descent. During the second stage, P is exploited to derive either several co-clusters or several sets of matched elements.

We share codes that implement our algorithms. A simulation study illustrates how they work and perform. A brief summary of the real data application in the motivating case-study further illustrates the applicability and interest of the algorithms.

This chapter is based on (Nguyen et al., 2023), joint work with W. Harchaoui, L. Mégret, C. Mendoza, O. Bouaziz, C. Neri, A. Chambaz. The project is funded by Université Paris

Cité thanks to a Ph.D. fellowship granted by Domaine d'Intérêt Majeur Math Innov (Région Île-de-France and Fondation Sciences Mathématiques de Paris).

> My main contribution has consisted in developing the methodology, formally and computationally, and performing the data analysis based on insights from Lucile Mégret and Chistian Neri on how mutant huntingtin may significantly influence expression patterns across CAG repeat alleles and age points in the brain of HD mice. The corresponding article has been submitted to the international *Journal of the Royal Statistical Society: Series C* (JRSS-C). A minor revision has been requested.

4.1 Introduction

The analysis of numerous omics data is a challenging task in biological research (Benayoun et al., 2019) and disease research (Langfelder et al., 2016; Maniatis et al., 2019). In disease research, omics data are increasingly available for the analysis of molecular pathology. This is notably illustrated by research on Huntington's Disease (HD): messenger-RNA (mRNA), micro-RNA (miRNA), protein data collectively quantifying several layers of molecular regulation in the brain of HD model knock-in mice (Langfelder et al., 2016, 2018) now compose one of the largest data set available to date to understand how neurodegenerative processes may work on a systems level. The data set is publicly available through the database repository Gene Expression Omnibus (GEO) and the HDinHD portal.

Encouraged by the promising findings of (Mégret et al., 2020), our ultimate goal is to shed light on the interaction between mRNAs and miRNAs based on data collected in the striatum (a brain region) of HD model knock-in mice (Langfelder et al., 2016, 2018). Each data point takes the form of multi-dimensional profile. The strong biological hypothesis is that if a miRNA induces the degradation of a target mRNA or blocks its translation into proteins, or both, then the profile of the former, say y, should be similar to minus the profile of the latter, say $-x$. We relax the hypothesis and consider that y is similar to $\theta(x)$ where θ is an affine transformation in a parametric class Θ that includes minus the identity and whose definition translates expert knowledge about the experiment that yields the data. Our study straightforwardly extends to the case that the relationship is known to belong to any parametric model. In order to identify groups of mRNAs and miRNAs that interact, we develop a co-clustering algorithm and a matching algorithm based on optimal transport (Peyré and Cuturi, 2019), spectral and block co-clustering, and a matching procedure tailored to our needs.

Spectral co-clustering (Dhillon, 2001) and block clustering (Brault et al., 2014; Govaert and Nadif, 2010) are two ways among many others to carry out co-clustering, an unsupervised learning task to cluster simultaneously the rows and columns of a matrix in order to obtain homogeneous blocks. There are many efficient approaches to solving the problem, often characterized as model-based or metric-based methods (Pontes et al., 2015).

In an enlightening article, Nazarov and Kreis (2021) review a variety of computational approaches to study how miRNAs "come together to regulate the expression of a gene or a group of genes". They identify three different families of methods: data-driven methods based on similarities, data-driven methods based on matrix factorization, and hybrid methods. Our algorithms belong to the first family. In view of (Nazarov and Kreis, 2021, Section 2.5 and Fig. 2), we do not rely on the standard similarity measures (Pearson and Spearman correlation coefficients; cosine similarity; mutual information) to define our similarity matrix but, instead, use optimal transport to derive it. Moreover, as in canonical

correlation analysis, we do not compare the raw mRNA and miRNA profiles x, y but, instead, we compare a data-driven transformation $\theta(x)$ and y, where θ is an affine transformation of x. Finally, as explained by Nazarov and Kreis (2021), our algorithms cannot discriminate between true interactions and fake interactions originating from common hidden regulators such as transcription factors. It is necessary to conduct a further biological analysis to identify the relevant findings.

The rest of the article is organized as follows. Section 4.2 describes the data we use. Section 4.3 presents a modicum of optimal transport theory. Section 4.4 introduces our algorithms. Section 4.5 evaluates the performances of the algorithms in various simulation settings. Section 4.6 illustrates the real data application.

4.2 Data

4.2.1 Presentation

The data analyzed herein cover RNA-seq data obtained in the striatum of the allelic series of HD knock-in mice (poly Q lengths: Q20, Q80, Q92, Q111, Q140, Q175) at 2-month, 6-month and 10-month of age. After preprocessing (Mégret et al., 2020, Methods section), the final data set consists of $M = 13,616$ mRNA profiles, $X := \{x_1, \dots, x_M\} \subset \mathbb{R}^d$, and of $N = 1,143$ miRNA profiles, $Y := \{y_1, \dots, y_N\} \subset \mathbb{R}^d$ with $d = 15$.

Informally, we look for couples $(m, n) \in [\![M]\!] \times [\![N]\!] := \{1, \dots, M\} \times \{1, \dots, N\}$ such that the nth miRNA induces the degradation of the mth mRNA or blocks its translation into proteins, or both. We are guided by the strong biological hypothesis that, if that is the case, then the profile y_n of the former is similar to minus the profile x_m of the latter – then x_m and y_n exhibit what we call a mirroring relationship. Of note, it is expected that a single miRNA can target several mRNAs.

The actual mirroring relationships can be more or less acute, for instance because of threshold effects, or of multiple miRNAs targeting the same mRNA, or of a single miRNA targeting several mRNAs. Therefore, instead of rigidly using comparisons between $-x_m$ and y_n, our algorithms will learn from the data a relevant transformation $\theta \in \Theta$ (in a parametric class Θ of transformations that includes minus the identity) and use comparisons between $\theta(x_m)$ and y_n.

Figure 4.1 exhibits two profiles x_m and y_n that showcase a mirrored similarity. The corresponding miRNA and mRNA, Mir20b (which may inhibit cerebral ischemia-induced inflammation in rats (Zhao et al., 2019)) and the Aryl-Hydrocarbon Receptor Repressor (Ahrr), are believed to interact in the striatum of HD model knock-in mice (Mégret et al., 2020).

4.2.2 A brief data analysis

So as to give a sense of the distribution of the data, we propose two kinds of visual summaries. The first one uses Lloyd's k-means algorithm (Lloyd, 1982) to build synthetic profiles representing the real profiles $x_1, \dots, x_M$ on the one hand and $y_1, \dots, y_N$ on the other hand. The second one uses kernel density estimators of the j-th component of $x_1, \dots, x_M$ on the one hand and of $y_1, \dots, y_N$ on the other hand, for each $1 \leq j \leq d$.

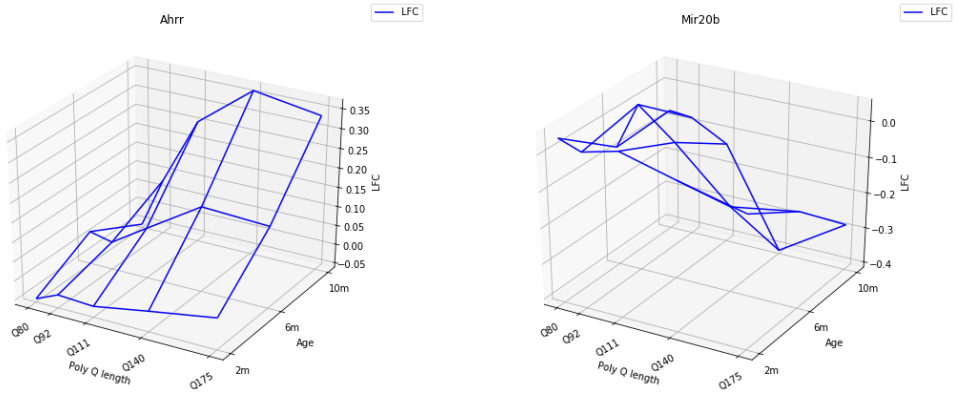

Figure 4.1 – Left: profile x_m of a mRNA (Ahrr). Right: profile y_n of a miRNA (Mir20b). It is believed that Mir20b targets Ahrr.

4.2.2.a Using k-means to cluster the mRNA and miRNA profiles

In Figure 4.2 we plot the synthetic mRNA profiles $\hat{x}_1, \ldots, \hat{x}_5$ of the 5 centroids obtained by running Lloyd's k-means algorithm on $x_1, \ldots, x_M$ with $k = 5$. Likewise, we plot in Figure 4.3 the synthetic miRNA profiles $\hat{y}_1, \ldots, \hat{y}_5$ of the 5 centroids obtained by running Lloyd's k-means algorithm on $y_1, \ldots, y_N$ with $k = 5$.

The 5 mRNA centroids correspond to 5319 ($\hat{x}_1$), 2097 ($\hat{x}_2$), 4688 ($\hat{x}_3$), 310 ($\hat{x}_4$) and 1202 ($\hat{x}_5$) mRNA profiles. The first and third centroids ($\hat{x}_1$ and $\hat{x}_3$), which represent 73% of the real mRNA profiles, are rather flat. The second and fourth centroids ($\hat{x}_2$ and $\hat{x}_4$), which represent 18% of the real mRNA profiles, are decreasing in poly Q length and age, in a more pronounced way for the latter than for the former. Finally, the fifth centroid ($\hat{x}_5$), which represents the remaining 9% of real mRNA profiles, is increasing in polyQ length and age.

The 5 miRNA centroids correspond to 872 ($\hat{y}_1$), 7 ($\hat{y}_2$), 80 ($\hat{y}_3$), 81 ($\hat{y}_4$) and 103 ($\hat{y}_5$) miRNA profiles. The first centroid ($\hat{y}_1$), which represents 76% of the real miRNA profiles, is rather flat. The second and fifth centroids ($\hat{y}_2$ and $\hat{y}_5$), which represent 10% of the real miRNA profiles, are increasing in poly Q length and age, in a more pronounced way for the former than for the latter. The fourth centroid ($\hat{y}_4$), which represents 7% of the real miRNA profiles, is decreasing in poly Q length and age. Finally, the third centroid ($\hat{y}_3$), which represents 7% of the real miRNA profiles, exhibits two peaks.

In Section 4.1, we stated the following biological hypothesis: if a miRNA induces the degradation of a target mRNA or blocks its translation into proteins, or both, then the profile of the former should be similar to minus the profile of the latter (a particular form of affine relationship). In view of this hypothesis, it is tempting to relate the synthetic miRNA profiles $\hat{y}_2$ and $\hat{y}_5$ to the synthetic mRNA profiles $\hat{x}_4$ and $\hat{x}_2$, respectively, and the synthetic miRNA profile $\hat{y}_4$ to the synthetic mRNA profile $\hat{x}_5$. Our objective is to identify groups of real mRNA and miRNA profiles that interact in this manner.

4.2.2.b Using kernel density estimators to study the marginal distributions of the mRNA and miRNA profiles

For each $1 \leq j \leq d$, we build the kernel density estimator of the j-th component of the mRNA profiles $x_1, \ldots, x_M$, using a Gaussian kernel and the default fine-tuning of the `density` function from the `stats` R-package (R Core Team, 2022), see Figure 4.4. We do the same

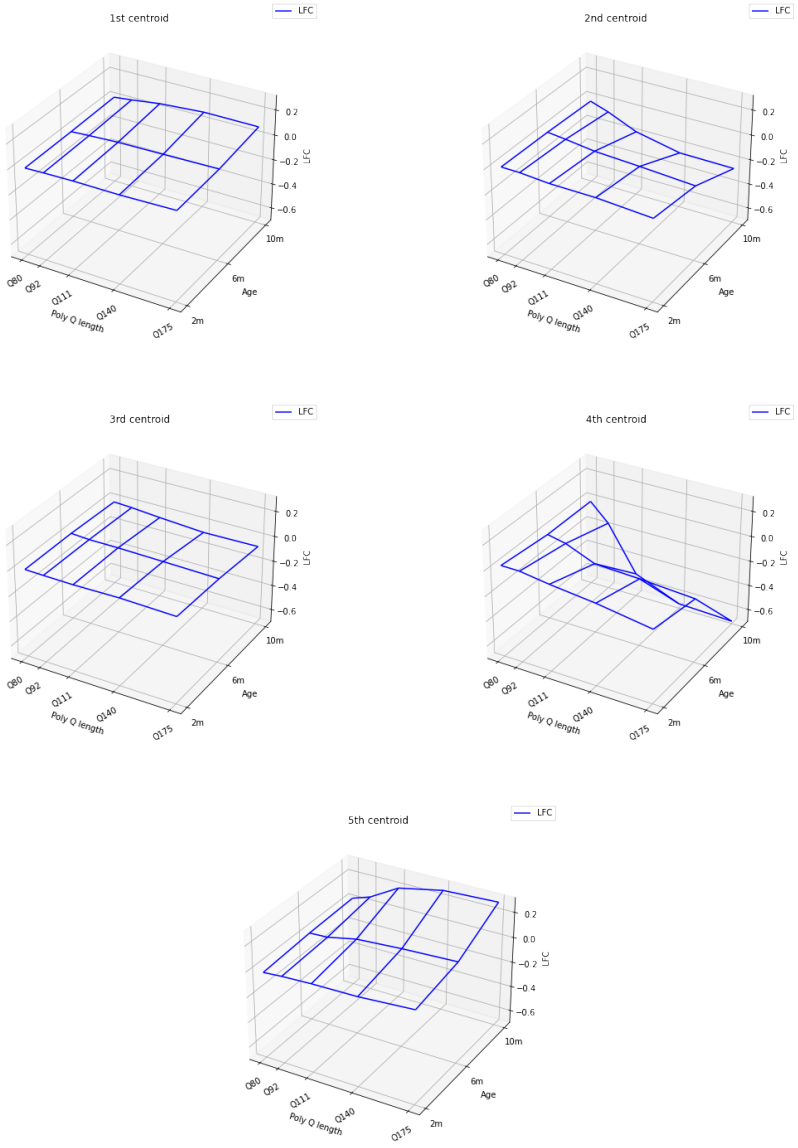

Figure 4.2 – Profiles $\hat{x}_1, \ldots, \hat{x}_5$ of the 5 centroids obtained by Lloyd's k-means algorithm on the mRNA profiles $x_1, \ldots, x_M$.

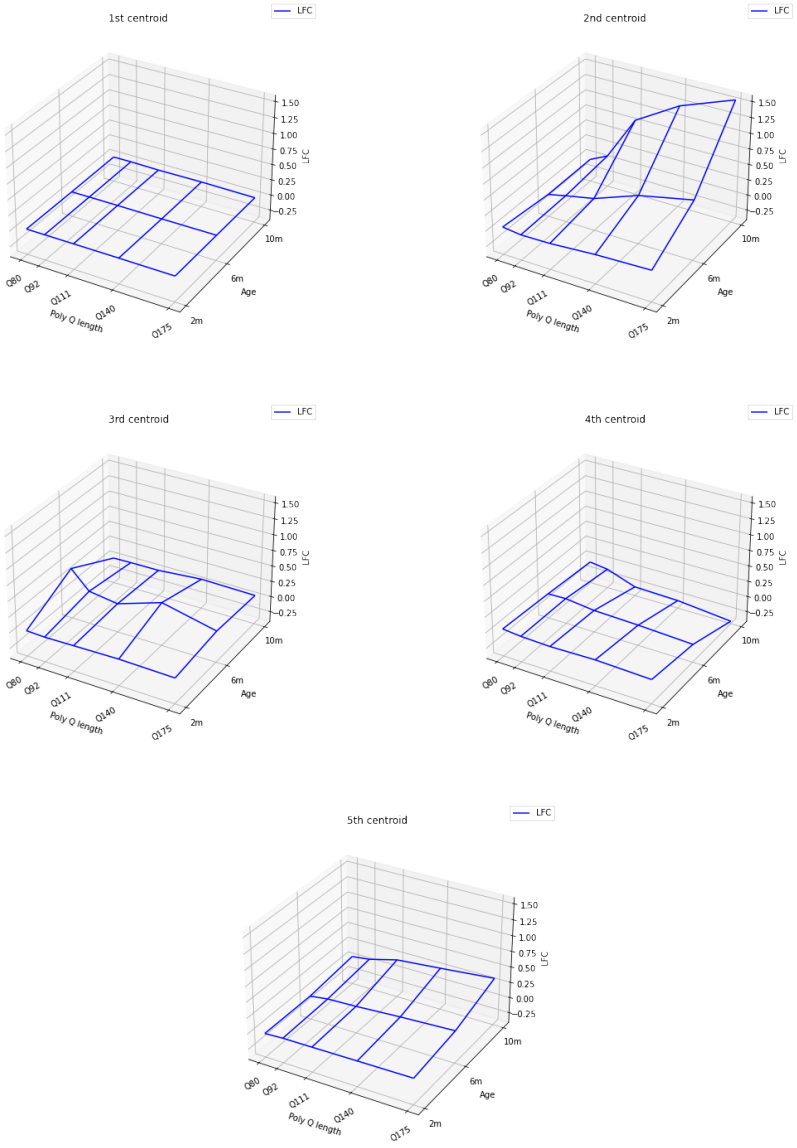

Figure 4.3 – Profiles $\hat{y}_1, \ldots, \hat{y}_5$ of the 5 centroids obtained by running Lloyd's k-means algorithm on the miRNA profiles $y_1, \ldots, y_N$.

poly Q length	Age 2	Age 6	Age 10
Q80	1	0.646	1.39
Q92	0.886	1.02	1.48
Q111	0.964	1.21	3.08
Q140	0.805	1.70	4.11
Q175	1.24	1.86	4.32

poly Q length	Age 2	Age 6	Age 10
Q80	1	2.35	1.03
Q92	0.516	1.06	0.956
Q111	0.655	0.722	2.15
Q140	0.698	1.92	2.72
Q175	0.588	1.80	3.34

Table 4.1 – For each level of poly Q length (Q80, Q92, Q111, Q140, Q175) and age (2, 6, 10 months) we computed the empirical standard deviation of mRNA (left) and miRNA (right) gene expressions, all normalized by the empirical standard deviation at poly Q length Q80 and 2 months of age (that is, by 0.0475 for mRNA and 0.0660 for miRNA).

for the miRNA profiles $y_1, \ldots, y_N$, see Figure 4.5. Both for mRNA and miRNA the kernel density estimates are systematically more concentrated around their means (all close to 0) than the corresponding Gaussian densities. Moreover, the kernel density estimates obtained from the M mRNA profiles are much smoother than those obtained from N miRNA profiles, a feature that could be simply explained by the fact that $M/N > 11$.

Table 4.1 reports, for each level of poly Q length (Q80, Q92, Q111, Q140, Q175) and age (2, 6, 10 months), the empirical standard deviation of mRNA (a) and miRNA (b) gene expressions, all normalized by the empirical standard deviation at poly Q length Q80 and 2 months of age (that is, by 0.0475 for mRNA and 0.0660 for miRNA). A clear pattern emerges from sub-Table 4.1 (a): except for poly Q length Q80, the poly Q length-specific empirical standard deviation increases as age increases. Likewise, except for age 2 months, the age-specific empirical standard deviation increases as poly Q length increases. On the contrary, no clear pattern emerges from sub-Table 4.1 (b) but the fact that, except for poly Q lengths Q80 and Q92, the poly Q length-specific empirical standard deviation increases as age increases. We do not comment on the empirical means because they are all very small compared to the corresponding empirical standard deviations.

4.3 Elements of optimal transport

Let $\Omega := \{\omega \in (\mathbb{R}_+)^M | \sum_{m\in[\![M]\!]} \omega_m = 1\}$ be the $(M-1)$-dimensional simplex and $\bar{\omega} := M^{-1}\mathbf{1}_M$, where $\mathbf{1}_M \in \mathbb{R}^M$ is the vector with all its entries equal to 1. For any $\omega \in \Omega$, define

$$\Pi(\omega) := \{P \in (\mathbb{R}_+)^{M\times N} | P\mathbf{1}_N = \omega, P^\top \mathbf{1}_M = N^{-1}\mathbf{1}_N\}$$

and let $\mu_X^\omega := \sum_{m\in[\![M]\!]} \omega_m \delta_{x_m}$, $\nu_Y := N^{-1}\sum_{n\in[\![N]\!]} \delta_{y_n}$ be the ω-weighted empirical measure attached to X and the empirical measure attached to Y. An element P of $\Pi(\omega)$ represents a joint law on $X \times Y$ with marginals μ_X^ω and ν_Y.

The celebrated Monge-Kantorovich problem (Peyré and Cuturi, 2019, Chapter 2) consists in finding a joint law over $X \times Y$ with marginals $\mu_X^{\bar{\omega}}$ and ν_Y that minimizes the expected cost of transport with respect to some cost function $c : X \times Y \to \mathbb{R}_+$. We focus on c given by $c(x, y) := \|x - y\|_2^2$ (the squared Euclidean norm in $\mathbb{R}^d$). Specifically, denoting $C_{X,Y} \in \mathbb{R}^{M\times N}$ the cost matrix given by $(C_{X,Y})_{mn} := c(x_m, y_n)$ for each $(m, n) \in [\![M]\!] \times [\![N]\!]$, the problem consists in solving $\min_{P\in\Pi(\bar{\omega})} \langle C_{X,Y}, P\rangle_F$ where $\langle C_{X,Y}, P\rangle_F := \sum_{(m,n)\in[\![M]\!]\times[\![N]\!]} (C_{X,Y})_{mn} P_{mn}$ is the P-specific expected cost of transport from X to Y.

It is well known that it is very rewarding from a computational viewpoint to consider a regularized version of the above problem (Peyré and Cuturi, 2019, Chapter 4). The penalty term is proportional to the discretized entropy of P, that is, to $E(P) :=$

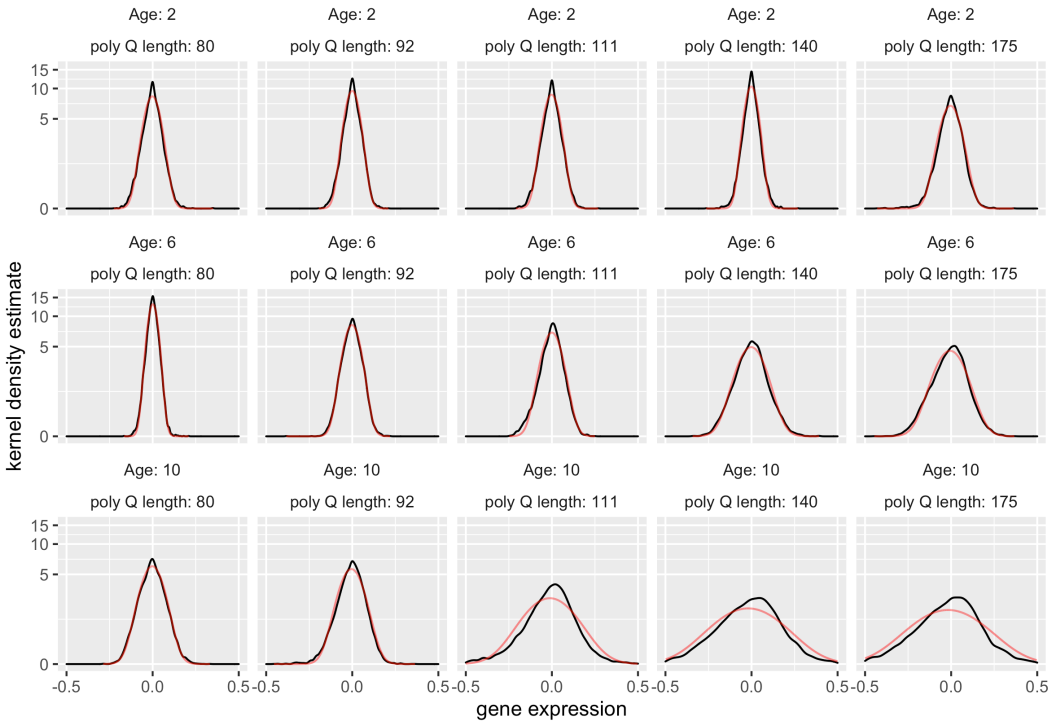

Figure 4.4 – In black, kernel density estimates of the densities of mRNA gene expression for each level of poly Q length (Q80, Q92, Q111, Q140, Q175) and age (2, 6, 10 months), zooming on the interval $[-0.5, 0.5]$ and using a $\log(1+\cdot)$-scale on the y-axis. In red, densities of the Gaussian laws with a mean and a variance equal to the empirical mean and variance computed in each stratum of data. Systematically, the kernel density estimates are more concentrated around their means than the corresponding Gaussian densities.

$-\sum_{(m,n)\in[\![M]\!]\times[\![N]\!]} P_{mn}(\log P_{mn} - 1)$. The regularized problem (presented here for any $\omega \in \Omega$ beyond the case $\omega = \bar{\omega}$) consists, for some user-supplied $\gamma > 0$, in finding P_γ that solves

$$\mathcal{W}_\gamma\left(\mu_X^\omega, \nu_Y\right) := \min_{P\in\Pi(\omega)} \left\{\langle C_{X,Y}, P\rangle_F - \gamma E(P)\right\}. \tag{4.1}$$

One of the advantages of entropic regularization is that one can solve (4.1) efficiently using the Sinkhorn-Knopp matrix scaling algorithm.

Finally, following (Genevay et al., 2018), we use $\mathcal{W}_\gamma$ to define the so called Sinkhorn loss between μ_X^ω (any $\omega \in \Omega$) and ν_Y as

$$\bar{\mathcal{W}}_\gamma\left(\mu_X^\omega, \nu_Y\right) := 2\,\mathcal{W}_\gamma\left(\mu_X^\omega, \nu_Y\right) - \mathcal{W}_\gamma\left(\mu_X^\omega, \mu_X^\omega\right) - \mathcal{W}_\gamma\left(\nu_Y, \nu_Y\right).$$

This loss interpolates between $\mathcal{W}_0\left(\mu_X^\omega, \nu_Y\right)$ and the maximum mean discrepancy of μ_X^ω relative to ν_Y (Genevay et al., 2018, Theorem 1). Paraphrasing the abstract of (Genevay et al., 2018), the interpolation allows to find "a sweet spot" leveraging the geometry of optimal transport and the favorable high-dimensional sample complexity of maximum mean discrepancy, which comes with unbiased gradient estimates.

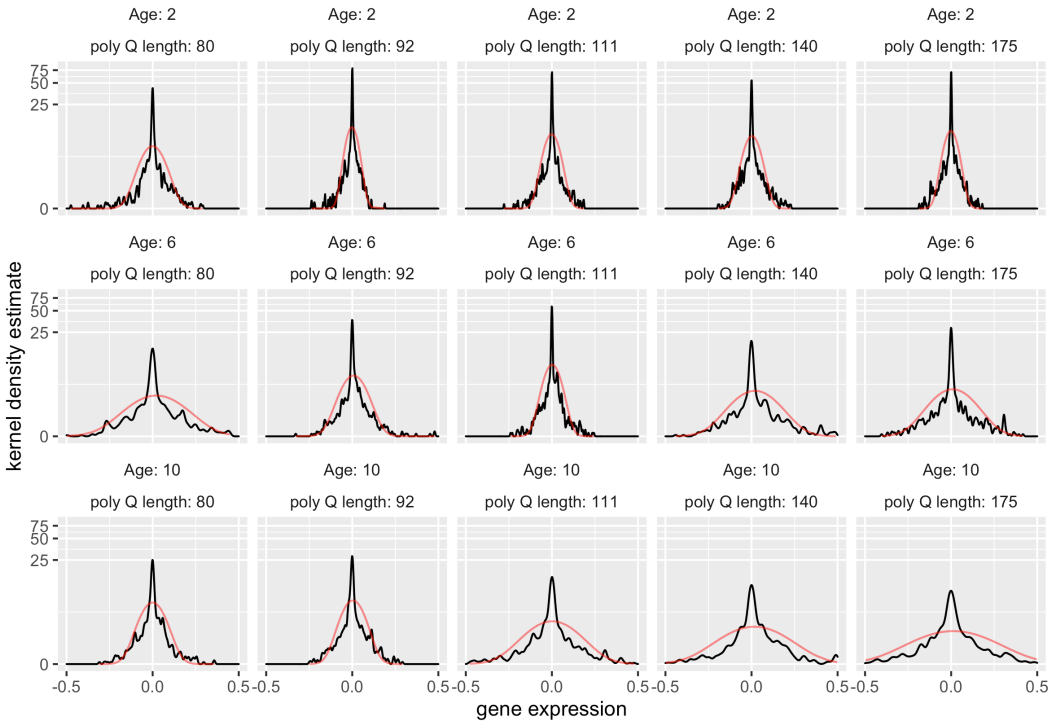

Figure 4.5 – In black, kernel density estimates of the densities of miRNA gene expression for each level of poly Q length (Q80, Q92, Q111, Q140, Q175) and age (2, 6, 10 months), zooming on the interval $[-0.5, 0.5]$ and using a $\log(1+\cdot)$-scale on the y-axis. In red, densities of the Gaussian laws with a mean and a variance equal to the empirical mean and variance computed in each stratum of data. Systematically, the kernel density estimates are more concentrated around their means than the corresponding Gaussian densities.

4.4 Optimal transport-based machine learning

In this section we introduce two co-clustering algorithms and one matching algorithm, all based on the solution of a master optimization program. The optimization program is presented in Section 4.4.1 and the algorithms are presented in Section 4.4.2.

4.4.1 Stage 1: the master optimization program and how to solve it

We introduce a parametric model Θ consisting of affine mappings $\theta : \mathbb{R}^d \to \mathbb{R}^d$ of the form $x \mapsto \theta(x) = \theta_1 x + \theta_2$, where $\theta_1 \in \mathbb{R}^{d\times d}$ and $\theta_2 \in \mathbb{R}^d$. The formal definition of Θ is given in Appendix 4.7. Each $\theta \in \Theta$ is a candidate to formalize the aforementioned mirroring relationship. The set Θ imposes constraints on the matrices θ_1, in particular that their diagonals are made of negative values. Of course, minus identity belongs to Θ. The parametrization is identifiable, in the sense that $\theta = \theta'$ implies $(\theta_1, \theta_2) = (\theta'_1, \theta'_2)$. It is noteworthy that *any* identifiable, regular model Θ could be used. We focus on Θ as defined in Appendix 4.7 because of the application that we consider in Section 4.6 (and in Section 4.5).

By analogy with Section 4.3 we introduce, for any $\theta \in \Theta$, $\omega \in \Omega$ and $\gamma > 0$, $\theta(X) := \{\theta(x_1), \ldots, \theta(x_M)\}$ the image of X by θ; the ω-weighted empirical measure attached to $\theta(X)$,

$\mu^{\omega}_{\theta(X)} := \sum_{m\in[\![M]\!]} \omega_m \delta_{\theta(x_m)}$; the cost matrix $C_{\theta(X),Y}$ given by $(C_{\theta(X),Y})_{mn} := c(\theta(x_m), y_n)$ for each $(m,n) \in [\![M]\!] \times [\![N]\!]$; and

$$\mathcal{W}_\gamma \left(\mu^{\omega}_{\theta(X)}, \nu_Y\right) = \min_{P\in\Pi(\omega)} \left\{\langle C_{\theta(X),Y}, P\rangle_F - \gamma E(P)\right\} \tag{4.2}$$

where $\langle C_{\theta(X),Y}, P\rangle_F := \sum_{(m,n)\in[\![M]\!]\times[\![N]\!]} (C_{\theta(X),Y})_{mn} P_{mn}$ is the P-specific expected cost of transport from $\theta(X)$ to Y.

Fix arbitrarily $\omega \in \Omega$. The first program that we introduce is the ω-specific program

$$\min_{\theta\in\Theta} \bar{\mathcal{W}}_\gamma \left(\mu^{\omega}_{\theta(X)}, \nu_Y\right), \tag{4.3}$$

where we are interested in the minimizer $\hat{\theta}$ that solves (4.3) *and* in the optimal joint matrix $\hat{P} \in \Pi(\omega)$ that solves

$$\min_{P\in\Pi(\omega)} \left\{\langle C_{\hat{\theta}(X),Y}, P\rangle_F - \gamma E(P)\right\}.$$

In words, we look for an ω-specific optimal mirroring function $\hat{\theta}$ and its ω-specific optimal transport plan $\hat{P}$.

How to choose ω? We decide to optimize with respect to ω as well. This additional optimization is relevant because we do not expect to associate a y_n to every x_m eventually at the co-clustering stage. So, our master program is

$$\min_{\omega\in\Omega} \min_{\theta\in\Theta} \bar{\mathcal{W}}_\gamma \left(\mu^{\omega}_{\theta(X)}, \nu_Y\right), \tag{4.4}$$

where we are interested in the minimizer $(\hat{\omega}, \hat{\theta})$ *and* in the optimal matrix $\hat{P} \in \Pi(\hat{\omega})$ that solves

$$\min_{P\in\Pi(\hat{\omega})} \left\{\langle C_{\hat{\theta}(X),Y}, P\rangle_F - \gamma E(P)\right\}. \tag{4.5}$$

We propose to solve (4.4) iteratively by updating ω and then θ. At round t, given ω_{t}, we make one step of mini-batch gradient descent to derive $\theta_{\mathrm{t}+1}$ from θ_{t} (here, we notably rely on the Sinkhorn-Knopp algorithm). Given $\theta_{\mathrm{t}+1}$, $\omega_{\mathrm{t}+1}$ is chosen proportional to the vector in $(\mathbb{R}_+)^M$ whose mth component equals $h^{-1} \sum_{n\in[\![N]\!]} \varphi((y_n - \theta_{\mathrm{t}+1}(x_m))/h)$ where φ is the standard normal density and h is the arithmetic mean of the $c(y_n, y_{n'})$ for all $n \neq n' \in [\![N]\!]$. Eventually, once the final round T is completed, we compute $\tilde{P} \in \Pi(\omega_{\mathrm{T}})$ that solves

$$\min_{P\in\Pi(\omega_{\mathrm{T}})} \left\{\langle C_{\theta_{\mathrm{T}}(X),Y}, P\rangle_F - \gamma E(P)\right\}$$

(again, we rely on the Sinkhorn-Knopp algorithm).

The algorithm to solve (4.4) is summarized in Procedure 1. We have no guarantee that it converges. Note, however, that using the Sinkhorn-Knopp algorithm to solve (4.5) for a given $(\hat{\omega}, \hat{\theta})$ is known to converge (Peyré and Cuturi, 2019, Theorem 4.2).

In light of (Alvarez-Melis, 2019, Section 1.3, page 25), we inject problem-specific knowledge onto two of the three main components of the transportation problem: the representation spaces (via the mapping θ) and the marginal constraints (via the weight ω), leaving aside the cost function. Furthermore, we resort to mini-batch gradient descent because the algorithmic complexity prevents the direct computation using the whole data set. A theoretical analysis of this practice is proposed in (Fatras et al., 2020).

We can now exploit $\tilde{P}$ so as to derive relevant associations between mRNAs and miRNAs. We propose two approaches. On the one hand, the first approach outputs *bona fide* co-clusters. We expect that the co-clusters can associate many mRNAs with many miRNAs, thus making it difficult to interpret and analyze the results. On the other hand, the second approach rather *matches* each mRNA with at most k miRNAs and each miRNA with at most k' mRNAs (k and k' are user-supplied integers). Details follow.

4.4.2 Stage 2: co-clustering or matching

4.4.2.a Co-clustering.

To carry out the co-clustering task once $\tilde{P}$ has been derived, we propose to rely either on spectral co-clustering (we will use the acronym SCC) (Dhillon, 2001), applying it once or twice, or co-clustering based on latent block models (Govaert and Nadif, 2010). Of course, any other co-clustering algorithm could be used as well. Specifically, we develop the following algorithms (the acronym WTOT stands for weighted transformation optimal transport).

WTOT-SCC1. Algorithm WTOT-SCC1 applies SCC *once* to build *bona fide* co-clusters based on $\tilde{P}$. It is required to provide a number of clusters. We rely on a criterion involving graph modularity to learn from the data a relevant number of clusters (Ailem et al., 2016, Sections 2 and 4).

In our simulation study, we also consider algorithm WTOT-SCC1*, an oracular version of WTOT-SCC1 that benefits from relying on the *true* number of clusters. This allows to assess how relevant is the learned number of clusters in WTOT-SCC1.

WTOT-SCC2. Algorithm WTOT-SCC2 applies SCC *twice* to build *bona fide* co-clusters based on $\tilde{P}$. It proceeds in three successive steps.

- In step 1, WTOT-SCC2 applies SCC a first time to derive an initial co-clustering. A relevant number of co-clusters is learned as in WTOT-SCC1.
- In step 2, WTOT-SCC2 selects and removes some rows and columns corresponding to mRNAs and miRNAs that are deemed irrelevant. The selection is based on a numerical criterion computed from $\tilde{P}$. In our simulation study (Section 4.5), all rows and columns that correspond to diagonal blocks with a variance larger than two times the overall variance of $\tilde{P}$ are selected and removed. In the real data application (Section 4.6), we implement and use a different procedure.
- In step 3, WTOT-SCC2 applies SCC a second time, the relevant number of co-clusters being learned as in WTOT-SCC1.

In our simulation study, we also consider algorithm WTOT-SCC2*, an oracular version of WTOT-SCC2 that is provided the *true* number of clusters for its third step. This allows to assess how relevant is the sub-procedure to learn the numbers of clusters in WTOT-SCC2.

WTOT-BC. Algorithm WTOT-BC applies the so called block clustering algorithm to build *bona fide* co-clusters based on $\tilde{P}$. It is required to provide the row- and column-specific numbers of clusters. We rely on an integrated completed likelihood criterion (Brault et al., 2014) to learn relevant values from the data.

The co-clusters obtained *via* WTOT-SCC1, WTOT-SCC2 or WTOT-BC should reveal the interplay between the (remaining, as far as WTOT-SCC2 is concerned) mRNAs and miRNAs in HD.

4.4.2.b Matching.

The larger $\tilde{P}_{mn}$ is, the more we are encouraged to believe that the profiles x_m and y_n reveal a strong relationship between the mth mRNA and the nth miRNA. This simple rule prompts the following matching procedure applied once $\tilde{P}$ has been derived.

WTOT-matching. Fix two integers $k, k' \geq 1$ and let $\tilde{\tau}$ be the quantile of order q of all the entries of $\tilde{P}$. For every $m \in [\![M]\!]$ and $n \in [\![N]\!]$, we introduce

$$\mathcal{N}_m^0 := \Big\{ n \in [\![N]\!] : \tilde{P}_{mn} \in \{\tilde{P}_{m(1)}, \ldots, \tilde{P}_{m(k)}\} \text{ and } \tilde{P}_{mn} \geq \tilde{\tau} \Big\},$$
$$\mathcal{M}_n^0 := \Big\{ m \in [\![M]\!] : \tilde{P}_{mn} \in \{\tilde{P}_{(1)n}, \ldots, \tilde{P}_{(k')n}\} \text{ and } \tilde{P}_{mn} \geq \tilde{\tau} \Big\}$$

where $\tilde{P}_{m(1)}, \ldots, \tilde{P}_{m(k)}$ are the k largest values among $\tilde{P}_{m1}, \ldots, \tilde{P}_{mN}$ and $\tilde{P}_{(1)n}, \ldots, \tilde{P}_{(k')m}$ are the k' largest values among $\tilde{P}_{1n}, \ldots, \tilde{P}_{Mn}$. For instance, $\mathcal{N}_m^0$ identifies the miRNAs that are the k more likely to have a strong relationship with the mth mRNA. However, this does not qualify them as relevant matches yet. In order to keep only matches that are really relevant, we also introduce, for each $m \in [\![M]\!]$ and $n \in [\![N]\!]$,

$$\mathcal{N}_m := \mathcal{N}_m^0 \cap \{n \in [\![N]\!] : m \in \mathcal{M}_n^0\},$$
$$\mathcal{M}_n := \mathcal{M}_n^0 \cap \{m \in [\![M]\!] : n \in \mathcal{N}_m^0\}.$$

Algorithm WTOT-matching outputs the collections $\{\mathcal{N}_m : m \in [\![M]\!]\}$ and $\{\mathcal{M}_n : n \in [\![N]\!]\}$.

Now if, for instance, $n \in \mathcal{N}_m$ then y_n is among the k miRNA profiles upon which $\tilde{P}$ puts more mass when it "transports" x_m onto Y *and* x_m is among the k' mRNA profiles upon which $\tilde{P}$ puts more mass when it "transports" y_n onto X.

Note that we expect that some $\mathcal{N}_m$ and $\mathcal{M}_n$ will be empty, depending on k and k'. The mRNAs and miRNAs worthy of interest are those for which $\mathcal{N}_m$ and $\mathcal{M}_n$ are not empty. The integers k and k' should be chosen relatively small, to make their interpretation and analysis feasible, but not too small because otherwise few matchings will be made.

In the simulation study, we use $k = k'$ between 2 and 200, depending on the simulation scheme. Moreover, we choose $q = 50\%$ so that $\tilde{\tau}$ is the median of the entries of $\tilde{P}$.

4.4.3 Implementation of the method

Our code is written in `python`. We adapt the Sinkhorn algorithm implemented by Aude Genevay and available here. The stochastic gradient descents relies on the machine learning framework `pytorch`. We use the implementation of SCC available in the `sklearn` `python` module. To learn a relevant number of clusters, we rely on the `coclust` `python` module. Finally, we rely on the `blockcluster` `R` package to carry out block clustering.

Our algorithm bears a similarity to the one developed in (Laclau et al., 2017). The main differences are *(i)* our use of the parametric model Θ and weights ω, *(ii)* the fact that we apply SCC or block clustering to the approximation of the optimal transport matrix $\tilde{P}$. Our algorithm also bears a similarity to (Heng et al., 2020), a fast and certifiable point cloud registration algorithm. We plan to study the similarities and differences closely.

4.5 Simulation study

To assess the performances of the algorithms described in Section 4.4.1, we conduct a simulation study in three parts. As we go on, the task gets more difficult. In all cases, the laws of the synthetic observations are mixtures of Gaussian laws. Overall 12 simulation scenarios are considered.

We think that the first two simulation schemes produce unrealistic data and, on the contrary, that the third simulation scheme produces somewhat realistic data. The diversity of the synthetic mRNA and miRNA profiles obtained by using Lloyd's k-means algorithm in order to summarize the variety of real profiles, see Section 4.2.2.a, encouraged us to rely on mixtures in order to simulate data. We chose mixtures of Gaussian laws because of their ubiquity and versatility.

In Section 4.5.4, the weights of the mixtures and parameters of the Gaussian laws are chosen by us. Moreover, the two mixtures (to simulate X and Y) share the same weights and induce a perfect mirroring relationship (details below), thus making the co-clustering task less difficult. In Section 4.5.5, the weights of the mixtures and parameters of the Gaussian laws are randomly generated. Moreover, the two mixtures do not share the same weights and do not induce a perfect mirroring relationship anymore, so that the co-clustering task is much more difficult. Finally, in Section 4.5.6, we use plus or minus real, randomly chosen miRNA profiles *and* $\mathbf{0}_d$ as means of the Gaussian laws to simulate X and Y, in such a way that there is no perfect mirroring relationship. We think that the corresponding co-clustering task is the most difficult of the three.

Section 4.5.1 briefly introduces two competing algorithms to identify matchings (Laclau et al., 2017). Section 4.5.2 lists all the algorithms that compete in the simulation study and Section 4.5.3 presents the measure of discrepancy between two co-clusterings and the matching criteria that we rely on to assess how well the algorithms perform. Sections 4.5.4, 4.5.5 and 4.5.6 present in turn the data-generating mechanisms and report the results in terms of co-clustering and matching performances.

4.5.1 Two ``Gromov-Wasserstein co-clustering" algorithms

We compare our algorithms with two co-clustering algorithms adapted from (Laclau et al., 2017). For self-containedness, we summarize here how these algorithms work.

The first step of both algorithms consists in computing the similarity matrices $K_X \in (\mathbb{R}_+)^{M\times M}$ and $K_Y \in (\mathbb{R}_+)^{N\times N}$ given by

$$(K_X)_{mm'} := \exp\left\{-\frac{\|x_m - x_{m'}\|_2^2}{2\ell_X^2}\right\} \quad (m, m' \in [\![M]\!]),$$
$$(K_Y)_{nn'} := \exp\left\{-\frac{\|y_n - y_{n'}\|_2^2}{2\ell_Y^2}\right\} \quad (n, n' \in [\![N]\!])$$

where ℓ_X (respectively, ℓ_Y) is the mean of all pairwise Euclidean distances between elements of X (respectively, of Y). The similarity matrices K_X and K_Y now represent X and Y through the lense of the so called radial basis function kernel.

For any integers $a, b \geq 1$ and pair of matrices $A \in \mathbb{R}^{a\times a}$ and $B \in \mathbb{R}^{b\times b}$, define

$$\begin{aligned}
\Pi_{a,b} &:= \left\{P \in (\mathbb{R}_+)^{a\times b} | P\mathbf{1}_b = a^{-1}\mathbf{1}_a, P^\top \mathbf{1}_a = b^{-1}\mathbf{1}_b\right\}, \\
\langle [A,B],[P,P]\rangle_F &:= \sum_{i,k\in[\![a]\!],j,\ell\in[\![b]\!]} (A_{ik} - B_{j\ell})^2 P_{ij} P_{k\ell} \quad (P \in \Pi_{a,b}), \\
\mathcal{GW}_\gamma(A,B) &:= \min_{P\in\Pi_{a,b}} \{\langle [A,B],[P,P]\rangle_F - \gamma E(P) \qquad (4.6)
\end{aligned}$$

where $E(P) := -\sum_{(i,j)\in[\![a]\!]\times[\![b]\!]} P_{ij}(\log P_{ij} - 1)$. The quantity $\mathcal{GW}_\gamma(A,B)$ is known in the literature as an entropic Gromov-Wasserstein discrepancy between A and B. It can be used to define an entropic Gromov-Wasserstein barycenter of A and B and its barycenter transport matrices. Specifically, setting $s = \lfloor \frac{1}{2}(a+b)\rfloor$ (one choice among many), $(\hat{\Gamma}, \hat{P}_A, \hat{P}_B) \in (\mathbb{R}_+)^{s\times s} \times \Pi_{s,a} \times \Pi_{s,b}$ that solves

$$\min_{\Gamma, P_A, P_B} \frac{1}{2}\left\{\left(\langle [\Gamma, A],[P_A,P_A]\rangle_F - \gamma E(P_A)\right) + \left(\langle [\Gamma,B],[P_B,P_B]\rangle_F - \gamma E(P_B)\right)\right\} \qquad (4.7)$$

(where (Γ, P_A, P_B) ranges over $(\mathbb{R}_+)^{s\times s} \times \Pi_{s,a} \times \Pi_{s,b}$) can be interpreted as a barycenter between A and B ($\hat{\Gamma}$) and the optimal transport matrices between $\hat{\Gamma}$ and A ($\hat{P}_A$) and between $\hat{\Gamma}$ and B ($\hat{P}_B$).

The second step of the algorithms consists either in solving numerically (4.6) with $(A,B) = (K_X, K_Y)$, yielding $\tilde{Q}$, or in solving numerically (4.7) with $(A,B) = (K_X, K_Y)$, yielding in particular the transport matrices $\tilde{Q}_X$ and $\tilde{Q}_Y$. We call CCOT-GWD and CCOT-GWB the corresponding algorithms. In both cases, the Sinkhorn-Knopp algorithm is used and provides solutions that decompose as

$$\begin{aligned}
\tilde{Q} &= \mathrm{diag}(\rho)\xi\,\mathrm{diag}(\rho'), \\
\tilde{Q}_X &= \mathrm{diag}(\rho_X)\xi_X\,\mathrm{diag}(\rho'_X), \\
\tilde{Q}_Y &= \mathrm{diag}(\rho_Y)\xi_Y\,\mathrm{diag}(\rho'_Y),
\end{aligned}$$

for some $\rho, \rho_X \in \mathbb{R}^M$, $\rho', \rho'_Y \in \mathbb{R}^N$, $\rho_X, \rho_Y \in \mathbb{R}^s$ and $\xi \in \mathbb{R}^{M\times N}, \xi_X \in \mathbb{R}^{s\times M}, \xi_Y \in \mathbb{R}^{s\times N}$ (Peyré et al., 2016).

The third and last step builds upon either (ρ, ρ') or (ρ'_X, ρ'_Y) to derive partitions of X and Y, by detecting "jumps" along the vectors. The two partitions finally yield a co-clustering.

4.5.2 Listing all competing algorithms

We run and compare algorithms WTOT-SCC1, WTOT-SCC2 (and their oracular counterparts WTOT-SCC1*, WTOT-SCC2*), WTOT-BC on the one hand (see Sections 4.4.2.a) and CCOT-GWD and CCOT-GWB on the other hand (see Section 4.5.1). In addition, we also run algorithm WTOT-matching (see Section 4.4.2.b).

For CCOT-GWD, we set $\gamma = 0.1$ in (4.6). For CCOT-GWB, we set $\gamma = 0.05$ in (4.7). We tried several values and chose the ones that yielded the smallest errors.

In view of Procedure 1, we choose $\widetilde{M}$ and $\widetilde{N}$ equal approximately $M/2$ and $N/2$ respectively, $(\eta, \gamma_0) = (1, 0)$ (no decay), $T = 500$, and an initial mapping θ_0 drawn randomly (see Appendix 4.7 for details).

We checked that varying $\widetilde{M}$ and $\widetilde{N}$ around $M/2$ and $N/2$ had little impact if any. Likewise, the randomly drawn initial mapping θ_0 had little impact if any. Moreover, varying $\underline{\gamma}$ in $[\frac{1}{2}\times\gamma^*; 2\times\gamma^*]$ with $\gamma^* = \mathrm{mean}\{\|x - x'\|_2 : x, x' \in X\}$ also had little impact if any. We

did not rigorously check the impact of the total number of iterations T, but we observed that numerical convergence seemed to be reached for fewer iterations than T. Finally, we did not challenge the choice of $h = \text{mean}\{\|y - y'\|_2 : y, y' \in Y\}$.

4.5.3 Assessing performances

A MEASURE OF DISCREPANCY BETWEEN TWO CO-CLUSTERINGS. In order to assess the quality of the co-clusterings that we derive, and to compare performances, we propose to rely on a commonly used measure of discrepancy between two co-clusterings. Its definition extends that of a measure of discrepancy between partitions that we first present.

Let z and z' be two partitions of the set $[\![M]\!]$ into K components, taking the form of matrices $z = (z_{mk})_{m\in[\![M]\!],k\in[\![K]\!]}$ and $z' = (z'_{mk})_{m\in[\![M]\!],k\in[\![K]\!]}$ with convention $z_{mk} = 1$ (respectively, $z'_{mk} = 1$) if m belongs to component k of z (respectively, z') and 0 otherwise. The corresponding confusion matrix $C(z, z') = (c_{k\ell})_{k,\ell\in[\![K]\!]}$ is given by $c_{k\ell} := \sum_{m\in[\![M]\!]} z_{mk}z'_{m\ell}$ (every $k, \ell \in [\![K]\!]$). Suppose that the labels of the partitions z and z' are such that

$$\text{Tr}(C(z, z')) = \max_{\sigma\in\Sigma_K} \text{Tr}(C(z, (z'_{m\sigma(k)})_{m\in[\![M]\!],k\in[\![K]\!]})),$$

where Σ_K is the set of permutations of the elements of $[\![K]\!]$. Then the proportion

$$\delta(z, z') := 1 - \frac{1}{M} \sum_{m\in[\![M]\!],k\in[\![K]\!]} z_{mk}z'_{mk} \tag{4.8}$$

is a natural measure of discrepancy between z and z'. As suggested earlier, the measure can be extended to compare pairs of partitions.

Consider now (z, w) and (z', w') two pairs of partitions, z and z' partitioning $[\![M]\!]$ into K components, w and w' partitioning $[\![N]\!]$ into L components. We represent (z, w) and (z', w') with

$$u = (u_{mnk\ell})_{m\in[\![M]\!],n\in[\![N]\!],k\in[\![K]\!],\ell\in[\![L]\!]}$$

and

$$u' = (u'_{mnk\ell})_{m\in[\![M]\!],n\in[\![N]\!],k\in[\![K]\!],\ell\in[\![L]\!]}$$

where $u_{mnk\ell} := z_{mk} \times w_{n\ell}$ and $u'_{mnk\ell} := z'_{mk} \times w'_{n\ell}$ (for every $m \in [\![M]\!], n \in [\![N]\!], k \in [\![K]\!], \ell \in [\![L]\!]$), supposing again that the labels of the partitions z, z' on the one hand and w, w' on the other hand maximize the traces of the confusion matrices $C(z, z')$ and $C(w, w')$ as above (then two pairs of partitions define without ambiguity a co-clustering). By analogy with (4.8), the proportion

$$\Delta((z, w), (z', w')) := 1 - \frac{1}{KL} \sum_{m\in[\![M]\!],n\in[\![N]\!],k\in[\![K]\!],\ell\in[\![L]\!]} u_{mnk\ell}u'_{mnk\ell} \tag{4.9}$$

is a measure of discrepancy between (z, w) and (z', w'). It can be shown that

$$\Delta((z, w), (z', w')) = \delta(z, z') + \delta(w, w') - \delta(z, z') \times \delta(w, w'). \tag{4.10}$$

In the rest of this section we report means and standard deviations, computed across 30 independent replications of each analysis, of the above measure of discrepancy between the derived partition/co-clustering and the true one.

MATCHING CRITERIA. Set arbitrarily $m \in [\![M]\!]$ and suppose that we have derived the subset $\mathcal{N}_m \subset [\![N]\!]$ that matches x_m to $\{y_n : n \in \mathcal{N}_m\}$. Suppose moreover that in reality x_m is matched to $\{y_n : n \in \mathcal{N}_m^\star\}$ for some $\mathcal{N}_m^\star \subset [\![N]\!]$. We propose to use three real-valued criteria to compare $\mathcal{N}_m$ with $\mathcal{N}_m^\star$.

Let $\mathrm{TP}_m := \mathrm{card}(\mathcal{N}_m \cap \mathcal{N}_m^\star)$, $\mathrm{FP}_m := \mathrm{card}(\mathcal{N}_m \cap (\mathcal{N}_m^\star)^c)$, $\mathrm{TN}_m := \mathrm{card}((\mathcal{N}_m)^c \cap (\mathcal{N}_m^\star)^c)$, $\mathrm{FN}_m := \mathrm{card}((\mathcal{N}_m)^c \cap \mathcal{N}_m^\star)$ be the numbers of true positives, false positives, true negatives and false negatives, respectively. The so called m-specific

- precision: $\mathrm{TP}_m/(\mathrm{TP}_m + \mathrm{FP}_m)$,
- sensitivity: $\mathrm{TP}_m/(\mathrm{TP}_m + \mathrm{FN}_m)$,
- specificity: $\mathrm{TN}_m/(\mathrm{TN}_m + \mathrm{FP}_m)$

quantify how similar are $\mathcal{N}_m$ and $\mathcal{N}_m^\star$, larger values indicating better concordance.

In the rest of this section we report means and standard deviations, computed across 30 independent replications of each analysis, of the average of the m-specific precision, sensitivity and specificity. We also report means and standard deviations, computed across the same 30 independent replications of each analysis, of

$$\tilde{k}_r := \frac{\sum_{m \in [\![M]\!]} \mathrm{card}(\mathcal{N}_m)}{\mathrm{card}(\{m \in [\![M]\!] : \mathcal{N}_m \neq \emptyset\})},$$
$$\tilde{k}_c := \frac{\sum_{n \in [\![N]\!]} \mathrm{card}(\mathcal{M}_n)}{\mathrm{card}(\{n \in [\![N]\!] : \mathcal{M}_n \neq \emptyset\})}$$

the row- and column-specific averages of the cardinalities of the sets $\mathcal{N}_m$ and $\mathcal{M}_n$ that are not empty.

4.5.4 First simulation study

SIMULATION SCHEME. For four different choices of the hyperparameters $M \geq 200, N \geq 200, K \geq 2, d \geq 2$, $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$, $\sigma \in \mathbb{R}_+^*$, $\alpha \in (\mathbb{R}_+)^K$ such that $\sum_{k \in [\![K]\!]} \alpha_k = 1$, we sample independently $x_1, \ldots, x_M$ from the mixture of Gaussian laws

$$\sum_{k \in [\![K]\!]} \alpha_k N(\mu_k, \sigma^2 \mathrm{Id}_d) \tag{4.11}$$

and $y_1, \ldots, y_N$ from

$$\sum_{k \in [\![K]\!]} \alpha_k N(-\mu_k, \sigma^2 \mathrm{Id}_d). \tag{4.12}$$

One way to sample x from the mixture (4.11) consists in sampling a latent label u in $[\![K]\!]$ from the multinomial law with parameter $(1; \alpha_1, \ldots, \alpha_K)$ then in sampling x from the Gaussian law $N(\mu_u, \sigma^2 \mathrm{Id}_d)$. Similarly, sampling y from the mixture (4.12) can be carried out by sampling a latent label v in $[\![K]\!]$ from the multinomial law with parameter $(1; \alpha_1, \ldots, \alpha_K)$ then by sampling y from the Gaussian law $N(-\mu_v, \sigma^2 \mathrm{Id}_d)$. We think of x and y as having a mirrored relationship if $u = v$. In this light, the challenge that we tackle consists in finding such relationships without having access to the latent labels.

Table 4.2 describes the four configurations that we investigate. Note that configuration A2 is more difficult to deal with than A1 because *(i)* the weights in α are balanced in the latter and unbalanced in the former, and *(ii)* because the variance σ^2 is smaller in A1 than in A2. Moreover, configurations A3 and A4 are more challenging than A2 because there is $K = 4$ components in the Gaussian mixture under A3 and A4 and $K = 3$ components under A2.

configuration	(M, N)	K	$\mu_1, \ldots, \mu_K$	σ^2	α
A1	$(200, 200)$	3	$\begin{pmatrix}4.0\\0.5\\1.5\end{pmatrix}, \begin{pmatrix}1.8\\4.5\\1.1\end{pmatrix}, \begin{pmatrix}1.5\\1.5\\5.5\end{pmatrix}$	0.10	$(1/3, 1/3, 1/3)$
A2	$(300, 300)$	3	$\begin{pmatrix}4.0\\0.5\\1.5\end{pmatrix}, \begin{pmatrix}1.8\\4.5\\5.1\end{pmatrix}, \begin{pmatrix}3.5\\1.5\\5.5\end{pmatrix}$	0.15	$(0.2, 0.3, 0.5)$
A3	$(400, 300)$	4	$\begin{pmatrix}4.0\\0.5\end{pmatrix}, \begin{pmatrix}0.5\\3.5\end{pmatrix}, \begin{pmatrix}7.5\\7.8\end{pmatrix}, \begin{pmatrix}0.5\\0.5\end{pmatrix}$	0.20	$(0.4, 0.2, 0.2, 0.2)$
A4	$(300, 300)$	4	$\begin{pmatrix}4.0\\0.5\end{pmatrix}, \begin{pmatrix}0.5\\3.5\end{pmatrix}, \begin{pmatrix}7.5\\7.8\end{pmatrix}, \begin{pmatrix}0.5\\0.5\end{pmatrix}$	0.10	$(0.5, 0.2, 0.1, 0.2)$

Table 4.2 – Four different configurations for the first simulation scheme. Configuration A1 is less challenging than A2 which is itself less challenging than A3 and A4.

RESULTS. Thirty times, independently, we simulated synthetic data sets X and Y under the simulation scheme described above, then we applied the various algorithms as presented in Section 4.5.2. We summarize the results in Tables 4.5, 4.6, and 4.7. Table 4.5 summarizes the results of the seven algorithms listed in Section 4.5.2 that rely on *bona fide* co-clustering algorithms (see Section 4.4.2.a), that is, of our algorithms WTOT-SCC1*, WTOT-SCC1, WTOT-SCC2*, WTOT-SCC2, WTOT-BC* and of algorithms CCOT-GWD and CCOT-GWB. As for Tables 4.6 and 4.7, they summarize the results of our algorithm that relies on matching (see Section 4.4.2.b).

Table 4.5. Except in configuration A1, where they perform equally well, our algorithms WTOT-SCC1, WTOT-SCC2 outperform their competitors CCOT-GWD and CCOT-GWB.

Recall that WTOT-SCC1 and WTOT-SCC2 learn the number of co-clusters. When they underestimate it, they pay a high price, partly explaining why the standard deviations are rather large. In order to assess how well they work relative to their counterparts which benefit from knowing in advance the true number of co-clusters, we can compare their measures of performance to those of algorithms WTOT-SCC1* and WTOT-SCC2*. In configurations A1 and A2, algorithms WTOT-SCC1, WTOT-SCC2 perform almost as well as WTOT-SCC1* and WTOT-SCC2*, respectively. In configuration A3, they are clearly outperformed. In configuration A4, algorithm WTOT-SCC1 performs better in average but not in standard deviation.

Finally, we note that algorithm WTOT-BC* outperforms all our other algorithms. Unfortunately, its counterpart that learns the number of co-clusters performs poorly (results not shown).

Tables 4.6 and 4.7. Table 4.6 illustrates the influence of $k = k'$ on the performances of algorithm WTOT-matching. In configuration A1, specificity is not impacted much by the value of $k = k'$, whereas precision decreases and sensitivity increases as $k = k'$ grows. More specifically, precision does not change much when one goes from $k = k' = 10$ to $k = k' = 75$ but it drops for larger values of $k = k'$. As for sensitivity, it increases dramatically when one goes from $k = k' = 10$ to $k = k' = 75$ and slightly for higher values of $k = k'$. Furthermore we note that, in configuration A1, when $k = k'$ equal either 65 or 75 and are thus closest to $N\alpha_\ell = M\alpha_\ell \approx 67$, $\tilde{k}_r$ is close to 67 and precision, sensitivity and specificity are quite satisfying. In configuration A4 (as

in configuration A1), specificity is not impacted much by the value of $k = k'$; on the contrary, precision decreases and sensitivity increases steadily as $k = k'$ grows. The best performances are achieved for $k = k' = 95$ and $k = k' = 150$, that is, when $k = k'$ get closer to $M \max_{i \leq 4}\{\alpha_i\} = N \max_{i \leq 4}\{\alpha_i\}$. As emphasized earlier, deriving relevant matchings is more difficult in configuration A4 than in configuration A1 because the weights given in parameter α are unbalanced in the former and balanced in the latter.

Table 4.7 summarizes the results of WTOT-matching in all configurations for a specific choice of $k = k'$ in terms of the row- and column-specific averages $\tilde{k}_r$ and $\tilde{k}_c$, precision, sensitivity and specificity. In each configuration, we chose the value of $k = k'$ among many retrospectively, so that the overall performance (in terms of precision, sensitivity and specificity) is good. The left-hand-side (m-specific) and right-hand-side (n-specific) tables in Table 4.7 are very similar. This does not come as a surprise because the first simulation scheme imposes symmetry.

4.5.5 Second simulation study

SIMULATION SCHEME. The second simulation scheme also relies on mixtures of Gaussian laws, but the means and weights are generated randomly from a Gaussian determinantal point process (DPP) for the former and from a Dirichlet law for the latter. More specifically, given the hyperparameters $M \geq 200, N \geq 200, K \geq L \geq 3$, $\sigma \in \mathbb{R}_+^*$,

1. we sample $\mu_1, \ldots, \mu_K$ from a Gaussian DPP on $[0,1]^2$ with a kernel proportional to $x \mapsto \exp(-\|x/0.05\|_2^2)$ conditionally on obtaining exactly K points (Lavancier et al., 2015; Baddeley and Turner, 2005);

2. independently, we sample $\alpha \in (\mathbb{R}_+)^K$ and $\beta \in (\mathbb{R}_+)^L$ from the Dirichlet laws with parameters $7\,\mathbf{1}_K$ and $7\,\mathbf{1}_L$;

3. we sample independently $x_1, \ldots, x_M$ from the mixture of Gaussian laws

$$\sum_{k \in [\![K]\!]} \alpha_k N(\mu_k, \sigma^2 \mathrm{Id}_2)$$

and $y_1, \ldots, y_N$ from

$$\sum_{k \in [\![L]\!]} \beta_k N(-\mu_k, \sigma^2 \mathrm{Id}_2).$$

We use a DPP to generate $\mu_1, \ldots, \mu_K$ to avoid the arbitrary choice of the mean parameters in such a way that the randomly picked $\mu_1, \ldots, \mu_K$ are dispersed in $[0,1]^2$ (because the DPP is a repulsive point process).

Table 4.3 describes the four configurations that we investigate. The larger L is the more challenging the configuration is. In configurations B2, B3, B4, it holds that $K = L+1$, hence the data points from the Kth cluster should not be matched. Moreover, for given (K, L) and (M, N), a configuration gets more challenging as its σ^2 parameter increases. It is noteworthy that the values of σ^2 as reported in Table 4.3 cannot be compared straightforwardly to those reported in Table 4.2, because $\mu_1, \ldots, \mu_K$ live in $[0,1]^2$ in the present simulation study whereas they do not in the simulation study of Section 4.5.4.

configuration	(M, N)	(K, L)	σ^2
B1	$(200, 200)$	$(3, 3)$	5×10^{-4}
B2	$(300, 300)$	$(7, 6)$	10^{-4}
B3	$(300, 300)$	$(16, 15)$	10^{-5}
B4	$(300, 300)$	$(16, 15)$	10^{-4}

Table 4.3 – Four different configurations for the second simulation scheme. The larger $\ell \in [4]$ is the more challenging configuration Bℓ is.

RESULTS. Thirty times, independently, we simulated synthetic data sets X and Y under the simulation scheme described above, then we applied the various algorithms as presented in Section 4.5.2. Table 4.8 summarizes the results of the seven algorithms listed in Section 4.5.2 that rely on *bona fide* co-clustering algorithms (see Section 4.4.2.a). Tables 4.9 and 4.10 summarize the results of our algorithm that relies on matching (see Section 4.4.2.b).

Table 4.8. We first note that WTOT-SCC1, WTOT-SCC2 and CCOT-GWD perform similarly in configurations B1 and B2, much better than CCOT-GWB, but less well than the oracular algorithms WTOT-SCC1*, WTOT-SCC2* and WTOT-BC*. More generally, across configurations B1, B2, B3, B4, the oracular algorithms WTOT-SCC1* and WTOT-SCC2* perform much better than the other algorithms (and WTOT-BC* fails to find a partition with the given number of co-clusters in B3 and B4). Moreover, WTOT-SCC1 and WTOT-SCC2 perform poorly in configurations B2, B3 and B4 though not as poorly as CCOT-GWD and CCOT-GWB in configurations B3 and B4. It seems that WTOT-SCC1 and WTOT-SCC2 fail to learn a "practical" number of co-clusters from $\tilde{P}$, in part because of those among $x_1, \ldots, x_M$ that are drawn from the Gaussian law $N(\mu_K, \sigma^2 \mathrm{Id}_2)$ when $K = L + 1$ (these data points should not be matched at all). The fact that WTOT-SCC1 and WTOT-SCC2 perform similarly in configurations B3 and B4 although σ^2 is 10 times larger in B4 than in B3 gives credit to the previous interpretation.

Tables 4.9 and 4.10. Table 4.9 illustrates the influence of $k = k'$ on the performances of algorithm WTOT-matching in configurations B1 and B4. In each configuration, the values of $k = k'$ are chosen in the vicinity of M/K (67 in configuration B1, 11 in configuration B4). We observe the same patterns in configurations B1 and B4: precision decreases (gradually) and specificity decreases (slightly) as $k = k'$ grows, while sensitivity increases (strongly in B1 and dramatically in B4).

Table 4.10 summarizes the results of WTOT-matching in configurations B1, B2, B3, B4 for a specific choice of $k = k'$ in terms of the row- and column-specific averages $\tilde{k}_r$ and $\tilde{k}_c$, precision, sensitivity and specificity. In each configuration, we chose the value of $k = k'$ among many retrospectively so that the overall performance (in terms of precision, sensitivity and specificity) is good. The left-hand-side (m-specific) and right-hand-side (n-specific) tables in Table 4.10 are very similar although $K > L$ in configuration B3 and B4. Interestingly, the fact that σ^2 is 10 times larger in configuration B4 than in B3 does not affect much the performance of the matching algorithm.

4.5.6 Third simulation study

SIMULATION SCHEME. The third simulation scheme aspires to generate synthetic data sets X and Y that are more similar to the real data sets than those generated in the two first simulation studies. Once again, we rely on mixtures of Gaussian laws. This time, however,

the various means are neither chosen arbitrarily (unlike in the first simulation study) nor drawn randomly (unlike in the second simulation study) but are sampled in the real collection of miRNAs. Moreover, the weights of the mixtures are random.

Specifically, given the hyperparameters $K \geq 3$, $\lambda_x, \lambda'_x \geq 0$, $\lambda_y, \lambda'_y \geq 0$ and $\sigma, \sigma' \in \mathbb{R}_+^*$ (with σ' much larger than σ),

1. we sample $\mu_1, \ldots, \mu_K$ uniformly without replacement from the collection of observed miRNA profiles conditionally on $\min_{k \neq k'} \|\mu_k - \mu_{k'}\|_2 \geq 2$;

2. independently, we sample independently $(m_1 - 1), \ldots, (m_K - 1)$ from the Poisson law with parameter λ_x, $(n_1 - 1), \ldots, (n_K - 1)$ from the Poisson law with parameter λ_y, $(m_{K+1} - 1)$ and $(n_{K+1} - 1)$ from the Poisson laws with parameter λ'_x and λ'_y;

3. for each $1 \leq k \leq K$, we sample independently $x_{k,1}, \ldots, x_{k,m_k}$ from the Gaussian law $N(\mu_k, \sigma^2 \mathrm{Id}_{18})$ and $y_{k,1}, \ldots, y_{k,n_k}$ from the Gaussian law $N(-\mu_k, \sigma^2 \mathrm{Id}_{18})$. Moreover, we also sample independently $x_{K+1,1}, \ldots, x_{K+1,m_{K+1}}$ and $y_{K+1,1}, \ldots, y_{K+1,n_{K+1}}$ from the Gaussian law $N(\mathbf{0}_{18}, (\sigma')^2 \mathrm{Id}_{18})$.

Here, we think of x and y as having a mirrored relationship if there exists $k \in [\![K]\!]$ such that x and y are drawn from the laws $N(\mu_k, \sigma^2 \mathrm{Id}_{18})$ and $N(-\mu_k, \sigma^2 \mathrm{Id}_{18})$. Furthermore, we view x and y drawn from the law $N(\mathbf{0}_{18}, (\sigma')^2 \mathrm{Id}_{18})$ as noise.

Table 4.4 describes the four configurations that we investigate. The larger K is the more challenging the configuration is.

configuration	(λ_x, λ_y)	(λ'_x, λ'_y)	K	(σ, σ')
C1	$(50, 50)$	$(50, 10)$	3	$(0.1, 5)$
C2	$(15, 15)$	$(0, 0)$	15	$(0.01, 5)$
C3	$(15, 15)$	$(30, 30)$	15	$(0.01, 5)$
C4	$(15, 15)$	$(30, 30)$	15	$(0.1, 5)$

Table 4.4 – Four different configurations for the third simulation scheme. The larger $\ell \in [4]$ is the more challenging configuration Cℓ is.

RESULTS. Thirty times, independently, we simulated synthetic data sets X and Y under the simulation scheme described above, then we applied the various algorithms as presented in Section 4.5.2. Table 4.11 summarizes the results of the seven algorithms listed in Section 4.5.2 that rely on *bona fide* co-clustering algorithms (see Section 4.4.2.a). Tables 4.12 and 4.13 summarize the results of our algorithm that relies on matching (see Section 4.4.2.b).

Table 4.11. We first focus on configuration C1. We note that WTOT-SCC1 and WTOT-SCC2 perform similarly, much better than CCOT-GWD and CCOT-GWB, better than the oracular algorithm WTOT-BC*, but not as well as the oracular algorithms WTOT-SCC1* and WTOT-SCC2*.

We now turn to configurations C2, C3 and C4. Configuration C3 is more challenging than configuration C2 because it shares the same hyperparameters as C2 except for (λ'_x, λ'_y) (which drives the number of noisy data points), set to $(0, 0)$ in C2 and to $(30, 30)$ in C3. Similarly, configuration C4 is more challenging than configuration C3 because it shares the same hyperparameters as C3 except for σ (the standard deviation of the Gaussian variations around the mean profiles), set to 0.01 in C3 and to 0.1 in

C4. The comparisons will not concern algorithms WTOT-BC* (which never converges in these simulations), CCOT-GWD and CCOT-GWB (which perform very poorly).

In configuration C2, in the absence of noisy data points, algorithm WTOT-SCC1 performs slightly better than WTOT-SCC2, as well as the oracular algorithm WTOT-SCC2*, and almost as well as the oracular algorithm WTOT-SCC1* (in average). In configurations C3 and C4, the introduction of noisy data points then the increase in variability strongly degrade the performances of WTOT-SCC1, WTOT-SCC1* and, to a lesser extent, those of WTOT-SCC2 and WTOT-SCC2*. Algorithm WTOT-SCC2 outperforms WTOT-SCC1 and the oracular algorithm WTOT-SCC1* too.

Tables 4.12 and 4.13. Table 4.12 illustrates the influence of $k = k'$ on the performances of algorithm WTOT-matching in configurations C1 and C4. In each configuration, the values $k = k'$ are chosen in the vicinity of λ_x or λ_y (50 in configuration C1, 15 in configuration C4). For specificity and sensitivity, we observe the same patterns in configurations C1 and C4: specificity is not impacted much as $k = k'$ grows whereas sensitivity increases dramatically. Precision remains high in configuration C1 for all choices of $k = k'$. In configuration C4, precision remains high for $k = k'$ ranging between 5 and 20, then it decreases when $k = k'$ grows from 25 to 30.

Table 4.13 summarizes the results of WTOT-matching in configurations C1, C2, C3, C4 for a specific choice of $k = k'$ in terms of the row- and column-specific averages $\tilde{k}_r$ and $\tilde{k}_c$, precision, sensitivity and specificity. In each configuration, we chose the value of $k = k'$ among many retrospectively, so that the overall performance (in terms of precision, sensitivity and specificity) is good. The left-hand-side (m-specific) and right-hand-side (n-specific) tables in Table 4.13 are very similar. In configurations C1 and C2, all precision, sensitivity and specificity are quite satisfying. In configurations C3, C4, sensitivity and specificity are quite satisfying as well while precision falls bellow 0.86.

4.6 Illustration on real data: matching mRNA and miRNA in Huntington's disease mice

Next, we apply algorithms WTOT-SCC2 and WTOT-matching to discover patterns hidden in RNA-seq data obtained in the striatum of HD model mice. As explained in Section 4.1, multidimensional mRNA and miRNA sequencing data were obtained in the striatum of these mice (Langfelder et al., 2016, 2018) and an earlier analysis of these data using shape analysis concepts (Mégret et al., 2020) has demonstrated their value.

4.6.1 Tuning

Specifically, in view of Procedure 1, we choose $\widetilde{M} = 1,024$, $\widetilde{N} = 512$, $T = 500$. The entries of the 3×5 matrices $\tilde{\theta}_1^a, \tilde{\theta}_1^b, \tilde{\theta}_1^c$ are constrained to take their values in $]-10, 0[$ (for WTOT-SCC2) or $]-2, 0[$ (for WTOT-matching), $]-0.2, 0.2[$ and $]-0.2, 0.2[$ respectively. We also choose $(\eta, \gamma_0) = (0.95, 3)$. Finally, the initial mapping θ_0 is drawn randomly.

Furthermore, regarding step 2 of algorithm WTOT-SCC2, we remove rows and columns based on the following loop: 100 times successively, *(i)* we compute the Kullback-Leibler divergence between each row (renormalized) and the uniform distribution then remove the 100 rows with the smallest divergences, then *(ii)* we compute the Kullback-Leibler divergence

between each column (renormalized) and the uniform distribution then remove the 5 columns with the smallest divergences. By doing so, we successively get rid of the rows and columns which, viewed as distributions, are too uniform and therefore deemed irrelevant. Finally, we remove all rows for which the (columnwise) sum of the remaining entries of $\tilde{P}$ is smaller than one tenth of the maximal (columnwise) sum, and all columns for which the (rowwise) sum of the remaining entries of $\tilde{P}$ is smaller than one tenth of the maximal (rowwise) sum.

4.6.2 Results

CO-CLUSTERING. The selection procedure (step 2 of WTOT-SCC2) keeps 3,409 mRNA profiles (among the 13,616 available in the data set) and 602 miRNA (among the 1,143 available in the data set). Eventually, algorithm WTOT-SCC2 outputs 8 co-clusters. The co-clusters's sizes (numbers of mRNA and miRNA gathered in each co-cluster) are $(321, 86)$, $(333, 30)$, $(261, 6)$, $(498, 125)$, $(127, 5)$, $(708, 203)$, $(703, 119)$, $(458, 28)$. Figure 4.6 represents the averages, computed across all blocks, of the entries of the matrix derived from the optimal transport matrix $\tilde{P}$ during step 2 of algorithm WTOT-SCC2 and after its rearrangement. Squares located on the diagonal tend to be slightly darker than the other squares. This reveals that, in average, a pair (x_m, y_n) of mRNA and miRNA gathered in a diagonal co-cluster tends to exhibit a mirrored relationship that is slightly stronger than those of the form $(x_m, y_{n'})$ or $(x_{m'}, y_n)$ which do not fall in the same co-cluster. However, few of the off-diagonal averages are small in comparison to the on-diagonal averages, a disappointing observation that comes on top of the fact that the co-clusters' sizes are so large that it is difficult to interpret the results. This makes it even more relevant to focus on algorithm WTOT-matching.

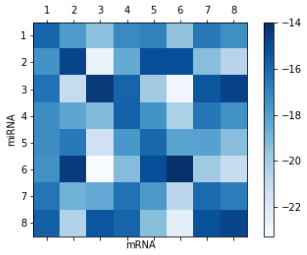

Figure 4.6 – Logarithms of the averages, computed across all blocks, of the entries of the matrix derived from the optimal transport matrix $\tilde{P}$ during step 2 of algorithm WTOT-SCC2 and after its rearrangement.

MATCHING. We run the WTOT-matching algorithm with $k = k' = 10$ and $q = 90\%$. For the anecdote, we observe $(\tilde{k}_r, \tilde{k}_c) \approx (1.82, 6.04)$ (recall that $\tilde{k}_r, \tilde{k}_c$ are the row- and column-specific averages of the cardinalities of the sets $\mathcal{N}_m$ and $\mathcal{M}_n$ that are not empty). We report the parameters that characterize the mapping $\widehat{\theta}$ in Appendix 4.7.

As an illustration, the mirrored profile (the opposite value of y_n) of the Mir20b miRNA is displayed in Figure 4.7 along with its three matched mRNAs (Ahrr, Cnih3 and Relb) obtained by running algorithm WTOT-matching algorithm with $k = k' = 10$. Recall that the original profile of Mir20b can be found in Figure 4.1.

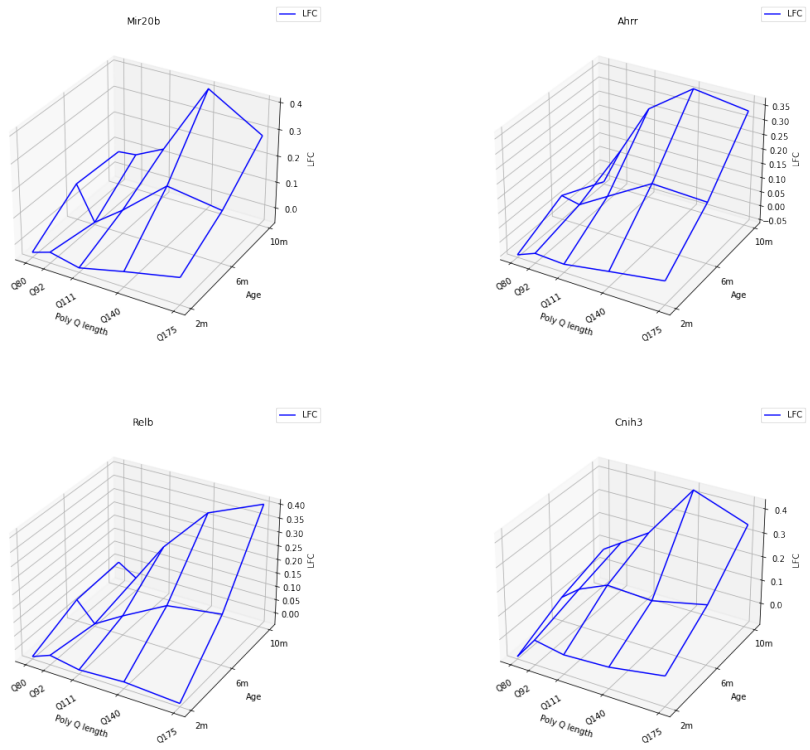

Figure 4.7 – Minus the profile $-y_n$ of the Mir20b miRNA (top left), and profiles x_m of its matched mRNAs, Ahrr (top right), Relb (bottom left) and Cnih3 (bottom right).

4.6.3 Biological analysis of the results

In an effort to guarantee biological relevance to the matchings, we only retain those showing evidence for binding sites as indicated in the databases TargetScan (Lewis et al., 2005), MicroCosm (Betel et al., 2010) and miRDB (Ding et al., 2016). Specifically, a pair (x, y) is retained if and only if the mRNA whose profile is x and the miRNA whose profile is y are both among the 27,355 mRNAs and 1,478 miRNAs appearing in the TargetScan, MicroCosm and miRDB databases. The 1,247 matchings retained out of the 7,521 output by the WTOT-matching algorithm are all presented on this page of the companion website. We stress that we would have obtained fewer matchings if we had excluded from the collections X and Y the profiles of mRNA or miRNA which do not appear in the databases.

Furthermore, we build upon two previous analyses of miRNA regulation in the striatum of HD knock-in-mice (Langfelder et al., 2018; Mégret et al., 2020) to comment on the biological relevance and novelty of our findings. The first analysis (Langfelder et al., 2018) relies on the WGCNA algorithm, a weighted gene co-expression network analysis which yields clusters of genes whose expression profiles are correlated. The second analysis (Mégret et al., 2020) relies on the MiRAMINT algorithm. MiRAMINT is a pipeline whose main steps consist in *(a)* carrying out a weighted gene co-expression network analysis, *(b)* using random forests to select candidate matchings, and *(c)* using Spearman's correlation test and a multiple testing procedure to identify the more reliable matchings. We highlight that

WGCNA outputs 1,583 mRNA-miRNA matchings showing evidence for binding sites in the databases TargetScan, MicroCosm and miRDB, which involve only 46 different miRNAs. As for MiRAMINT, it only outputs 31 matchings of which 20 show evidence for binding sites in the databases TargetScan, MicroCosm and miRDB, involving 14 different miRNAs. The 31 mRNA-miRNA matchings output by MiRAMINT are all presented on this webpage.

ANALYZING THE OVERLAPS. Three mRNA-miRNA matchings are retained both by the WTOT-matching and WGCNA algorithms: Mir186-Chl1, Mir132-Fam196b, Mir212-Fam196b. No matchings are retained both by the WTOT-matching and the MiRAMINT algorithms. One pair is retained both by the MiRAMINT and WGCNA algorithms: Mir132-Pafah121.

Figure 4.8 in Appendix 4.7 presents two Venn diagrams summarizing the overlaps between the sets of miRNAs (respectively, mRNAs) which belong to a pair output by the WGCNA, MiRAMINT and WTOT-matching algorithms. On the one hand, focusing on miRNAs, 13/14 (respectively, 29/46) miRNAs involved in a mRNA-miRNA pair output by MiRAMINT (respectively, WGCNA) are among the miRNAs involved in a mRNA-miRNA pair output by WTOT-matching. On the other hand, focusing on mRNAs, 1/20 (respectively, 100/1,583) miRNAs involved in a mRNA-miRNA pair output by MiRAMINT (respectively, WGCNA) are among the miRNAs involved in a mRNA-miRNA pair output by WTOT-matching. We carry out one-sided Fisher's exact tests to quantify to what extent the overlaps reflect an agreement between two algorithms (using the 1,478 miRNAs and 27,355 mRNAs appearing in the TargetScan, MicroCosm and miRDB databases as reference populations). The p-value of the test comparing WTOT-matching and MiRAMINT equals 0.45. The other p-values are smaller than 10^{-6}.

It is desirable to identify miRNAs that are particularly susceptible to play a distinct role in HD in mice. To do so, we evaluate two simple criteria on the mRNAs associated to each miRNA (the miRNAs with no matched mRNAs are obviously less interesting in our study). The criteria assess to what extent a mRNA profile is "monotonic" and, on the contrary, to what extent it is "peaked", accounting for the amplitude of log-fold change. Formally, rewriting each profile $x \in \mathbb{R}^{15}$ as a matrix $(\tilde{x}_{tq})_{t\in[\![3]\!],q\in[\![5]\!]}$, the first criterion is the minimum (relative to time t) of the absolute values of the slopes of the regression lines of the sets $\{(q, \tilde{x}_{tq}) : q \in [\![5]\!]\}$ and the second criterion is $\max_{q\in[\![5]\!]}(\tilde{x}_{1q} - \tilde{x}_{2q}) \times (\tilde{x}_{2q} - \tilde{x}_{3q})$. By convention, a miRNA profile is labeled monotonic (respectively, peaked) if at least one of its associated mRNA profiles is such that its first (respectively, second) criterion is larger than 95% (respectively, smaller than 99%) of the similar criteria. Moreover, all mRNA profiles x appearing in a pair (x, y) are labeled like y. We stress that no mRNA labeling conflicts occur.

Below, we reproduce the same analysis as above focusing in turn on mRNA-miRNA matchings labeled as peaked, monotonic, and neither peaked nor monotonic.

Peaked profiles. Figure 4.9 in Appendix 4.7 presents two Venn diagrams summarizing the overlaps between the sets of miRNAs (respectively, mRNAs) which belong to a pair output by the WGCNA, MiRAMINT and WTOT-matching algorithms, *looking at the WTOT-matching matchings labeled as peaked*. None of the 17 miRNAs and none of the 12 mRNAs involved in a mRNA-miRNA pair output by WTOT-matching are involved in a mRNA-miRNA pair output by the WGCNA or MiRAMINT algorithms.

The take-home message is that the WTOT-algorithm retains mRNA-miRNA matchings that we label as peaked whereas neither the WGCNA nor the MiRAMINT algorithms do.

Monotonic profiles. Figure 4.10 in Appendix 4.7 presents two Venn diagrams summarizing the overlaps between the sets of miRNAs (respectively, mRNAs) which belong to a pair output by the WGCNA, MiRAMINT and WTOT-matching algorithms, *looking at the WTOT-matching matchings labeled as monotonic*. On the one hand, focusing on miRNAs, 8/14 (respectively, 9/46) miRNAs involved in a mRNA-miRNA pair output by MiRAMINT (respectively, WGCNA) are among the miRNAs involved in a mRNA-miRNA pair output by WTOT-matching. On the other hand, focusing on mRNAs, 0/20 (respectively, $14/1,583$) miRNAs involved in a mRNA-miRNA pair output by MiRAMINT (respectively, WGCNA) are among the miRNAs involved in a mRNA-miRNA pair output by WTOT-matching. We carry out one-sided Fisher's exact tests to quantify to what extent the overlaps reflect an agreement between two algorithms (using the 1,478 miRNAs and 27,355 mRNAs appearing in the TargetScan, MicroCosm and miRDB databases as reference populations), excluding the comparison of the MiRAMINT and WTOT-matching algorithms in mRNAs (due to an empty intersection). The p-values are smaller than 10^{-5}.

The take-home message is that, in matchings that we label as monotonic, the agreement between the WTOT-matching and WGCNA algorithms is better than that between the WTOT-matching and MiRAMINT algorithms.

Neither peaked nor monotonic profiles. Finally, Figure 4.11 in Appendix 4.7 presents two Venn diagrams summarizing the overlaps between the sets of miRNAs (respectively, mRNAs) which belong to a pair output by the WGCNA, MiRAMINT and WTOT-matching algorithms and labeled neither as peaked nor monotonic. On the one hand, focusing on miRNAs, 12/14 (respectively, 28/46) miRNAs involved in a mRNA-miRNA pair output by MiRAMINT (respectively, WGCNA) are among the miRNAs involved in a mRNA-miRNA pair output by WTOT-matching. On the other hand, focusing on mRNAs, 1/20 (respectively, $86/1,583$) miRNAs involved in a mRNA-miRNA pair output by MiRAMINT (respectively, WGCNA) are among the miRNAs involved in a mRNA-miRNA pair output by WTOT-matching. We carry out one-sided Fisher's exact tests to quantify to what extent the overlaps reflect an agreement between two algorithms (using the 1,478 miRNAs and 27,355 mRNAs appearing in the TargetScan, MicroCosm and miRDB databases as reference populations), excluding the comparison of the MiRAMINT and WTOT-matching algorithms in mRNAs (due to an intersection reduced to a singleton). The p-value are smaller than 10^{-5}.

The take-home message is that, in matchings that we label as neither peaked nor monotonic, the agreement between the WTOT-matching and WGCNA algorithms is better than that between the WTOT-matching and MiRAMINT algorithms.

ENRICHMENT ANALYSIS. Next, we assess and compare the biological significance of the mRNAs retained by the WGCNA, MiRAMINT and WTOT-matching algorithms. To do so we carry out an enrichment analysis using the EnrichR tools (Chen et al., 2013; Kuleshov et al., 2016; Xie et al., 2021). We consider only top annotations (balancing a small p-value and a large number of hits) as provided by Gene Ontology data (biological process, cellular content) and KEGG data. When necessary, only the top 40 hits are considered so as to guarantee a sufficient level of biological precision. Pubmed searches are also used to assess the biological significance of predicted miRNA regulation.

Figures 4.12, 4.13 and 4.14 in Appendix 4.7 present the mRNA-miRNA networks based on the mRNA-miRNA matchings output by the WTOT-matching algorithm, focusing on the matchings which are labeled as peaked, monotonic and neither peaked nor monotonic

(in that order). The mRNAs and miRNAs retained by the WGCNA and MiRAMINT algorithms are colored. The enrichment analysis reveals

- that the mRNA-miRNA matchings output by the WGCNA algorithm are primarily annotated for *axonogenesis*[*], which relates to cytoskeleton dynamics and cell morphology;
- that the matchings output by the MiRAMINT algorithm are primarily annotated for *regulation of defense response to virus by host*[†], which relates to stress response and innate immunity;
- that the matchings output by the WTOT-matching algorithm are primarily annotated for *extracellular matrix organization* (which relates to cell identity)[‡], due to the matchings labeled as neither peaked nor monotonic, and secondarily annotated for *mitigation of host antiviral defense response*[§], due to the matchings labeled as monotonic, and for *conventional motile cilium*[¶], due to the matchings labeled as peaked.

Although the numbers of hits in some of these annotations are small, they suggest that the WTOT-matching algorithm is able to uncover a role of miRNA regulation in responding to mutant huntingtin that was not detected by the WGCNA and MiRAMINT algorithms (despite the large number of mRNAs retained by the former).

We now interpret the above results from a biological viewpoint. Recall that the peaked and monotonic profiles are especially interesting because they are more susceptible to correspond to mRNAs and miRNAs that play a distinct role in HD in mice. Extracellular matrix organization (the primary annotation of the matchings output by the WTOT-matching algorithm, driven by the mRNA-miRNA matchings labeled as neither peaked nor monotonic) is known to be regulated by miRNAs (Rutnam et al., 2013) and HD mutations are known to strongly affect neuronal identity via down-regulating a large number of cell identity genes (Achour et al., 2015). Mitigation of host antiviral defense response (the first secondary annotation of the matchings output by the WTOT-matching algorithm, due to the mRNA-miRNA matchings labeled monotonic) is similar to the primary annotation of the matchings output by the MiRAMINT algorithm. Finally, conventional motile cilium (the second secondary annotation of the matchings output by the WTOT-matching algorithm, due to the mRNA-miRNA matchings labeled peaked) is a new finding.

Additionally, although miRNA levels and regulation in response to mutant huntingtin is anticipated to be dependent on cellular context and could be differentially influenced across murine models of HD, it is noticeable that the analysis of miRNA regulation in the striatum of HD knock-in mice based on the WTOT-matching algorithm retained several miRNAs that are altered in the striatum of other types of HD mice such as BACHD (Olmo et al., 2021) or altered in the human HD caudate nucleus (Petry et al., 2022) such as for example Mir100, Mir127, Mir132, Mir 212 and Mir133, supporting the relevance of our findings for the study of molecular regulation in mouse and human HD.

[*]GO:0007409, de novo generation of a long process of a neuron, including the terminal branched region. Refers to the morphogenesis or creation of shape or form of the developing axon, which carries efferent (outgoing) action potential from the cell body towards target cells.

[†]GO:0050691, any host process that modulates the frequency, rate or extent of the antiviral response of a host cell or organism.

[‡]GO:0030198, a process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of an extracellular matrix.

[§]GO:0050690, evasion by virus of host immune response.

[¶]GO:0097729, a motile cilium where the axoneme has a ring of 9 outer microtubules doublets plus 2 central micro tubules.

We believe that these facts substantiate our claim that the WTOT-matching algorithm strikes a good balance between the low and high selectivity of the WGCNA and MiRAMINT algorithms. Moreover, our findings related to striatal alterations in HD mice lead to reconsidering the formerly-expressed view on a limited role of miRNA regulation in the striatum of HD mice on a systems level (Mégret et al., 2020).

4.7 Appendix

PARAMETRIC MODEL Θ. Introduced in Section 4.4.1, the parametric model Θ consists of affine mappings $\theta : \mathbb{R}^d \to \mathbb{R}^d$ of the form $x \mapsto \theta_1 x + \theta_2$, where θ_1 takes its values in a subset T_1 of $\mathbb{R}^{d\times d}$ and θ_2 takes its values in $\mathbb{R}^d$ (without any constraint). It is easier to describe the set of linear mappings $\{x \mapsto \theta_1 x : \theta_1 \in T_1\}$ after a reparametrization.

In the rest of this section only, we rewrite the mRNA and miRNA profiles $x, y \in \mathbb{R}^d$ under the form of $d_1 \times d_2$ matrices $\tilde{x} = (\tilde{x}_{tq})_{t\in[\![d_1]\!],q\in[\![d_2]\!]}$ and $\tilde{y} = (\tilde{y}_{tq})_{t\in[\![d_1]\!],q\in[\![d_2]\!]}$. For each $t \in [\![d_1]\!]$ and $q \in [\![d_2]\!]$, $\tilde{x}_{t\bullet}$ and $\tilde{x}_{\bullet q}$ are the tth row and qth column of $\tilde{x}$. Here, indices t and q correspond to the age and CAG lengths of the mice whose RNA sequencing yielded $\tilde{x}_{tq}$ and $\tilde{y}_{tq}$.

The definition of T_1 should formalize what we consider to be a (plausible) mirroring relationship. The simplest mirroring relationship is $y = -x$ or, equivalently, $\tilde{y} = -\tilde{x}$. The equality is of course too stringent/rigid, and the definition of T_1 is driven by our wish to relax it.

Biological arguments encourage us to consider that y and x exhibit a (plausible) mirroring relationship if, for each (t, q) ($t \in [\![d_1]\!]$, $q \in [\![d_2]\!]$), $\tilde{y}_{tq}$ is strongly negatively correlated with $\tilde{x}_{tq}$, mainly, and (positively or negatively) correlated with $\tilde{x}_{(t-1)q}$ (if $t > 1$) and/or with $\tilde{x}_{t(q-1)}$ (if $q > 1$), secondarily. We thus formalize $\{x \mapsto \theta_1 x : \theta_1 \in T_1\}$ as the set of all linear mappings of the form

$$x \mapsto \tilde{\theta}_1^a \odot \tilde{x} + \tilde{\theta}_1^b \odot \begin{pmatrix} \mathbf{0}_{d_2}^\top \\ \tilde{x}_{1\bullet} \\ \vdots \\ \tilde{x}_{(d_1-1)\bullet} \end{pmatrix} + \tilde{\theta}_1^c \odot \begin{pmatrix} \mathbf{0}_{d_1} & \tilde{x}_{\bullet 1} & \cdots & \tilde{x}_{\bullet(d_2-1)} \end{pmatrix}$$

where $\tilde{\theta}_1^a$ and $\tilde{\theta}_1^b, \tilde{\theta}_1^c$ are $d_1 \times d_2$ matrices (here, $\odot$ is the componentwise multiplication). The entries of $\tilde{\theta}_1^a$ correspond to comparisons between $\tilde{x}_{tq}$ and $\tilde{y}_{tq}$ (same poly Q length q and age t). The entries of $\tilde{\theta}_1^b$ (whose first row consists of 0s) correspond to comparisons between $\tilde{x}_{(t-1)q}$ and $\tilde{y}_{tq}$ (same poly Q length q, different age t). The entries of $\tilde{\theta}_1^c$ (whose first column consists of 0s) correspond to comparisons between $\tilde{x}_{t(q-1)}$ and $\tilde{y}_{tq}$ (different poly Q length q, same age t)

In the simulation study presented in Section 4.5, the entries of $\tilde{\theta}_1^a$ are constrained to take their values in the interval $]-5, 0[$ while those of $\tilde{\theta}_1^b, \tilde{\theta}_1^c$ are constrained to take their values in $]-1/2, 1/2[$. In the simulation study presented in Section 4.5, the entries of $\tilde{\theta}_1^a$ are constrained to take their values in the interval $]-5, 0[$ while those of $\tilde{\theta}_1^b, \tilde{\theta}_1^c$ are constrained to take their values in $]-1/2, 1/2[$. The initial mapping is drawn randomly by sampling the entries of $\tilde{\theta}_1^a$ independently and uniformly in $]-5, 0[$ and, independently, by sampling the entries of $\tilde{\theta}_1^b$ and $\tilde{\theta}_1^c$ independently and uniformly in $]-1/2, 1/2[$.

In the illustration of the WTOT-matching algorithm presented in Section 4.6.2, the mapping $\widehat{\theta}$ is parametrized by $\tilde{\theta}$ given by

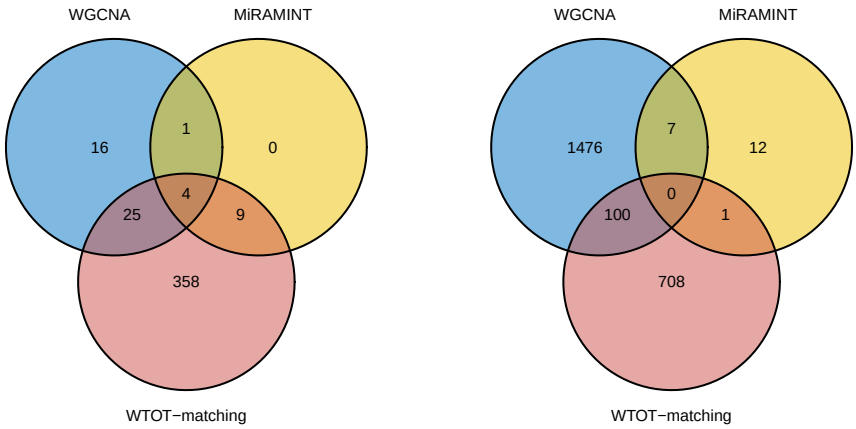

Figure 4.8 – Venn diagrams summarizing the overlaps between the sets of miRNAs (left) and mRNAs (right) which belong to a pair output by the WGCNA, MiRAMINT and WTOT-matching algorithms.

$$\tilde{\theta}_1^a = \begin{pmatrix} -0.88 & -1.47 & -0.73 \\ -0.59 & -0.90 & -0.89 \\ -0.62 & -0.70 & -1.17 \\ -0.97 & -1.30 & -0.95 \\ -0.56 & -1.16 & -1.24 \end{pmatrix}, \quad \tilde{\theta}_1^b = \begin{pmatrix} 0 & 0 & 0 \\ 0.13 & -0.19 & 0.13 \\ 0.17 & 0.09 & 0.13 \\ 0.19 & 0.09 & -0.00 \\ 0.18 & 0.15 & 0.08 \end{pmatrix},$$

$$\tilde{\theta}_1^c = \begin{pmatrix} 0 & 0.18 & -0.18 \\ 0 & 0.19 & 0.17 \\ 0 & 0.04 & 0.15 \\ 0 & 0.05 & 0.11 \\ 0 & 0.18 & 0.14 \end{pmatrix}, \quad \theta_2 = \begin{pmatrix} -0.01 & 0.01 & -0.00 \\ 0.00 & 0.01 & 0.01 \\ 0.00 & 0.01 & 0.00 \\ 0.01 & 0.01 & 0.01 \\ -0.01 & 0.01 & 0.01 \end{pmatrix}$$

(the numbers are rounded to two decimal places). We note that:

- On the one hand, the entries of $\tilde{\theta}_1^a$ are distributed around -1. On the other hand, the entries of θ_2 are small. This is in line with the *strong* biological hypothesis (that is, if a miRNA induces the degradation of a target mRNA or blocks its translation into proteins, or both, then the profile of the former should be similar to minus the profile of the latter).
- The entries of $\tilde{\theta}_1^b$ and $\tilde{\theta}_1^c$ are small.

	the WTOT(...) algorithms					the CCOT(...) algorithms	
	WTOT-SCC1*	WTOT-SCC1	WTOT-SCC2*	WTOT-SCC2	WTOT-BC*	CCOT-GWD	CCOT-GWB
A1	0	0.068 ± 0.126	0	0.068 ± 0.126	0	0.054 ± 0.14	0.092 ± 0.15
A2	0 ± 0.001	0.014 ± 0.029	0 ± 0.001	0.016 ± 0.035	0.033 ± 0.125	0.105 ± 0.13	0.121 ± 0.146
A3	0.005 ± 0.005	0.189 ± 0.175	0.0182 ± 0.033	0.233 ± 0.179	0.029 ± 0.087	0.612 ± 0.03	0.532 ± 0.068
A4	0.326 ± 0.064	0.282 ± 0.232	0.257 ± 0.256	0.393 ± 0.164	0.05 ± 0.093	0.507 ± 0.123	0.522 ± 0.116

Table 4.5 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of the co-clustering discrepancy obtained for configurations A1, A2, A3, A4.

3cm

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
A1	10	7.825 ± 0.091	1.0 ± 0.0	0.118 ± 0.001	1.0 ± 0.0
A1	35	29.373 ± 0.261	1.0 ± 0.0	0.442 ± 0.003	1.0 ± 0.0
A1	65	60.649 ± 0.998	0.999 ± 0.002	0.913 ± 0.014	1.0 ± 0.0
A1	75	67.418 ± 0.9	0.981 ± 0.006	0.991 ± 0.013	0.994 ± 0.002
A1	95	76.335 ± 1.282	0.888 ± 0.014	1.0 ± 0.0	0.957 ± 0.005
A1	150	97.049 ± 1.182	0.727 ± 0.012	1.0 ± 0.0	0.879 ± 0.005

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
A4	10	6.964 ± 0.161	0.998 ± 0.003	0.089 ± 0.003	1.0 ± 0.0
A4	35	28.632 ± 0.668	0.995 ± 0.009	0.374 ± 0.01	1.0 ± 0.0
A4	65	54.653 ± 0.927	0.986 ± 0.011	0.668 ± 0.018	0.998 ± 0.002
A4	75	61.193 ± 0.724	0.963 ± 0.016	0.709 ± 0.022	0.993 ± 0.003
A4	95	75.856 ± 0.749	0.893 ± 0.017	0.768 ± 0.022	0.975 ± 0.003
A4	150	121.273 ± 3.63	0.783 ± 0.025	0.976 ± 0.023	0.936 ± 0.011

Table 4.6 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of $\tilde{k}_r$, precision, sensitivity and specificity of the m-specific matchings averaged across all mRNAs for configuration A1 (left) and A4 (right).

3cm

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
A1	75	67.418 ± 0.9	0.981 ± 0.006	0.991 ± 0.013	0.994 ± 0.002
A2	130	100.217 ± 2.127	0.976 ± 0.017	0.894 ± 0.027	0.995 ± 0.004
A3	120	82.764 ± 1.105	0.881 ± 0.015	0.902 ± 0.025	0.968 ± 0.004
A4	120	97.561 ± 1.836	0.821 ± 0.015	0.853 ± 0.025	0.95 ± 0.005

	$k = k'$	$\tilde{k}_c$	precision	sensitivity	specificity
A1	75	67.418 ± 0.9	0.982 ± 0.006	0.991 ± 0.015	0.994 ± 0.002
A2	130	100.217 ± 2.127	0.984 ± 0.012	0.894 ± 0.028	0.995 ± 0.004
A3	120	110.352 ± 1.473	0.878 ± 0.017	0.9 ± 0.024	0.967 ± 0.004
A4	120	97.561 ± 1.836	0.84 ± 0.018	0.853 ± 0.026	0.951 ± 0.004

Table 4.7 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of $\tilde{k}_r$ or $\tilde{k}_c$, precision, sensitivity and specificity of the m-specific matchings (left) and n-specific matchings (right) averaged across all mRNAs (left) and all miRNAs (right).

	the WTOT(...) algorithms					the CCOT(...) algorithms	
	WTOT-SCC1*	WTOT-SCC1	WTOT-SCC2*	WTOT-SCC2	WTOT-BC*	CCOT-GWD	CCOT-GWB
B1	0.062 ± 0.151	0.204 ± 0.221	0.082 ± 0.161	0.204 ± 0.221	0.049 ± 0.125	0.276 ± 0.204	0.53 ± 0.168
B2	0.114 ± 0.108	0.418 ± 0.265	0.178 ± 0.207	0.455 ± 0.258	0.382 ± 0.121	0.477 ± 0.14	0.523 ± 0.115
B3	0.175 ± 0.086	0.724 ± 0.236	0.163 ± 0.082	0.775 ± 0.176	–	0.858 ± 0.042	0.867 ± 0.044
B4	0.174 ± 0.092	0.747 ± 0.196	0.171 ± 0.112	0.782 ± 0.159	–	0.882 ± 0.041	0.883 ± 0.04

Table 4.8 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of the co-clustering discrepancy obtained for configurations B1, B2, B3, B4.

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
B1	60	48.578 ± 5.201	0.885 ± 0.209	0.658 ± 0.191	0.985 ± 0.025
B1	80	63.96 ± 6.126	0.851 ± 0.199	0.816 ± 0.222	0.968 ± 0.03
B1	85	67.537 ± 6.193	0.837 ± 0.193	0.842 ± 0.222	0.961 ± 0.031
B1	90	71.214 ± 6.208	0.823 ± 0.186	0.864 ± 0.219	0.953 ± 0.031
B1	110	85.833 ± 6.358	0.753 ± 0.156	0.918 ± 0.202	0.913 ± 0.029

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
B4	10	6.78 ± 0.259	0.926 ± 0.102	0.321 ± 0.046	0.999 ± 0.001
B4	20	15.163 ± 0.619	0.873 ± 0.091	0.72 ± 0.087	0.996 ± 0.003
B4	25	19.033 ± 0.784	0.817 ± 0.084	0.837 ± 0.084	0.991 ± 0.004
B4	30	22.889 ± 0.997	0.754 ± 0.076	0.907 ± 0.077	0.984 ± 0.005
B4	40	31.118 ± 1.086	0.618 ± 0.053	0.969 ± 0.049	0.963 ± 0.005

Table 4.9 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of $\tilde{k}_r$, precision, sensitivity and specificity of the m-specific matchings averaged across all mRNAs for configurations B1 (left) and B4 (right).

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
B1	85	67.537 ± 6.193	0.837 ± 0.193	0.842 ± 0.222	0.961 ± 0.031
B2	60	48.282 ± 3.449	0.751 ± 0.194	0.838 ± 0.2	0.979 ± 0.022
B3	25	19.546 ± 1.151	0.833 ± 0.136	0.837 ± 0.152	0.992 ± 0.006
B4	25	19.033 ± 0.784	0.817 ± 0.084	0.837 ± 0.084	0.991 ± 0.004

	$k = k'$	$\tilde{k}_c$	precision	sensitivity	specificity
B1	85	63.732 ± 8.642	0.844 ± 0.175	0.836 ± 0.229	0.96 ± 0.033
B2	60	44.349 ± 2.495	0.792 ± 0.218	0.819 ± 0.227	0.971 ± 0.024
B3	25	18.766 ± 0.97	0.847 ± 0.125	0.833 ± 0.152	0.991 ± 0.005
B4	25	18.833 ± 0.793	0.834 ± 0.087	0.827 ± 0.099	0.99 ± 0.005

Table 4.10 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of $\tilde{k}_r$ or $\tilde{k}_c$, precision, sensitivity and specificity of the m-specific matchings (left) and n-specific matchings (right) averaged across all mRNAs (left) and all miRNAs (right).

	the WTOT(...) algorithms					the CCOT-(...) algorithms	
	WTOT-SCC1*	WTOT-SCC1	WTOT-SCC2*	WTOT-SCC2	WTOT-BC*	CCOT-GWD	CCOT-GWB
C1	0.106 ± 0.1	0.203 ± 0.135	0.101 ± 0.056	0.194 ± 0.116	0.265 ± 0.255	0.496 ± 0.16	0.902 ± 0.007
C2	0.209 ± 0.131	0.252 ± 0.182	0.262 ± 0.141	0.345 ± 0.205	–	0.938 ± 0.023	0.971 ± 0.026
C3	0.609 ± 0.113	0.693 ± 0.154	0.385 ± 0.151	0.521 ± 0.198	–	0.926 ± 0.027	0.987 ± 0.002
C4	0.63 ± 0.141	0.751 ± 0.145	0.435 ± 0.197	0.6 ± 0.233	–	0.939 ± 0.027	0.987 ± 0.002

Table 4.11 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of the co-clustering discrepancy obtained for configurations C1, C2, C3, C4.

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
C1	10	7.748 ± 0.446	0.973 ± 0.03	0.156 ± 0.01	1.0 ± 0.0
C1	30	25.888 ± 1.418	0.972 ± 0.029	0.526 ± 0.032	1.0 ± 0.0
C1	50	45.521 ± 2.441	0.944 ± 0.025	0.916 ± 0.04	1.0 ± 0.001
C1	55	49.108 ± 3.018	0.93 ± 0.021	0.972 ± 0.025	0.999 ± 0.002
C1	60	51.365 ± 3.335	0.919 ± 0.024	0.993 ± 0.011	0.997 ± 0.004
C1	70	55.296 ± 3.312	0.881 ± 0.034	1.0 ± 0.0	0.985 ± 0.01

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
C4	5	3.293 ± 0.096	0.895 ± 0.023	0.185 ± 0.012	1.0 ± 0.0
C4	10	7.278 ± 0.303	0.899 ± 0.022	0.474 ± 0.029	1.0 ± 0.0
C4	15	11.982 ± 0.578	0.888 ± 0.02	0.787 ± 0.04	1.0 ± 0.0
C4	20	15.935 ± 0.864	0.843 ± 0.022	0.96 ± 0.023	0.997 ± 0.001
C4	25	19.138 ± 0.89	0.762 ± 0.032	0.997 ± 0.005	0.989 ± 0.003
C4	30	22.578 ± 1.11	0.671 ± 0.04	1.0 ± 0.0	0.978 ± 0.004

Table 4.12 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of $\tilde{k}_r$, precision, sensitivity and specificity of the m-specific matchings averaged across all mRNAs for configurations C1 (left) and C4 (right).

	$k = k'$	$\tilde{k}_r$	precision	sensitivity	specificity
C1	55	49.108 ± 3.018	0.93 ± 0.021	0.972 ± 0.025	0.999 ± 0.002
C2	20	16.203 ± 0.956	0.955 ± 0.016	0.965 ± 0.021	0.997 ± 0.001
C3	20	15.552 ± 0.877	0.854 ± 0.024	0.968 ± 0.019	0.997 ± 0.001
C4	20	15.935 ± 0.864	0.843 ± 0.022	0.96 ± 0.023	0.997 ± 0.001

	$k = k'$	$\tilde{k}_c$	precision	sensitivity	specificity
C1	55	49.056 ± 3.461	0.898 ± 0.06	0.971 ± 0.026	0.981 ± 0.009
C2	20	16.371 ± 0.812	0.953 ± 0.018	0.963 ± 0.023	0.997 ± 0.001
C3	20	15.879 ± 0.691	0.804 ± 0.025	0.969 ± 0.018	0.993 ± 0.001
C4	20	15.867 ± 0.635	0.812 ± 0.032	0.961 ± 0.021	0.993 ± 0.002

Table 4.13 – Mean ($\pm$ standard deviation) computed across the 30 independent replications of $\tilde{k}_r$ or $\tilde{k}_c$, precision, sensitivity and specificity of the m-specific matchings (left) and n-specific matchings (right) averaged across all mRNAs (left) and all miRNAs (right).

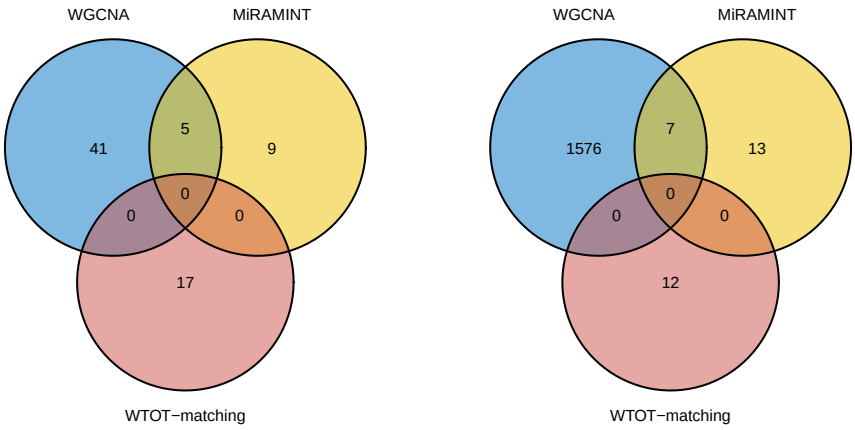

Figure 4.9 – Venn diagrams summarizing the overlaps between the sets of miRNAs (left) and mRNAs (right) which belong to a pair output by the WGCNA, MiRAMINT and WTOT-matching algorithms, *focusing on the WTOT-matching matchings labeled as peaked*.

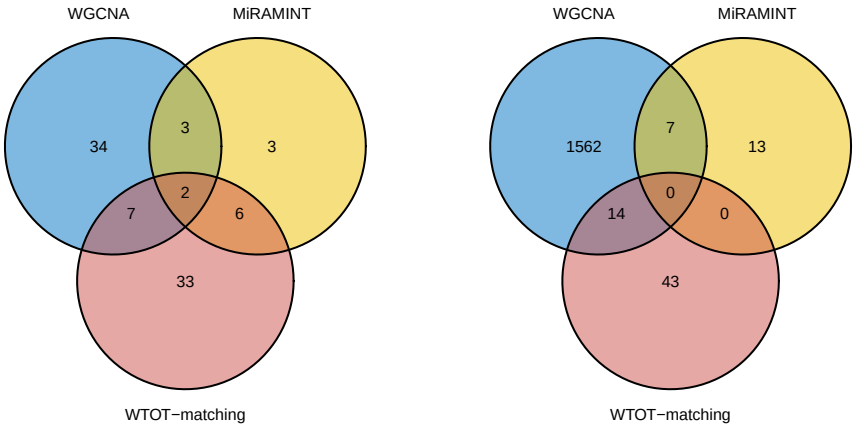

Figure 4.10 – Venn diagrams summarizing the overlaps between the sets of miRNAs (left) and mRNAs (right) which belong to a pair output by the WGCNA, MiRAMINT and WTOT-matching algorithms, *focusing on the WTOT-matching matchings labeled as monotonic*.

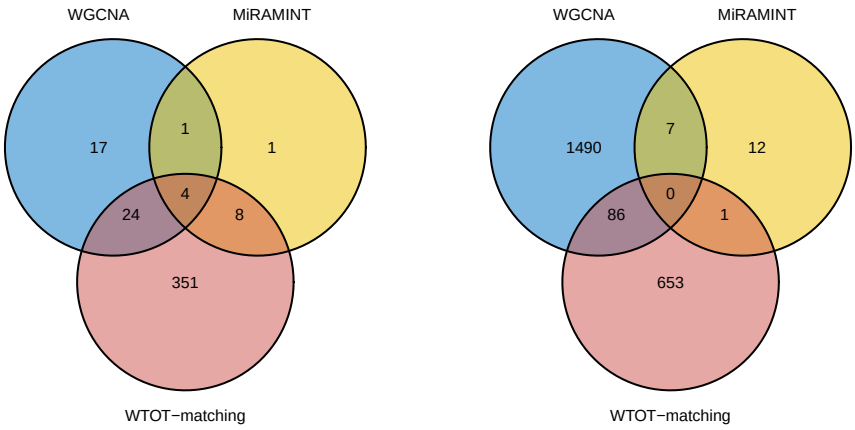

Figure 4.11 – Venn diagrams summarizing the overlaps between the sets of miRNAs (left) and mRNAs (right) which belong to a pair output by the WGCNA, MiRAMINT and WTOT-matching algorithms, *focusing on the WTOT-matching matchings which are labeled as neither peaked nor monotonic*.

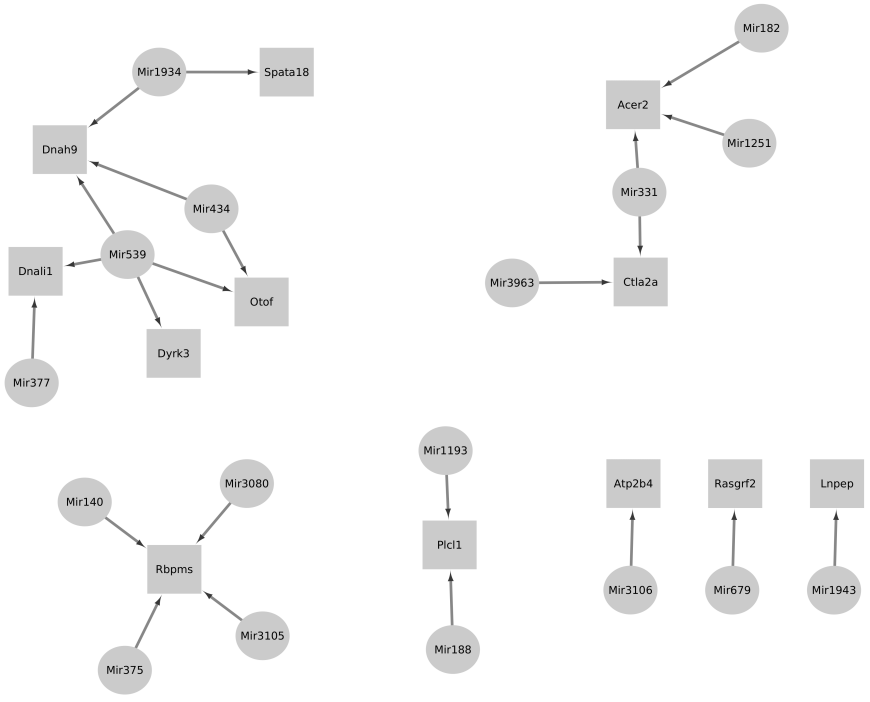

Figure 4.12 – The mRNA-miRNA networks based on the mRNA-miRNA matchings output by the WTOT-matching algorithm, *focusing on the matchings which are labeled as peaked*. Disks correspond to miRNAs and squares to mRNAs. The top annotation is *conventional motile cilium* (GO:0097729, 3 hits).

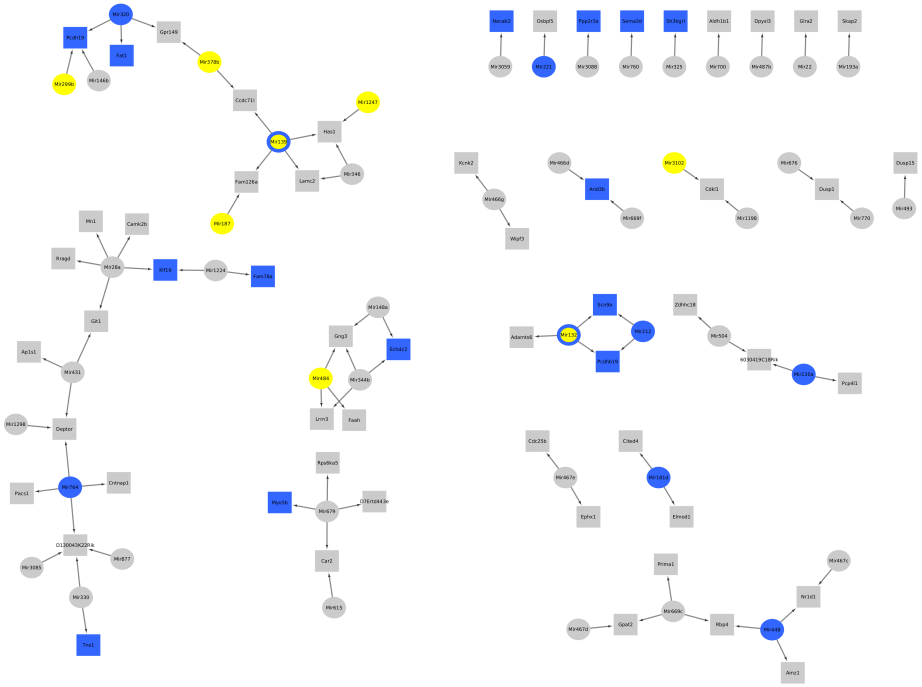

Figure 4.13 – The mRNA-miRNA networks based on the mRNA-miRNA matchings output by the WTOT-matching algorithm, *focusing on the matchings which are labeled as monotonic*. Disks correspond to miRNAs and squares to mRNAs. Elements also retained by the WGCNA algorithm (respectively, the MiRAMINT algorithm) are indicated in blue (respectively, yellow). The top annotation is *mitigation of host antiviral defense response* (GO:0050690, 2 hits).

Algorithm 1 *Master optimal transport algorithm.*

Input: X, Y, minibatch sizes $\widetilde{M}, \widetilde{N}$, decay rate $\eta \in]0,1]$, initial regularization parameter γ_0, initial mapping $\theta_0 \in \Theta$, maximal number of iterations T

Output: Transport coupling $\tilde{P}_T \in (\mathbb{R}_+)^{M \times N}$, mapping $\theta_T \in \Theta$, weight ω_T

Compute:

- $\underline{\gamma} = \text{mean}\{\|x - x'\|_2 : x, x' \in X\}$ {for entropy regularization}
- $h = \text{mean}\{\|y - y'\|_2 : y, y' \in Y\}$ {for window calibration}

Set $t \leftarrow 0$

Set stop $\leftarrow$ FALSE

while $\neg$ stop or $t < T$ **do**

$\gamma_t \leftarrow \max(\gamma_0 \times \eta^t, \underline{\gamma})$

Sample uniformly a minibatch of $\widetilde{M}$ observations $\tilde{x}_{1:\widetilde{M}} := (\tilde{x}_1, \ldots, \tilde{x}_{\widetilde{M}})$ from X

Sample uniformly a minibatch of $\widetilde{N}$ observations $\tilde{y}_{1:\widetilde{N}} := (\tilde{y}_1, \ldots, \tilde{y}_{\widetilde{N}})$ from Y

Define and compute $\theta_t(\tilde{x}_{1:\widetilde{M}}) := \big(\theta_t(\tilde{x}_1), \ldots, \theta_t(\tilde{x}_{\widetilde{M}})\big)$

Define and compute $\omega_t \in (\mathbb{R}_+)^{\widetilde{M}}$ such that $\sum_{m \in [\![\widetilde{M}]\!]} (\omega_t)_m = 1$ by setting

$$(\omega_t)_m \propto \sum_{n \in [\![\widetilde{N}]\!]} \varphi\left(\frac{\tilde{y}_n - \theta_t(\tilde{x}_m)}{h}\right) \quad (\text{all } m \in [\![\widetilde{M}]\!])$$

where φ is the standard normal density

Define $\mu^{\omega_t}_{\theta_t(\tilde{x}_{1:\widetilde{M}})}$, the ω_t-weighted empirical measure attached to $\theta_t(\tilde{x}_{1:\widetilde{M}})$, and $\nu_{\tilde{y}_{1:\widetilde{N}}}$, the empirical measure attached to $\tilde{y}_{1:\widetilde{N}}$

Compute $\text{Loss}_t = \bar{\mathcal{W}}_{\gamma_t}\left(\mu^{\omega_t}_{\theta_t(\tilde{x}_{1:\widetilde{M}})}, \nu_{\tilde{y}_{1:\widetilde{N}}}\right)$ and ∇Loss_t, the gradient of Loss_t relative to the parameter defining θ_t {relies on the Sinkhorn-Knopp algorithm}

Update the parameter defining θ_t by performing one step of stochastic gradient descent, yielding θ_{t+1}

Check stopping criterion and update stop variable accordingly

$t \leftarrow t + 1$

end while

Set $\theta_T \leftarrow \theta_{t-1}$

Set $\gamma_T \leftarrow \gamma_{t-1}$

Define and compute $\omega_T \in (\mathbb{R}_+)^M$ such that $\sum_{m \in [\![M]\!]} (\omega_T)_m = 1$ by setting

$$(\omega_T)_m \propto \sum_{n \in [\![N]\!]} \varphi\left(\frac{y_n - \theta_T(x_m)}{h}\right) \quad (\text{all } m \in [\![M]\!])$$

Compute $\tilde{P}_T \in \Pi(\omega_T)$ solving $\min_{P \in \Pi(\omega_T)} \mathcal{W}_{\gamma_T}\left(\mu^{\omega_T}_{\theta_T(X)}, \nu_Y\right)$

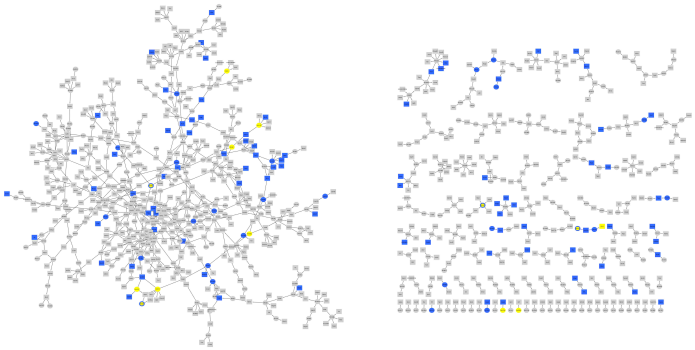

Figure 4.14 – The mRNA-miRNA networks based on the mRNA-miRNA matchings output by the WTOT-matching algorithm, *focusing on the matchings which are labeled as neither peaked nor monotonic*. Disks correspond to miRNAs and squares to mRNAs. Elements also retained by the WGCNA algorithm (respectively, the MiRAMINT algorithm) are indicated in blue (respectively, yellow). The top annotation is *extracellular matrix organization* (GO:0030198, 22 hits).

5

Making sparse predictions, and anticipating the requests of declaration of natural disasters for a drought event in France

Drought events, the phenomenon of clay swelling and shrinking in humid and dry conditions, rank as the second most costly natural disaster within the French legal framework of the natural disaster compensation scheme. A critical aspect of the national compensation scheme involves cities submitting requests for the government declaration of natural disaster for a drought event as a key step. This chapter is dedicated to the challenge that we take up of forecasting which cities will submit such requests.

The problem can be tackled as a classification task, leveraging the power of classification algorithms. Taking a slightly different perspective, we introduce an alternative procedure that hinges on OT theory and iPiano, an inertial proximal algorithm for nonconvex optimization. The optimization problem is designed so as to yield a sparse vector of predictions because it is known that relatively few cities will submit requests. Additionally, we develop a hybrid procedure that synergistically combines and utilizes both types of predictions (that is, those made based on classification algorithms and those yielded by the alternative procedure), resulting in enhanced forecasting accuracy.

A simulation study illustrates the procedures. The real data application is presented and discussed in details. The convergence of the iPiano algorithm is established, using the notion of o-minimal structures.

This chapter is based on a joint work with G. Ecoto (Ph.D. candidate under the supervision of A. Chambaz) and A. Chambaz. The project was funded by Université Paris Cité and Caisse Centrale de Réassurance.

> My main contribution has consisted in developing the methodology, formally and computationally, and performing the data analysis with Geoffrey Ecoto. This chapter will

be submitted soon as a technical report then to an international journal.

5.1 Introduction

We define a drought event in this study as the phenomenon of clay shrinking and swelling during a calendar year. For a comprehensive introduction to drought events and their economic consequences, we refer to (Charpentier et al., 2022b, Sections 1 and 2). In brief, the clay in the soil undergoes alternating shrinkage and swelling in dry and humid conditions, leading to instabilities and cracks in buildings. The costs incurred by these cracks are covered by all private property insurance policies (MTES, 2016). As 90% of the French natural disasters insurance market is reinsured by Caisse Centrale de Réassurance (henceforth abbreviated as CCR) (CCR, 2022), a public-sector reinsurer providing coverage against natural catastrophes and uninsurable risks, the French state ultimately bears the risk.

Due to intricacies of the French legal framework (known as the natural disasters compensation scheme, see Charpentier et al., 2022b, Section 2.1), two prerequisites must be met in order to initiate the compensation scheme. Firstly, the property that has been lost and/or damaged must be covered by a property and casualty insurance policy, which is a condition of private nature. Secondly, a government decree declaring a natural disaster must be published in the Official Journal, which is a condition of public nature. The responsibility of initiating the request for the government declaration of a natural disaster for the cities they administer lies with the mayors. Of note, we adopt here and henceforth the term "city" regardless of the size of the *commune*, encompassing a wide range from small hamlets to large urban centers.

Forecasting the cost of drought events in France is a critical task for CCR. CCR currently addresses two sub-problems separately: sub-problem 1 involves predicting which cities will submit a request for the government declaration of natural disaster for a drought event, while sub-problem 2 is centered on predicting the cost of a drought event for those cities that have already obtained the government declaration of natural disaster for a drought event. In this study, we concentrate on sub-problem 1. (Ecoto et al., 2021; Ecoto and Chambaz, 2022) focus on sub-problem 2. In contrast, (Chatelain and Loisel, 2021) takes on both sub-problems simultaneously. On the other hand, (Charpentier et al., 2022b; Heranval et al., 2022) predict which cities will experience claims (a proxy for sub-problem 1) and subsequently estimate the cost for these cities. We acknowledge that the problem we address in this study is, therefore, more narrowly focused than those studied in (Chatelain and Loisel, 2021; Charpentier et al., 2022b; Heranval et al., 2022).

Quoting (Logar and van den Bergh, 2011, page 4, first paragraph), "[t]he existing literature on the costs of drought [events] is scarce, fragmented and heterogeneous and there is a need for comprehensive costs estimations to help designing effective policy responses." To the best of our knowledge, (Chatelain and Loisel, 2021; Charpentier et al., 2022b; Heranval et al., 2022; Ecoto et al., 2021; Ecoto and Chambaz, 2022) are the only five references available about the prediction of the cost of drought events, thus susceptible to address the problem of predicting which cities will submit a request for the government declaration of natural disaster for a drought event. It is worth noting that studies conducted by insurance companies are often kept confidential, further emphasizing the scarcity of available literature on this subject.

In (Chatelain and Loisel, 2021), the authors use Generalized Linear Models (GLM) and the extreme gradient boosting algorithm to predict which cities will submit a request for the government declaration of natural disaster for a drought event (see Section 3.1 therein). We

also tackle the problem as a classification task, leveraging the power of classification algorithms. However, taking a slightly different perspective, our main contribution consists in introducing an alternative procedure that hinges on optimal transport theory and an inertial proximal algorithm for nonconvex optimization. The optimization problem is designed so as to yield a sparse vector of predictions because it is known that relatively few cities will submit requests. Additionally, we develop a hybrid procedure that synergistically combines and utilizes both types of predictions (that is, those made based on classification algorithms and those yielded by the alternative procedure).

The rest of the study is organized as follows. Section 5.2 introduces the data set that we obtained by merging several data sets, some of which either provided by CCR's cedents* while others were collected from other trusted sources. This section also outlines the statistical challenge that we undertake and presents insights into the data. Section 5.3 is a modicum of optimal transport theory. Section 5.4 exposes our novel procedure to make sparse predictions and discusses how to solve the nonconvex optimization task that sits at its core using the algorithm iPiano (Ochs et al., 2015), from both theoretical and computational perspectives. Section 5.5 presents a simulation study and introduces the hybrid procedure. Section 5.6 describes the full-fledged application to the challenge of forecasting which cities will submit a request for the government declaration of natural disaster for a drought event. In the appendix, Section 5.7 gathers the proofs of the convergence of the iPiano algorithm using a theorem proven in (Ochs et al., 2015). The Kurdyka-Lojasiewicz property (Attouch et al., 2010) and notion of o-minimal structures (Wilkie, 1996) play a central role.

5.2 Data and statistical challenge

5.2.1 Presentation of the data, first pass

The data set is obtained by merging several data sets, either provided by CCR's cedents or collected from other sources, namely the National Institute for Statistical and Economic Studies (Insee), Geographic National Institute (IGN), French Geological Survey (BRGM) and Météo-France. While there are numerous similarities between the present data set and the one comprehensively presented and used in (Ecoto and Chambaz, 2022, see Section 2), there are also major differences.

From now on, France refers to *Metropolitan* or *Mainland* France, and the adjective French to what is related to France with the restricted acceptation of the word. This is justified because drought events are not a threat in Overseas France (essentially because there is little clay in these parts of the country).

The experimental units are the French cities. Each of them can contribute a data structure for a given year t (by convention, $t = 1, 2, 3$ respectively correspond to years 2019, 2020 and 2021) and a given week u (the integer $u \in \mathcal{U}_t \subset \mathbb{N}^*$ being the number of weeks starting from the first week of year t, with $44 \leq u \leq 85$). A data structure encompasses multiple aspects of a city's profile, aiming to provide a comprehensive representation of its context and potential triggers for requesting the government declaration of natural disaster for a drought event. It consists of the following blocks of variables:

City description (16 variables). This block provides detailed information about the city, covering various aspects such as housing stock age, housing stock exposure to clay-

*A cedent is a party in an insurance contract that passes the financial obligation for certain potential losses to the insurer. In return for bearing a particular risk of loss, the cedent pays a reinsurance premium.

shrinkage-swelling hazard, and climatic zone. By capturing these variables, a holistic understanding of the city's characteristics is obtained.

City exposure to drought events (25 variables). The variables within this block outline the city's exposure to drought events. They build upon the Soil Wetness Index (SWI), and include an indicator of whether or not the city is eligible for the government declaration of natural disaster for a drought event.

City history of requests (12 variables). This block provides a record of the city's previous requests for the government declaration of natural disaster for a drought event, including information on the success or failure of the requests. The record gives us insight into the city's decision-making process, intentions and actions regarding the submission of a request for the government declaration of natural disaster for a drought event.

City current request status (1 variable). This variable indicates whether or not the city submitted a request for the government declaration of natural disaster for a drought event for year t during week u or before.

City's vicinity description (13 variables). This block focuses on the city's surroundings. It provides information about the neighboring cities' claims and requests for the government declaration of natural disaster for a drought event.

5.2.2 Presentation of the data, second pass

DESCRIPTION OF A CITY. The description of a city notably consists of its population, of the (estimated) number of houses located within the city's limits (the estimation is based on census data: Insee, 2000), of the city's average altitude and area (source: IGN, 2018), house density (defined as the ratio of the number of houses to the city's area), and proportions of buildings built prior to 1949, between 1950 and 1974, between 1975 and 1989, and after 1989 (the proportions are computed based on data found in Insee, 2000). In addition, the description of the city also includes the proportions of houses located within the city's limits that fall in each of the four clay-shrinkage-swelling hazard categories (as defined by, and obtained from BRGM: MI, 2019); the city's seismic zone (a four-category variable attributed to each city by the French *Code de l'environnement*); the climatic zone of the city's department (the French State attributes to each department this five-category variable; a department is a level of government between the administrative regions and communes).

Up to now, the variables that we listed are essentially static. The description of the city is completed by the (estimated) insured sum corresponding to the houses located within its limits. The estimations are based on data from Insee and portfolios data provided by CCR's cedents. This last piece of information depends on the year, but the variations from one year to another are limited.

To conclude, let us stress that the age of the housing stock is used here as a proxy for the house building technology, an important factor to consider because some buildings are more vulnerable than others (France Assureurs, 2022, page 28). Furthermore, accounting for clay concentration is mandatory since it is the clay present in the soil that, by shrinking and swelling in dry and humid conditions, creates instabilities and generates cracks in buildings.

DESCRIPTION OF A CITY'S EXPOSURE TO DROUGHT EVENTS. The description of a city's exposure to drought events builds upon the SWI in a manner presented almost comprehensively

in (Ecoto and Chambaz, 2022, Section 2.3.2). For self-containedness, we recall here the main elements of the presentation.

Provided by Météo-France since 1959, the SWI data consist of time series of values (one value every ten-day period) ranging between -3.33 (very dry soil) and 2.33 (very wet soil). There are as many SWI time series as the number of 8×8 km^2 squares used by Météo-France to partition the French territory.

Note that for any year t and week $u \in \mathcal{U}_t \cap [\![44, 52]\!]$ (that is, before the end of year t), we necessarily have access to fewer than 37 values of the SWI for year t. We use a prediction model to predict future values of the SWI so that all the time series of SWI cover the whole year. As u increases, the predicted values are replaced by the actual values provided by Météo-France, until the complete time series for year t are all observed.

For every year t and every city, we then derive a city-specific SWI time series by taking the convex average of the possibly completed SWI time series attached to the squares that overlap the city's area, the weights being proportional to the areas of the intersections. The description of a city's exposure to drought events for year t builds upon the corresponding SWI time series. It notably consists of the minimum value of the SWI time series, of the overall average of the time series, of the averages restricted to the first, second, third and fourth quarters of year t respectively (that is, January-March, April-June, July-September, October-December), and of the averages restricted to the unions of the second and third quarters (April-September) or of the first, second and third quarters (January-September). The description is complemented by measures of how exceptional the monthly and quarterly average SWI (say $\overline{\text{SWI}}$) are relative to historical SWI data. Specifically, for every month (respectively, every quarter), we compute the empirical cumulative distribution function of the monthly (respectively, quarterly) average SWI using all data for the city of interest from 1959 to 2009 and then evaluate that function at $\overline{\text{SWI}}$. The smaller is the resulting proportion, the more pronounced is the soil dryness and, conversely, the larger is the resulting proportion, the more pronounced is the soil wetness. Moreover, the description includes an indicator of whether or not the city is eligible for a government declaration of natural disaster for a drought event.

This description holds utmost relevance as it focuses on the critical role of soil humidity in causing the shrinkage and swelling of clay, eventually leading to instabilities and the formation of cracks in buildings.

REQUESTS FOR THE GOVERNMENT DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT. Being the secretary of the Commission Interministérielle Catastrophe Naturelle, CCR has been having access, since 1989, to the requests for the government declaration of natural disaster for a drought event as they accrue. Formally, a city can submit a request for the government declaration of natural disaster for a drought event for year t until the end of June of year $(t + 2)$. However, anticipating which cities will submit a request for year t is only a necessity typically between the months of November of year t and of September of year $(t + 1)$.

DESCRIPTION OF A CITY'S REQUEST HISTORY. Given a year t and a week u, the (t, u)-specific description of a city's request history consists of t and u, of the overall number of French cities that submitted a request for year t during week u or before, and of the ratio of the logarithm of that overall number to u. In addition, the description includes the number of requests submitted by the city since 1990 (respectively, between years $(t - 4)$ and t), the number of times the city obtained the government declaration of natural disaster for a drought event since 1990 (respectively, between years $(t - 4)$ and t), and the ratio of the

aforementioned number of requests submitted by the city since 1990 to the number of years between 1990 and year t. Moreover, the description includes an indicator of whether or not the city was denied the government declaration of natural disaster for a drought event on year $(t-1)$, and the numbers of denied requests between $(t-2)$ and $(t-1)$ and between $(t-4)$ and $(t-1)$.

This description holds significant relevance, primarily due to its ability to provide valuable insights into the city's inclination to submit a request for a government declaration of natural disaster for a drought event. By examining the city's historical pattern of submitting such requests since 1990 or within the previous five years, regardless of their success, we can gather essential information about the city's familiarity with the administrative procedure. Additionally, this serves as a proxy for assessing the city's exposure to drought events.

DESCRIPTION OF A CITY'S VICINITY. Using the flux of requests, we compile a collection of variables describing the vicinity of a city. The variables concern either the neighboring cities or, more broadly, the cities in the same department. Given a year t and a week u, the (t,u)-specific collection notably consists of the following five numbers: the number of neighboring cities that requested the government declaration of natural disaster for a drought event for year t during week u or before, the number of neighboring cities (respectively, of cities in the same department) that submitted such a request *for the first time* for year t, and the number of neighboring cities (respectively, of cities in the same department) that submitted such a request *for the first time* between years $(t-4)$ and t. The collection is complemented by the ratios of the four last numbers to either the number of neighboring cities or the number of cities in the same department. In addition, the collection also includes the number of claims for year t made during week u or before by the neighboring cities (respectively, by the cities of the same department), and the ratio of that number to the number of neighboring cities (respectively, of cities in the same department).

To conclude, it is important to emphasize the potential relevance of these variables for several compelling reasons. For instance, it is common for mayors of neighboring cities to exchange information, particularly if their cities are part of the same federation of municipalities. This interconnectedness means that if a city submits a request for a government declaration of natural disaster for a drought event, then that raises the likelihood that neighboring cities will do the same, either in the same year or later. Furthermore, it is worth noting that drought events are not necessarily confined to a single city's territory. Even if the mayors do not actively share information, the occurrence of a drought event in one city that prompts the submission of a request for a government declaration of natural disaster for a drought event increases the likelihood that a similar drought event has taken place in nearby areas. Consequently, the likelihood of submitting a request for such a declaration also increases in those affected vicinity areas.

5.2.3 The statistical challenge and some facts about the data

As elaborated in Section 5.2.1, each French city can contribute a data structure for a given year t and a given week u (the integer u being the number of weeks starting from the first week of year t). It is worth mentioning that the composition of the set of French cities undergoes slight changes from one year to another. To address this variability, we define $\mathcal{A}_t$ as the set of cities for year t (with the aforementioned convention $t = 1, 2, 3$ for years 2019, 2020 and 2021, respectively). Furthermore, we introduce $\mathcal{U}_t$ as the comprehensive list of weeks during which CCR received the latest submissions of a request for the government declaration of natural disaster for a drought event for year t, encompassing a period of up

to 85 weeks following the first week of year t.

We report that $\operatorname{card}\mathcal{A}_1 = \operatorname{card}\mathcal{A}_2 = 34,841$ and $\operatorname{card}\mathcal{A}_3 = 34,836$. Moreover,

$$\mathcal{U}_1 = \{44, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 59, 60, 61, 69, 75\},$$
$$\mathcal{U}_2 = \{48, 49, 50, 51, 53, 54, 55, 56, 58, 59, 60, 61, 63, 65, 67, 68, 69, 70, 71, 73, 75, 78, 81, 85\},$$
$$\mathcal{U}_3 = \{49, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 71, 72, 73, 77, 78\}.$$

For every year $t = 1, 2, 3$ and each week $u \in \mathcal{U}_t$, we let

- $\xi_{\alpha,t,u} \in \mathcal{X} \subset \mathbb{R}^d$ be city α's vector of covariates on week u relative to year t (for any city $\alpha \in \mathcal{A}_t$);
- $\zeta_{\alpha,t,u} \in \{0, 1\}$ be the indicator equal to 1 if and only if (iff) city α submitted a request *before or during* week u relative to year t (for any city $\alpha \in \mathcal{A}_t$);
- $u^- := \max\{\nu \in \mathcal{U}_t : \nu < u\}$ index the week before u in $\mathcal{U}_t$ (with convention $u^- = 0$ if $u = \min\mathcal{U}_t$), so that $(\zeta_{\alpha,t,u} - \zeta_{\alpha,t,u^-}) \in \{0, 1\}$ equals 1 iff city α submitted a request during week u relative to year t (for any city $\alpha \in \mathcal{A}_t$, with convention $\zeta_{\alpha,t,0} = 0$).

In addition we also define, for each year $t = 1, 2, 3$ and any city $\alpha \in \mathcal{A}_t$, $\zeta_{\alpha,t} \in \{0, 1\}$, the indicator equal to 1 iff city α submitted a request relative to year t (possibly after the week $\max\mathcal{U}_t$). Note that $\zeta_{\alpha,t} \geq \max_{u\in\mathcal{U}_t} \zeta_{\alpha,t,u}$. In words, some cities may submit a request for the government declaration of natural disaster for a drought event relative to year t beyond week $\max\mathcal{U}_t$. This fact is discussed further in the next paragraph.

Table 5.1 reports the quartiles of the sets

$$\left\{ \sum_{\alpha\in\mathcal{A}_t} (\zeta_{\alpha,t,u} - \zeta_{\alpha,t,u^-}) : u \in \mathcal{U}_t \right\}, \quad t = 1, 2, 3,$$

that is, the quartiles of the sets of the week-specific numbers of new requests for the government declaration of natural disaster for a drought event relative to year t, for $t = 1, 2, 3$. Table 5.1 also reports the initial numbers and proportions of requests for the government declaration of natural disaster for a drought event relative to year t (that is, $\sum_{\alpha\in\mathcal{A}_t} \zeta_{\alpha,t,\min\mathcal{U}_t}$ and $\sum_{\alpha\in\mathcal{A}_t} \zeta_{\alpha,t,\min\mathcal{U}_t} / \operatorname{card}\mathcal{A}_t$), their overall numbers and proportions at week $\max\mathcal{U}_t$ (that is, $\sum_{\alpha\in\mathcal{A}_t} \zeta_{\alpha,t,\max\mathcal{U}_t}$ and $\sum_{\alpha\in\mathcal{A}_t} \zeta_{\alpha,t,\max\mathcal{U}_t} / \operatorname{card}\mathcal{A}_t$), and the overall numbers and proportions of requests for the government declaration of natural disaster for a drought event relative to year t (that is, $\sum_{\alpha\in\mathcal{A}_t} \zeta_{\alpha,t}$ and $\sum_{\alpha\in\mathcal{A}_t} \zeta_{\alpha,t} / \operatorname{card}\mathcal{A}_t$), for $t = 1, 2, 3$. We emphasize that only 12.5% (776/6240), 11.0% (589/5335) and 4.8% (81/1696) of the requests for the government declaration of natural disaster for a drought event relative to year t were already submitted at week $\min\mathcal{U}_t$, while only 82% (5142/6240), 92.9% (4958/5335) and 69.0% (1169/1696) of the overall numbers of requests for the government declaration of natural disaster for a drought event relative to year t were submitted at week $\max\mathcal{U}_t$, for $t = 1, 2, 3$. Moreover, between the first and last weeks $\min\mathcal{U}_t$ and $\max\mathcal{U}_t$, the median numbers of newly submitted requests corresponded to 4.7% (245/5142), 3.3% (166/4958) and 4% (47/1169) of the overall numbers of requests at week $\max\mathcal{U}_t$, for $t = 1, 2, 3$.

Our ultimate objective is to achieve sequential forecasting of which cities will submit a request for the government declaration of natural disaster for a drought event leveraging past data and, in particular, knowing which cities already did. Formally, our objective is the following: for every $u \in \mathcal{U}_3$, leveraging past observations, that is

$$\{(\xi_{\alpha,t,\nu}, \zeta_{\alpha,t,\nu}, \zeta_{\alpha,t}) : t = 1, 2, \alpha \in \mathcal{A}_t, \nu \in \mathcal{U}_t \text{ st } \zeta_{\alpha,t,\nu} = 0 \text{ or } (\zeta_{\alpha,t,\nu^-}, \zeta_{\alpha,t,\nu}) = (0, 1)\}$$

numbers of new requests $(\sum_{\alpha\in\mathcal{A}_t}(\zeta_{\alpha,t,u}-\zeta_{\alpha,t,u^-}),\ u\in\mathcal{U}_t)$	2019 $(t=1)$	2020 $(t=2)$	2021 $(t=3)$
minimum	104	41	10
1st quartile	138	75	32
median	245	166	47
3rd quartile	386	208	69
maximum	776	589	129
initial number (and proportion) of requests $(\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t,\min\mathcal{U}_t})$	776 (2.2%)	589 (1.7%)	81 (0.2%)
overall number (and proportion) of requests $(\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t,\max\mathcal{U}_t})$	5142 (14.8%)	4958 (14.2%)	1169 (3.3%)
overall number (and proportion) of requests $(\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t})$	6240 (17.9%)	5335 (15.3%)	1696 (4.9%)

Table 5.1 – Summary measures of the sets $\{\sum_{\alpha\in\mathcal{A}_t}(\zeta_{\alpha,t,u}-\zeta_{\alpha,t,u^-}) : u\in\mathcal{U}_t\}$ $(t=1,2,3)$, that is, of the numbers of new requests for the government declaration of natural disaster for a drought event as weeks go by, for years 2019, 2020 and 2021 respectively. In addition, the overall numbers $\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t}$ and proportions $\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t}/\operatorname{card}\mathcal{A}_t$ $(t=1,2,3)$ of requests for the government declaration of natural disaster for a drought event relative to year t are also reported for years 2019, 2020 and 2021.

if $u=\min\mathcal{U}_3$ and otherwise

$$\{(\xi_{\alpha,t,\nu},\zeta_{\alpha,t,\nu},\zeta_{\alpha,t}) : t=1,2,\alpha\in\mathcal{A}_t,\nu\in\mathcal{U}_t \text{ st } \zeta_{\alpha,t,\nu}=0 \text{ or } (\zeta_{\alpha,t,\nu^-},\zeta_{\alpha,t,\nu})=(0,1)\}$$
$$\cup\{(\xi_{\alpha,3,\nu},\zeta_{\alpha,3,\nu},0) : \alpha\in\mathcal{A}_3,\nu\in\mathcal{U}_3,\nu<u \text{ st } \zeta_{\alpha,3,\nu}=0 \text{ or } (\zeta_{\alpha,3,\nu^-},\zeta_{\alpha,3,\nu})=(0,1)\}, \quad (5.1)$$

we wish to predict $\zeta_{\alpha,3}$ using $\xi_{\alpha,3,u}$ for every $\alpha\in\mathcal{A}_3$ such that $\zeta_{\alpha,3,u}=0$. Of note, the set defined in (5.1) when $u=\max\mathcal{U}_3$ consists of more than 2.05 million triplets. Moreover, we will not apply thresholding to the estimated probabilities with the aim of generating binary labels.

The focus on "making sparse predictions" which is explicit in the title of the manuscript is justified by the last row of Table 5.1: in 2019, 2020 and 2021, the proportions of cities that eventually submitted a request for the government declaration of natural disaster for a drought event were respectively 17.9%, 15.3% and 4.9%. Finally, promoting 0-predictions as part of the control of the sparsity of a set of predictions $\{\widehat{\zeta}^u_{\alpha,3} : \alpha\in\mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0\}$ for a week $u\in\mathcal{U}_3$ holds merit in itself. Indeed, denoting by $\mathrm{IS}_{\alpha,3}$ the 2021 (estimated) insured sum corresponding to the houses located within the limits of any city $\alpha\in\mathcal{A}_3$ (one of the entries of $\xi_{\alpha,3,u}$, see Section 5.2.2), the sum

$$\sum_{\alpha\in\mathcal{A}_3}\mathrm{IS}_{\alpha,3}\,\mathbf{1}\{\zeta_{\alpha,3,u}=1\}+\sum_{\alpha\in\mathcal{A}_3}\widehat{\zeta}^u_{\alpha,3}\mathrm{IS}_{\alpha,3}\,\mathbf{1}\{\zeta_{\alpha,3,u}=0\} \quad (5.2)$$

may be used as an estimator of 2021 drought events overall cost. The contribution to (5.2) of a single city $\alpha\in\mathcal{A}_3$ with a large $\mathrm{IS}_{\alpha,3}$ may be significant even if its prediction $\widehat{\zeta}^u_{\alpha,3}$ is small but not 0. In addition, the contribution to (5.2) of many cities with moderate insured sums may be significant even if their prediction are small but not 0.

5.3 A modicum of optimal transport theory

This section introduces the few tools from optimal transport theory that will be instrumental in developing our novel procedure in the next section.

Fix arbitrarily two integers $R, R' \geq 2$. Let $\mathbf{z} := (z_1, \ldots, z_R)$ and $\mathbf{z}' := (z'_1, \ldots, z'_{R'})$ be two collections of elements of a space $\mathcal{Z}$. Let $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ map any couple (z, z') to a nonnegative number interpreted as the cost to move z to z', a cost function. The cost function c induces the $R \times R'$ matrix $C(\mathbf{z}, \mathbf{z}') \in \mathbb{R}_+^{R \times R'}$ whose (r, r')-specific component $(C(\mathbf{z}, \mathbf{z}'))_{r,r'} := c(z_r, z'_{r'})$ is interpreted as the cost to move z_r to $z'_{r'}$ (relative to c).

Let $\Pi_{R,R'} := \{P \in \mathbb{R}_+^{R \times R'} : P\,\mathbf{1}_{R'} = \frac{1}{R}\,\mathbf{1}_R, P^\top \mathbf{1}_R = \frac{1}{R'}\,\mathbf{1}_R\}$ represent the joint laws on $[\![R]\!] \times [\![R']\!]$ with uniform marginal laws, where $[\![d]\!] := \{1, \ldots, d\}$ for every integer $d \geq 1$. For each $P \in \Pi_{R,R'}$, let

$$E(P) := -\sum_{r \in [\![R]\!], r' \in [\![R']\!]} P_{r,r'} \log P_{r,r'}$$

denote the entropy of P. For every $P \in \Pi_{R,R'}$ and $C \in \mathbb{R}_+^{R \times R'}$, let

$$\langle P, C \rangle := \sum_{r \in [\![R]\!], r' \in [\![R']\!]} P_{r,r'} \times C_{r,r'}.$$

When $C = C(\mathbf{z}, \mathbf{z}')$, $\langle P, C \rangle$ is interpreted as the (P, C)-specific cost to transport $\mathbf{z}$ onto $\mathbf{z}'$.

For any $\gamma > 0$ and $C \in \mathbb{R}_+^{R \times R'}$, introduce

$$\mathcal{W}_\gamma(C) := \min_{P \in \Pi_{R,R'}} \left[\langle P, C \rangle - \gamma E(P)\right]. \tag{5.3}$$

In particular, when $C = C(\mathbf{z}, \mathbf{z}')$, $\mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}'))$ is the γ-regularized optimal cost to transport $\mathbf{z}$ onto $\mathbf{z}'$, abbreviated to "the γ-regularized OT cost". Considering the γ-regularized OT cost $\mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}'))$ instead of the regular OT cost $\mathcal{W}_0(C(\mathbf{z}, \mathbf{z}'))$ (defined as in (5.3) with $\gamma = 0$) has two important merits (Peyré and Cuturi, 2019, Chapters 3, 4, 9). First, $\mathbb{R}_+^{R \times R'} \ni C \mapsto \mathcal{W}_0(C) \in \mathbb{R}$ is not differentiable whereas $\mathbb{R}_+^{R \times R'} \ni C \mapsto \mathcal{W}_\gamma(C) \in \mathbb{R}$ is differentiable. Second, for any $C \in \mathbb{R}_+^{R \times R'}$, computing $\mathcal{W}_0(C)$ requires solving a costly linear program via network simplex methods whereas computing $\mathcal{W}_\gamma(C)$ can be performed easily thanks to the so-called Sinkhorn algorithm (Cuturi, 2013b).

Finally, we use the γ-regularized OT cost to define the γ-regularized Sinkhorn cost

$$\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}') := \mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}')) - \tfrac{1}{2}\left[\mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z})) + \mathcal{W}_\gamma(C(\mathbf{z}', \mathbf{z}'))\right]$$

(the dependence of $\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}')$ on the cost function c is hidden). By (Feydy et al., 2019b, Theorem 1), $\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}') \geq \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}) = 0$. Moreover, we stress that $\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}')$ can be computed with little additional computational cost compared to $\mathcal{W}_\gamma(\mathbf{z}, \mathbf{z}')$.

5.4 Making sparse predictions

The procedure we are about to present is funded on two core ideas. Firstly, we aim to predict whether a city will submit a request for the government declaration of natural disaster for a drought event by employing an interpretable comparison of the city's covariates with those of other cities whose submission status may be already known. Secondly, we want to have a control on the sparsity of the set of predictions and encourage 0-predictions, which correspond to cases where we predict that a city will not submit a request.

5.4.1 Translation to an optimization problem

As elaborated in Section 5.2.3, our objective is to predict $\zeta_{\alpha,3}$ based on $\xi_{\alpha,3,u}$ for every $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$, using past observations (5.1), and so repeatedly for each $u \in \mathcal{U}_3$. In the rest of the study, it will be convenient to denote generically $\{(x_m, y_m) : m \in [\![M]\!]\} \subset \mathcal{X} \times \{0,1\}$ and $\{(x'_n, y'_n) : n \in [\![N]\!]\} \subset \mathcal{X} \times \{0,1\}$ two collections of couples for which it is desired to predict y'_n based on x'_n, for every $n \in [\![N]\!]$, using past observations (x_1, y_1), ..., (x_M, y_M). To do so, we propose to solve the following optimization problem:

$$\underset{\theta \in \mathbb{R}^N}{\arg\min} \left\{ \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta)) + g_\tau(\theta) \right\}, \tag{5.4}$$

where

- for all $\theta \in \mathbb{R}^N$,

$$\mathbf{z} := ((x_1, y_1), \ldots, (x_M, y_M)), \quad \mathbf{z}'(\theta) := ((x'_1, \theta_1), \ldots, (x'_N, \theta_N));$$

- the cost function $c : (\mathcal{X} \times \mathbb{R}) \times (\mathcal{X} \times \mathbb{R}) \to \mathbb{R}_+$ is given by

$$c((x, y), (x', \theta)) := \mathrm{dis}(x, x')^2 + (y - \theta)^2 \tag{5.5}$$

  for a distance or dissimilarity dis on $\mathcal{X}$;

- g_τ is a convex function given by either $g_\tau(\theta) := \tau\|\theta\|_1 + \mathbf{I}\{\theta \in [0,1]^N\}$, with $\|\theta\|_1 := \sum_{n \in [\![N]\!]} |\theta_n|$, or $g_\tau(\theta) := \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$, where $\mathbf{I}\{A\}$ equals 0 if A is true and $+\infty$ otherwise;

- $\gamma, \tau > 0$ are some user-supplied constants.

A few comments are in order. Firstly, the argmin in (5.4) is over $\mathbb{R}^N$ but could equivalently be over $[0,1]^N$ (even if the term $\mathbf{I}\{\theta \in [0,1]^N\}$ was dropped from the definitions of $g_\tau(\theta)$). We thus view θ_n as the probability that the city described by x'_n will submit a request of the government declaration of natural disaster for a drought event.

Secondly, though hidden in the notation, the cost function c obviously plays a pivotal role. It operationalizes the core idea of making predictions based on comparisons between the covariates of different cities.

Thirdly, for both choices of g_τ, the ℓ^1-norm of θ can be seen as a measure of sparsity of θ, a substitute for the integer $\mathrm{card}\{n \in [\![N]\!] : \theta_n \neq 0\}$. Incorporating the penalization term $+g_\tau(\theta)$ operationalizes the core idea of promoting sparse solutions, aligning with our prior understanding that only a limited number of cities will eventually submit a request of the government declaration of natural disaster for a drought event (see Table 5.1 for the actual numbers and proportions of cities that did in 2019, 2020 and 2021). Finally, the case where $g_\tau(\theta) = \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$ is quite interesting because, as we will see, there is a natural way to select τ.

5.4.2 On solving (5.4)

Solving (5.4) is not straightforward, in part because the criterion to minimize is the sum of the non-convex differentiable function $f : \theta \mapsto \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta))$ (see Section 5.7.1.b) and of the convex non-differentiable function g_τ. Luckily, we can rely on the so-called iPiano algorithm (Ochs et al., 2015) which was developed precisely to deal with such optimization problems.

The iPiano algorithm starts from an initial $\theta^{-1} = \theta^0 \in]0,1[^N$ and the update scheme informally writes as (below, α, β are positive constants)

$$\theta^{k+1} = \mathrm{Prox}_{\alpha g_\tau}\left(\theta^k - \alpha \nabla f(\theta^k) + \beta(\theta^k - \theta^{k-1})\right), \tag{5.6}$$

where the proximal map $\mathrm{Prox}_{\alpha g_\tau}$ is defined by

$$\mathrm{Prox}_{\alpha g_\tau}(t) := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^N} \left\{ \tfrac{1}{2}\|\theta - t\|_2^2 + \alpha g_\tau(\theta) \right\}. \tag{5.7}$$

On the one hand, if $g_\tau(\theta) = \tau\|\theta\|_1 + \mathbf{I}\{\theta \in [0,1]^N\}$ then (5.7) is simply given by

$$\left(\mathrm{Prox}_{\alpha g_\tau}(t)\right)_n = \min\{(|t_n| - \alpha\tau)_+, 1\}.$$

In particular, if $t \in [0,1]^N$ then $(\mathrm{Prox}_{\alpha g_\tau}(t))_n = (t_n - \alpha\tau)_+$ for every $n \in [\![N]\!]$. On the other hand, if $g_\tau(\theta) = \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$ then the proximal map is the Euclidean projection onto the ℓ^1-ball centered at 0 and with radius τ. An efficient algorithm is available to implement this projection (Duchi et al., 2008).

Moreover, following (Cuturi and Doucet, 2014, Section 4.3), we show in Section 5.7.1.b that the gradient of f is given by

$$\begin{aligned}
\nabla f(\theta) &= \nabla \mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}'(\theta)) - \tfrac{1}{2}\nabla \mathcal{W}_\gamma\left(C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))\right) \\
&= 2(\tfrac{1}{N}\theta - \widehat{P}_\theta^\top y) - (\tfrac{2}{N}\theta - (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta) \\
&= -2\widehat{P}_\theta^\top y + (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta
\end{aligned} \tag{5.8}$$

with

$$\widehat{P}_\theta = \operatorname*{arg\,min}_{P \in \Pi_{M,N}} \left\{ \langle P, C(\mathbf{z}, \mathbf{z}'(\theta))\rangle - \gamma E(P) \right\}, \tag{5.9}$$

$$\widehat{Q}_\theta = \operatorname*{arg\,min}_{P \in \Pi_{N,N}} \left\{ \langle P, C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))\rangle - \gamma E(P) \right\}. \tag{5.10}$$

We check that the assumptions of (Ochs et al., 2015, Theorems 4.9 and 4.14) are met by proving that f is C^1-smooth with an L-Lipschitz gradient on $\mathrm{dom}\, g_\tau$ and that $(f + g_\tau)$ satisfies the Kurdyka-Lojasiewicz property on its domain (the proof is presented in Section 5.7). Therefore we can assert that

- the sequence $(\theta^k)_{k\geq 0}$ converges to a critical point of $\theta \mapsto f(\theta) + g_\tau(\theta)$;
- $\min_{k \leq K} \|\theta^{k+1} - \theta^k\|_2^2 = O(K^{-1})$;
- if we set $r(\theta) := \theta - \mathrm{Prox}_{\alpha g_\tau}(\theta - \alpha\nabla f(\theta))$, then $\min_{k \leq K} \|r(\theta^k)\|_2^2 = O(K^{-1})$.

The so-called proximal residual $r(\theta)$ is interesting because $r(\theta) = 0$ means that the first-order optimality condition is met at θ. Indeed (denoting by $\partial\ell(x)$ either the subdifferential of the convex function ℓ at x or the limiting-subdifferential of the proper lower semicontinuous function ℓ at x, see Section 5.7.2.a), $r(\theta) = 0$ iff

$$\begin{aligned}
\theta = \mathrm{Prox}_{\alpha g_\tau}(\theta - \theta\nabla f(\theta)) \quad &\text{iff} \quad 0 \in \partial\left(\tfrac{1}{2}\|\theta - \alpha\nabla f(\theta) - \cdot\|_2^2 + \alpha g_\tau\right)(\theta) \\
&\text{iff} \quad 0 \in \{\theta - (\theta - \alpha\nabla f(\theta))\} + \alpha\partial g_\tau(\theta) \\
&\text{iff} \quad 0 \in \{\alpha\nabla f(\theta)\} + \alpha\partial g_\tau(\theta) \\
&\text{iff} \quad 0 \in \partial(f + g_\tau)(\theta).
\end{aligned}$$

5.4.3 Implementation of the ``OT-procedure"

Algorithm 2 solves (5.4) by using the iPiano algorithm and a mini-batch procedure to cope with situations where M and N are large. From now on, running the OT-procedure will mean applying Algorithm 2.

Algorithm 2 A mini-batch version of the inertial proximal algorithm for nonconvex optimization (iPiano) tailored to solve (5.4). For any vector $\theta \in \mathbb{R}^N$ and subset $\mathcal{N}$ of $[\![N]\!]$, we denote $\theta|_{\mathcal{N}} := (\theta_n)_{n\in\mathcal{N}} \in \mathbb{R}^{\operatorname{card}\mathcal{N}}$.

Input: Data $\{(x_m, y_m) : m \in [\![M]\!]\}$, $\{x'_n : n \in [\![N]\!]\}$; regularization parameter $\gamma > 0$, constraint $\tau > 0$; learning rate $\alpha > 0$, momentum parameter $\beta \geq 0$; batch size $B \in \mathbb{N}^*$, number of iterations $T \in \mathbb{N}^*$

Output: Proposed optimizer θ^T

Sample $\theta^{-1} \in \mathbb{R}^N$ with independent components drawn from the uniform law on $[0, 0.01]$
Set $\theta^{-1} \leftarrow 0.5 + \theta^{-1}$ and $\theta^0 \leftarrow \theta^{-1}$
Set $t \leftarrow 0$
while $t < T$ **do**
  Independently, sample uniformly without replacement $\mathcal{M} \subset [\![M]\!]$, $\mathcal{N} \subset [\![N]\!]$ of cardinality B
  Set $\mathbf{z} \leftarrow ((x_m, y_m) : m \in \mathcal{M})$ and $\mathbf{z}'(\theta^t|_{\mathcal{N}}) \leftarrow ((x'_n, \theta^t_n) : n \in \mathcal{N})$
  Compute $F(\theta^t|_{\mathcal{N}}) = \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta^t|_{\mathcal{N}}))$ using Sinkhorn's algorithm
  Compute $\nabla F(\theta^t|_{\mathcal{N}})$ using automatic differentiation
  Set $\theta^{t+1} \leftarrow \theta^t$ and update $\theta^{t+1}|_{\mathcal{N}} \leftarrow \theta^{t+1}|_{\mathcal{N}} - \alpha\nabla F(\theta^t|_{\mathcal{N}}) + \beta(\theta^t|_{\mathcal{N}} - \theta^{t-1}|_{\mathcal{N}})$
  Update $\theta^{t+1} \leftarrow \operatorname{Prox}_{\alpha g_\tau}(\theta^{t+1})$
  Update $t \leftarrow t + 1$
end while

We wrote a `python/pytorch` program that implements Algorithm 2. It will be made available soon. The program hinges on the `GeomLoss` package (Feydy et al., 2019a) which provides a very fast GPU implementation of the Sinkhorn algorithm (Cuturi, 2013b).

In Section 5.5, we conduct a simple simulation study in a simple context where $\mathcal{X} = \mathbb{R}^2$ and both M and N are relatively small. We compare the results obtained by aggregating the predictions acquired from classification algorithms with those achieved through the OT-procedure. Notably, we report how we select the pivotal cost function (5.5), g_τ and the hyperparameters (γ, α, β) of Algorithm 2. Moreover, we also introduce the hybrid procedure which synergistically combines and utilizes the two types of predictions.

Section 5.6 is dedicated to the challenging task of forecasting the requests of the government declaration of natural disaster for a drought event. This real-world application poses greater challenges than the simulation study. Tangibly, these challenges arise because $\mathcal{X} \subset \mathbb{R}^d$ is a relatively high-dimensional space ($d = 67$) and both M and N are large. Intangibly, the intricacies lie in the mechanisms that determine whether a request is submitted or not.

We compare the results obtained from a classification algorithm with those achieved through the OT-procedure and the hybrid procedure. Regarding the OT-procedure, we notably rely on HYPERBAND (Li et al., 2018), a bandit-based approach to hyperparameter optimization, to define the pivotal cost function, and on a simple grid search to then fine-tune the hyperparameters (γ, α, β) of Algorithm 2.

5.5 A simple simulation study, introducing the "hybrid procedure"

5.5.1 Simulated data

For any $p \in (0,1)$, let P_p be the law on $\mathbb{R}^2 \times \{0,1\}$ such that

- if R and A are independently drawn from the $\chi^2(1)$ law and from the uniform law on $[0, 2\pi]$, if $X = (R\cos(A), R\sin(A))$ and if, conditionally on X, Y is drawn from the Bernoulli law with parameter $\mathrm{expit}(\mathrm{cst}(p) + R)$, then the joint law of (X, Y) is P_p;
- the above constant $\mathrm{cst}(p) \in \mathbb{R}$ is defined in such a way that $E_{P_p}(Y) = P_p(Y = 1) = p$.

For instance, $\mathrm{cst}(15\%) \approx -3.13$, $\mathrm{cst}(10\%) \approx -3.83$ and $\mathrm{cst}(5\%) \approx -5.00$. Note that, for any $p \in (0,1)$, under P_p, the further X is from 0 the more likely it is that $Y = 1$.

We generate independently $L = 30$ data sets as follows. For each $\ell \in [\![L]\!]$, for every $p \in \{15\%, 10\%, 5\%\}$, we independently sample $n = 1000$ independent copies of (X, Y) under P_p. We thus obtain $M = 3n$ couples $(x_{m,\ell}, y_{m,\ell})$. Moreover, we also sample independently $n = 1000$ independent copies of (X, Y) from the law P_p with $p = 5\%$. We thus obtain $N = n$ couples $(x'_{n,\ell}, y'_{n,\ell})$. Our objective is to recover, for each $\ell \in [\![L]\!]$, the vector $(y'_{n,\ell})_{n\in[\![N]\!]}$ based on $\{(x_{m,\ell}, y_{m,\ell}) : m \in [\![M]\!]\}$ and on $(x'_{n,\ell})_{n\in[\![N]\!]}$.

5.5.2 Fine-tuning the OT-procedure

Let us first describe how we fine-tune the OT-procedure in order to predict $(y'_{n,\ell})_{n\in[\![N]\!]}$ by solving (5.4) with $(x_m, y_m) = (x_{m,\ell}, y_{m,\ell})$ and $(x'_n, y'_n) = (x'_{n,\ell}, y'_{n,\ell})$ for all $m \in [\![M]\!]$ and $n \in [\![N]\!]$, for each $\ell \in [\![L]\!]$ in turn. On the one hand, we select the cost function $c : (\mathbb{R}^2 \times \{0,1\}) \times (\mathbb{R}^2 \times \{0,1\}) \to \mathbb{R}_+$ (5.5) given by

$$c((x_1, x_2, y), (x'_1, x'_2, y')) := 100 \times \left| \sqrt{x_1^2 + x_2^2} - \sqrt{(x'_1)^2 + (x'_2)^2} \right| + (y - y')^2.$$

Admittedly, this puts us in a favorable position because the true conditional probability of the event $Y = 1$ given X only depends on $\sqrt{X_1^2 + X_2^2}$. On the other hand, we choose the function $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$ for a τ whose choice is explained in Section 5.5.3. Furthermore, in view of Algorithm 2, we set $\gamma = 10^{-3}$, $\alpha = 10^{-3}$, $\beta = 10^{-4}$, $B = 128$ and $T = 2000$.

5.5.3 Alternative, classification-based approaches

As an alternative approach, we also consider training an algorithm using $\{(x_{m,\ell}, y_{m,\ell}) : m \in [\![M]\!]\}$ in order to learn to classify each $x'_{n,\ell}$ individually ($n \in [\![N]\!]$), for every $\ell \in [\![L]\!]$ in turn. Instead of selecting one algorithm, we rely on super learning to learn and train a meta-algorithm that builds upon several algorithms to classify at least as well as (and sometimes better than) all the candidate algorithms (van der Laan et al., 2007; Polley et al., 2021, 2011, and references therein). We rely on four individual algorithms to learn the conditional probability of the event $Y = 1$ given X: an algorithm that approximates it under the form of a constant function (in X); an algorithm that learns which element of the working model $\{x \mapsto \mathrm{expit}(t_0 + t_1 x_1 + t_2 x_2) : t \in \mathbb{R}^3\}$ best approximates it (see `stats::glm`); an algorithm that approximates it under the form of a tree, using the covariates X_1 and X_2 (see `rpart::rpart`); an algorithm that approximates it under the form of a random forest, using the covariates X_1 and X_2 (see `ranger::ranger`) – more details are given below.

In addition, we consider a second super learning procedure to learn the conditional probability of the event $Y = 1$ given X by relying on: an algorithm that approximates it under the form of a constant function (in X); an algorithm that learns which element of the working model $\{x \mapsto \text{expit}(t_0 + t_1x_1 + t_2x_2 + t_3\sqrt{x_1^2 + x_2^2}) : t \in \mathbb{R}^4\}$ best approximates it (see `stats::glm`); an algorithm that approximates it under the form of a tree, using the covariates X_1, X_2 and $\sqrt{X_1^2 + X_2^2} = R$ (see `rpart::rpart`); an algorithm that approximates it under the form of a random forest, using the covariates X_1, X_2 and R (see `ranger::ranger`). We expect the second super learner to perform better than the first one because it can use the relevant covariate R.

We use the `SuperLearner R` package (R Core Team, 2022; Polley et al., 2021) to implement and train the super learners. For both super learning procedures, we rely on V-fold cross validation with $V = 10$ folds and use the default hyperparameters specified in `SuperLearner::SL.glm`, `SuperLearner::SL.rpart` (Therneau and Atkinson, 2019) and `SuperLearner::SL.ranger` (Wright and Ziegler, 2017).

5.5.4 Results, introducing the "hybrid procedure"

For each $\ell \in [\![L]\!]$, we train the two super learners and denote by $\widehat{y}'^{\text{SL}_1}_{n,\ell}$ and $\widehat{y}'^{\text{SL}_2}_{n,\ell}$ the estimates of the conditional probabilities that $Y = 1$ given $X = x'_{n,\ell}$ that they output for each $n \in [\![N]\!]$. Next, we set $\tau = \|\widehat{y}'^{\text{SL}_2}_{n,\ell}\|_1$ for the OT-procedure, run it, and denote by $\widehat{y}'^{\text{OT}}_{n,\ell}$ the estimates of the conditional probability that $Y = 1$ given $X = x'_{n,\ell}$ for each $n \in [\![N]\!]$ that it yields.

Before discussing the results, we introduce a fourth procedure that we aptly refer to as the "hybrid procedure" because it builds upon the OT-procedure and the second super learning procedure. Specifically, the hybrid procedure produces estimates of the above conditional probabilities which are merely defined as the geometric means of the estimates output by the second super learner and yielded by the OT-procedure. Hereafter, these estimates are denoted by $\widehat{y}'^{\text{HYB}}_{n,\ell} := (\widehat{y}'^{\text{SL}_2}_{n,\ell} \times \widehat{y}'^{\text{OT}}_{n,\ell})^{1/2}$ for every $n \in [\![N]\!]$.

Figure 5.1 provides insights into the predictions $\{\widehat{y}'^{\bullet}_{n,\ell} : n \in [\![N]\!]\}$ where the symbol $\bullet$ stands for $\text{SL}_1, \text{SL}_2, \text{OT}, \text{HYB}$. On the one hand, the empirical cumulative distribution functions (ecdfs) plotted in the left-hand side panel of Figure 5.1 reveal that the predictions $\widehat{y}'^{\text{OT}}_{n,\ell}$ for $(n, \ell) \in [\![N]\!] \times [\![L]\!]$ such that $y_{n,\ell} = 0$ are often (17%) equal to 0 and are generally more concentrated around 0 than the other predictions (the red ecdf dominates the others). In stark contrast, the predictions $\widehat{y}'^{\text{SL}_1}_{n,\ell}$ and $\widehat{y}'^{\text{SL}_2}_{n,\ell}$ for the same couples (n, ℓ) are bounded away from 0 (being larger than 1.56% and 1.35%, respectively). On the other hand, the ecdfs plotted in the right-hand side panel of Figure 5.1 reveal that the predictions $\widehat{y}'^{\text{OT}}_{n,\ell}$ for $(n, \ell) \in [\![N]\!] \times [\![L]\!]$ such that $y_{n,\ell} = 1$ can be equal to 0 (2.7%) and are generally smaller than the other predictions (the red ecdf dominates the others again). They also show that the second super learner outperforms the first one in the sense that the maximum gap between their ecdfs is large (a Kolmogorov-Smirnov viewpoint). Furthermore, by conducting a comparison across panels we discern the notable and desirable trend wherein the predictions $\{\widehat{y}'^{\bullet}_{n,\ell} : n \in [\![N]\!], \ell \in [\![L]\!] \text{ st } y'_{n,\ell} = y\}$ exhibit larger values when $y = 1$ as opposed to when $y = 0$. In conclusion, the hybrid predictions seem to strike a fine balance between the predictions output by the second super learner and the OT-procedure.

In order to complement this first analysis, we employ mean squared error (MSE) as a measure of performance and compute, for each $\ell \in [\![L]\!]$,

$$\text{MSE}^{\bullet}_{\ell} := \frac{1}{N} \sum_{n \in [\![N]\!]} (y'_{n,\ell} - \widehat{y}'^{\bullet}_{n,\ell})^2 \tag{5.11}$$

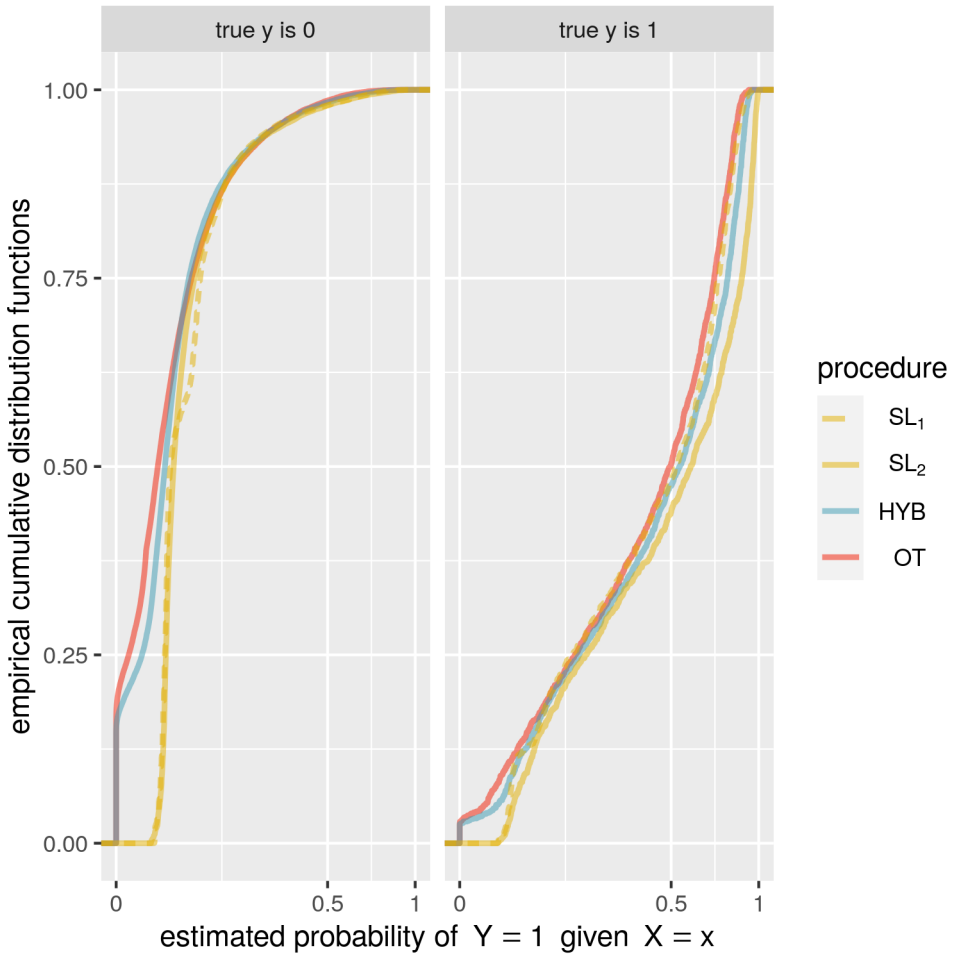

Figure 5.1 – Empirical cumulative distribution functions of the sets $\{\widehat{y}'^{\bullet}_{n,\ell} : \ell \in [\![L]\!], n \in [\![N]\!] \text{ st } y'_{n,\ell} = y\}$ for $y = 0$ (left-hand side panel) and $y = 1$ (right-hand side panel), where the symbol $\bullet$ stands for $\mathsf{SL}_1, \mathsf{SL}_2, \mathsf{OT}, \mathsf{HYB}$.

where we substitute $\mathrm{SL}_1, \mathrm{SL}_2, \mathrm{OT}, \mathrm{HYB}$ for the symbol $\bullet$. The average and standard deviations of these numbers are reported in Table 5.2. There is no stark differences in terms of standard deviations. In terms of average, the estimates yielded by the OT-procedure outperform those obtained by super learning. However, it is the hybrid procedure that emerges as the top performer. Figure 5.2 allows us to go beyond comparisons in average. More that two thirds of the points are situated to the left of the black vertical line, meaning that $\mathrm{MSE}_\ell^{\mathrm{OT}}$ is smaller than $\mathrm{MSE}_\ell^{\mathrm{SL}_2}$ for the corresponding ℓs. Likewise, 29 out of 30 blue points are situated below the horizontal black line, meaning that $\mathrm{MSE}_\ell^{\mathrm{HYB}}$ is smaller than $\mathrm{MSE}_\ell^{\mathrm{SL}_2}$ for the corresponding ℓs, while 24 out of 30 red points are situated below the horizontal black line, meaning that $\mathrm{MSE}_\ell^{\mathrm{HYB}}$ is smaller than $\mathrm{MSE}_\ell^{\mathrm{OT}}$ for the corresponding ℓs. In particular, the average pattern unveiled by Table 5.2 remains consistent even before averaging: the hybrid procedure exhibits superior performance, surpassing the OT-procedure, which in turn outperforms the second super learning procedure.

	MSE	
procedure	average	std. deviation
SL_1	0.0361	0.0046
SL_2	0.0345	0.0048
HYB	**0.0330**	0.0045
OT	0.0337	**0.0044**

Table 5.2 – Averages and standard deviations of the mean squared errors $\{\mathrm{MSE}_\ell^\bullet : \ell \in [\![L]\!]\}$ (5.11) where the symbol $\bullet$ stands for $\mathrm{SL}_1, \mathrm{SL}_2, \mathrm{OT}, \mathrm{HYB}$ and $L = 30$. See also Figure 5.2. In each column, the smallest value stands out in bold characters.

5.6 Forecasting the requests of the government declaration of natural disaster for a drought event in France

5.6.1 Fine-tuning the OT-procedure

DEFINING A COST FUNCTION. To begin with, we address the challenge of defining a cost function $c : (\mathcal{X} \times \mathbb{R}) \times (\mathcal{X} \times \mathbb{R}) \to \mathbb{R}_+$ (5.5). In view of the description of a generic vector of covariates $x \in \mathcal{X}$ made in Section 5.2.1, let us rewrite $x := (x_{[1]}, \ldots, x_{[4]})$ where $x_{[1]}$, $x_{[2]}$, $x_{[3]}$ and $x_{[4]}$ respectively regroup the covariates that collectively describe the corresponding city ($x_{[1]}$, 16 variables) and its exposure to drought events ($x_{[2]}$, 25 variables), provide a history of its past requests of declaration of natural disaster for a drought event, successful or not ($x_{[3]}$, 13 variables), and describe the city's vicinity ($x_{[4]}$, 13 variables).

Let $\bar{\xi}_1$ and std_1 be the vectors whose components are the component-specific mean and standard deviation of

$$\{\xi_{\alpha,1,u} : \alpha \in \mathcal{A}_1, u \in \mathcal{U}_1 \text{ st } \zeta_{\alpha,t,u} = 0 \text{ or } (\zeta_{\alpha,t,u^-}, \zeta_{\alpha,t,u}) = (0,1)\} \subset \mathcal{X},$$

that is, the set of covariates corresponding to year 2019, and let $\bar{\zeta}_1$ be the $\|\cdot\|_1$-norm of $\{\zeta_{\alpha,1} : \alpha \in \mathcal{A}_1\}$, that is, the number of cities which made a request for year 2019. For any generic vector of covariates $x \in \mathcal{X}$, denote (using the entrywise division of vectors)

$$\widetilde{x} := \frac{x - \bar{\xi}_1}{\mathrm{std}_1}. \tag{5.12}$$

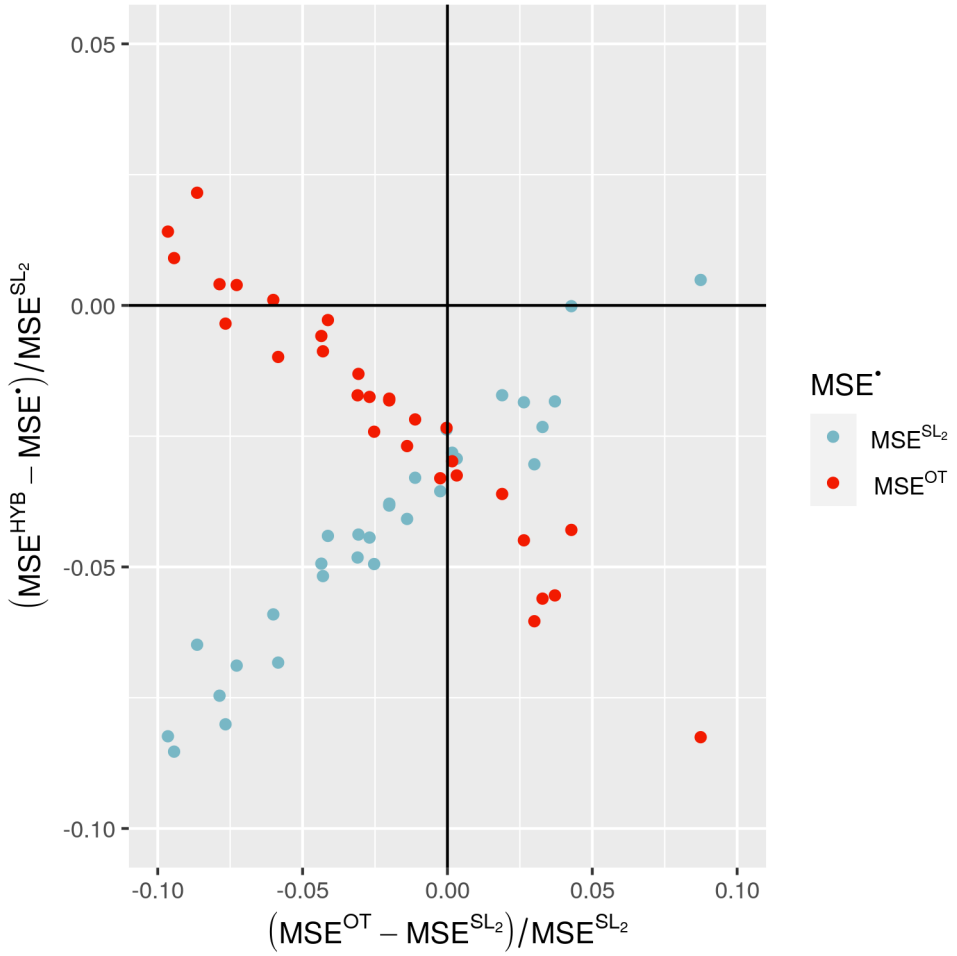

Figure 5.2 – Scatterplot of $(\mathrm{MSE}_\ell^{\mathrm{HYB}} - \mathrm{MSE}_\ell^{\bullet})/\mathrm{MSE}_\ell^{\mathrm{SL_2}}$ against $(\mathrm{MSE}_\ell^{\mathrm{OT}} - \mathrm{MSE}_\ell^{\mathrm{SL_2}})/\mathrm{MSE}_\ell^{\mathrm{SL_2}}$ ($\ell \in [\![30]\!]$) where the symbol $\bullet$ stands for $\mathrm{SL_2}$ (blue) or OT (red). See also Table 5.2.

We select a cost function in the parametric set $\{c_a : a \in \mathbb{R}^5_+\}$ where, for any $a \in \mathbb{R}^5_+$ and $x, x' \in \mathcal{X}$, $y, y' \in \mathbb{R}$,

$$c_a((x,y),(x',y')) := \sum_{k=1}^{4} a_k \|\widetilde{x}_{[k]} - \widetilde{x}'_{[k]}\|_2^2 + a_5(y-y')^2. \tag{5.13}$$

To do so, we rely on HYPERBAND, an algorithm which reformulates hyperparameter optimization as a pure-exploration, adaptive resource allocation problem addressing how to allocate resources among randomly generated hyperparameter configurations (Li et al., 2018). Specifically, in view of (5.4), we set $\gamma = 10^{-2}$, $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$ with $\tau = \bar{\zeta}_1$ and, in view of (5.6) and Algorithm 2 in Section 5.4.3, we set

$$\{(x_m, y_m) : m \in [\![M]\!]\} = \{(\xi_{\alpha,1,75}, \zeta_{\alpha,1}) : \alpha \in \mathcal{A}_1 \text{ st } \zeta_{\alpha,2,75} = 0 \text{ or } (\zeta_{\alpha,2,75^-}, \zeta_{\alpha,2,75}) = (0,1)\}, \tag{5.14}$$

$$\{x'_n : n \in [\![N]\!]\} = \{\xi_{\alpha,2,85} : \alpha \in \mathcal{A}_2 \text{ st } \zeta_{\alpha,2,85} = 0\}, \tag{5.15}$$

$\alpha = 10^{-3}$, $\beta = 10^{-4}$ and $B = 128$. In words, setting (5.14) and (5.15) means that we exploit the data associated with the last week relative to year 2019 (that is, the $(75 - 52) = 23$rd week of 2020) to predict which cities will submit a request for the government declaration of natural disaster for a drought event for year 2020 during the last week relative to year 2020 (that is, the $(85 - 52) = 33$rd week of 2021). As for the random generation of configurations $a = (a_1, a_2, a_3, a_4, a_5) \in \mathbb{R}^5_+$, we sample independently a_5 uniformly on $[1/5, 10]$ and (a_1, a_2, a_3, a_4) from the law of $73 \times \exp(Z)/\|\exp(Z)\|_1$ with Z drawn in $\mathbb{R}^4$ from the centered Gaussian law with identity covariance matrix and where the exponential is applied elementwise.

Moreover, in view of (Li et al., 2018, Algorithm 1, page 8), we set the maximum amount of resource that can be allocated to a single configuration (that is, the maximum number of iterations in Algorithm 2 that can be allocated to a randomly generated candidate $a \in \mathbb{R}^5_+$) to $R = 3000$ and the parameter controlling the proportion of configurations discarded in each round of SUCCESSIVEHALVING to $\eta = 10$. For this specific couple (R, η), HYPERBAND consists of 4 independent "brackets" which we present in Table 5.3. In the bracket indexed by $s = 0$, $n_{0,0} = 4$ different $a \in \mathbb{R}^5_+$s (that is, configurations) are independently randomly generated; then each is allocated $r_{0,0} = 3000$ iterations in Algorithm 2 and associated with a score, a notion that we will clarify in the next paragraph. In the brackets indexed by $s \in \{1,2,3\}$, $n_{s,0}$ different $a \in \mathbb{R}^5_+$s are independently randomly generated; then, each is allocated $r_{s,0}$ iterations of Algorithm 2 and associated with a score. Next, recursively for $i = 1, \ldots, s$, each of the $n_{s,i}$ configurations with the smallest scores is allocated $r_{s,i}$ iterations of Algorithm 2 and associated with a new score.

	brackets							
	$s=3$		$s=2$		$s=1$		$s=0$	
i	$n_{3,i}$	$r_{3,i}$	$n_{2,i}$	$r_{2,i}$	$n_{1,i}$	$r_{1,i}$	$n_{0,i}$	$r_{0,i}$
0	1000	3	134	30	20	300	4	3000
1	100	30	13	300	2	3000		
2	10	300	1	3000				
3	4	3000						

Table 5.3 – Resource allocations and numbers of configurations $((r_{s,i}, n_{s,i}),\ i \in \{0, \ldots, s\})$ in each bracket $s \in \{0,1,2,3\}$ of the HYPERBAND procedure.

It only remains to clarify what are the aforementioned scores. For any configuration a randomly generated and tested while running HYPERBAND, let us denote by $\widehat{\zeta}_{\alpha,2}^{\mathrm{OT},85}(a)$ the predicted probability output by Algorithm 2 that city α will eventually submit a request for the government declaration of natural disaster for a drought event for year 2020 for every $\alpha \in \mathcal{A}_2$ such that $\zeta_{\alpha,2,85} = 0$. The score associated with a is the MSE score

$$\frac{1}{N} \sum_{\alpha \in \mathcal{A}_2} (\widehat{\zeta}_{\alpha,2}^{\mathrm{OT},85}(a) - \zeta_{\alpha,2})^2 \, \mathbf{1}\{\zeta_{\alpha,2,85} = 0\}. \tag{5.16}$$

This completes the description of the HYPERBAND algorithm that we run to select a cost function of the form (5.13). Eventually, we select c_a with $a \approx (16.75, 18.74, 30.57, 6.94, 0.34)$ (entries rounded to two decimal places).

RELATIVE IMPORTANCE OF THE FOUR GROUPS OF COVARIATES CONCERNING THE SELECTED COST FUNCTION. To discuss the relative importance of each term in (5.13) with this choice of a, we sample uniformly without replacement $M = B = 128$ elements $x_1, \ldots, x_m, \ldots, x_M$ from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\} \subset \mathcal{X}$ and, independently, $N = B = 128$ elements $x'_1, \ldots, x'_n, \ldots, x'_N$ from $\{\xi_{\alpha,2,48} : \alpha \in \mathcal{A}\}$ (recall that $\min \mathcal{U}_1 = 44$ and $\min \mathcal{U}_2 = 48$). In view of (5.12), each x_m yields $\widetilde{x}_{m,[1]}, \widetilde{x}_{m,[2]}, \widetilde{x}_{m,[3]}, \widetilde{x}_{m,[4]}$ and each x'_n yields $\widetilde{x}'_{n,[1]}, \widetilde{x}'_{n,[2]}, \widetilde{x}'_{n,[3]}, \widetilde{x}'_{n,[4]}$. We then compute the quartiles of the sets $\{\|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 : m \in [\![M]\!], n \in [\![N]\!]\}$ $(k = 1, 2, 3, 4)$, which we report in Table 5.4.

Looking at Table 5.4 it seems that, for any $x, x' \in \mathcal{X}$ viewed as two cities' vectors of covariates, the sum $\sum_{k=1}^4 a_k \|x_{[k]} - x'_{[k]}\|_2^2$ (the left-hand side sum in (5.13)) is mainly driven, in decreasing order of importance, by $x_{[2]}, x'_{[2]}$ (the groups of 25 covariates describing the cities' exposures to drought events), $x_{[3]}, x'_{[3]}$ (the groups of 13 covariates describing the cities' histories of requests of declaration of natural disaster for a drought event), $x_{[1]}, x'_{[1]}$ (the groups of 16 covariates describing the cities) and $x_{[4]}, x'_{[4]}$ (the groups of 13 covariates describing the cities' vicinities). This is confirmed by Figure 5.3.

Figure 5.3 represents the cumulative distribution functions of the sets $\{\mathrm{cst}_{m,n} \times a_k \|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 : m, n \in [\![128]\!]\}$ $(k = 1, 2, 3, 4)$ where each $\mathrm{cst}_{m,n}$ (any $m, n \in [\![128]\!]$) is defined as

$$\mathrm{cst}_{m,n} := \left(\sum_{k=1}^4 a_k \|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 \right)^{-1}.$$

The more a cumulative distribution function is shifted to the right the more a generic sum $\sum_{k=1}^4 a_k \|x_{[k]} - x'_{[k]}\|_2^2$ (for any $x, x' \in \mathcal{X}$, the left-hand side sum in (5.13)) is driven by the corresponding groups of covariates. By this criterion, we recover the ordering suggested by Table 5.4.

SETTING THE REMAINING HYPERPARAMETERS. Once the cost function is defined, we carry out a grid search to select values for γ (the regularization parameter in (5.4)), α and β (the learning rate and momentum parameters in Algorithm 2), with

$$(\gamma, \alpha, \beta) \in \{10^{-2}, 10^{-1}, 1\} \times \{10^{-3}, 5 \times 10^{-3}\} \times \{10^{-4}, 5 \times 10^{-4}\}.$$

For each possible triplet (γ, α, β), we run Algorithm 2 with $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$ where $\tau = \bar{\zeta}_1$, (5.14), (5.15), $B = 128$ and collect the predicted probability $\widehat{\zeta}_{\alpha,2}^{\mathrm{OT},85}(\gamma, \alpha, \beta)$ that city α will eventually submit a request for the government declaration of natural disaster for a drought event for year 2020 for every $\alpha \in \mathcal{A}_2$ such that

covariates describing: $(\widetilde{x}_{[k]})$	a city $(k=1)$	its exposure to drought events $(k=2)$	its request history $(k=3)$	its vicinity $(k=4)$
minimum	0.40	2.80	2.01	0.00
1st quartile	5.25	7.41	2.01	1.26
median	6.20	8.75	3.69	2.35
3rd quartile	7.25	10.22	6.20	3.83
maximum	15.94	20.78	15.80	20.18
a	16.75	18.74	30.57	6.94

Table 5.4 – Quartiles of the sets $\{\|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 : m, n \in [\![128]\!]\}$ $(k = 1, 2, 3, 4)$ where $\widetilde{x}_1, \ldots, \widetilde{x}_{128}$ and $\widetilde{x}'_1, \ldots, \widetilde{x}'_{128}$ are derived from $x_1, \ldots, x_{128}$ and $x'_1, \ldots, x'_{128}$ which are independently sampled, uniformly without replacement, from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\}$ and $\{\xi_{\alpha,2,48} : \alpha \in \mathcal{A}\}$. The last row recalls the four first entries of a selected based on the HYPERBAND algorithm. See also Figure 5.3.

$\zeta_{\alpha,2,85} = 0$. The score associated with (γ, α, β) is the MSE score defined as in (5.16) with $\widehat{\zeta}^{\mathrm{OT},85}_{\alpha,2}(\gamma, \alpha, \beta)$ substituted for $\widehat{\zeta}^{\mathrm{OT},85}_{\alpha,2}(a)$. We select the triplet whose score is the smallest: $(\gamma, \alpha, \beta) = (10^{-2}, 10^{-3}, 10^{-4})$.

5.6.2 Alternative, classification-based approaches

As in the simulation study presented in Section 5.5, we also develop an alternative approach to predicting the requests of the government declaration of natural disaster for a drought event. We consider five individual algorithms in order to learn to classify each x'_n $(n \in [\![N]\!])$ using $\{(x_m, y_m) : m \in [\![M]\!]\}$. From a probabilistic viewpoint, the first algorithm, CST, approximates the conditional probability that $Y = 1$ given X under the form of a constant function (in X); the second algorithm, GLM, learns which element in a linear working model best approximates it (see `stats::glm`); the third algorithm, RPART, approximates it under the form of a tree (see `rpart::rpart`); the fourth algorithm, RANGER, approximates it under the form of a random forest (see `ranger::ranger`); the fifth algorithm, KNN, uses the nearest labelled neighbors of any x to estimate the conditional probability at $X = x$. More specifically, the linear working model at the core of GLM regresses Y linearly onto each component of X, treating as categorical variables the covariates characterizing a city's seismic and climatic zones, and uses a logit link function. RPART relies on the default hyperparameters specified in `rpart::rpart.control` (Therneau and Atkinson, 2019). RANGER uses the Gini splitting rule while the other hyperparameters are set to their default values specified in `ranger::ranger` (Wright and Ziegler, 2017). As for KNN, it relies on the `python` class `sklearn.neighbors.KNeighborsClassifier` (Buitinck et al., 2013) and uses $k = 100$ neighbors, uniform weights, the ball tree algorithm (Liu et al., 2006, to handle the large learning data set) with a leaf size set to 30 and the weighted Euclidean $(x, x') \mapsto \|\tilde{x} - \tilde{x}'\|_2$.

We adopt a sequential learning viewpoint. Firstly, we train the five algorithms using all the data relative to year 2019, that is

$$\begin{aligned}&\{(x_m, y_m) : m \in [\![M]\!]\}\\ &= \{(\xi_{\alpha,1,u}, \zeta_{\alpha,1}) : \alpha \in \mathcal{A}_1, u \in \mathcal{U}_1 \text{ st } \zeta_{\alpha,1,u} = 0 \text{ or } (\zeta_{\alpha,1,u^-}, \zeta_{\alpha,1,u}) = (0,1)\},\end{aligned}$$

yielding five functions $\widehat{\zeta}^{\bullet}_1 : \mathcal{X} \to [0,1]$, where the symbol $\bullet$ stands for CST, GLM, RPART, RANGER or KNN. Secondly, for each algorithm in turn, we compute the predicted probabilities of submitting a request relative to year 2020 for every week $u \in \mathcal{U}_2$ and all cities

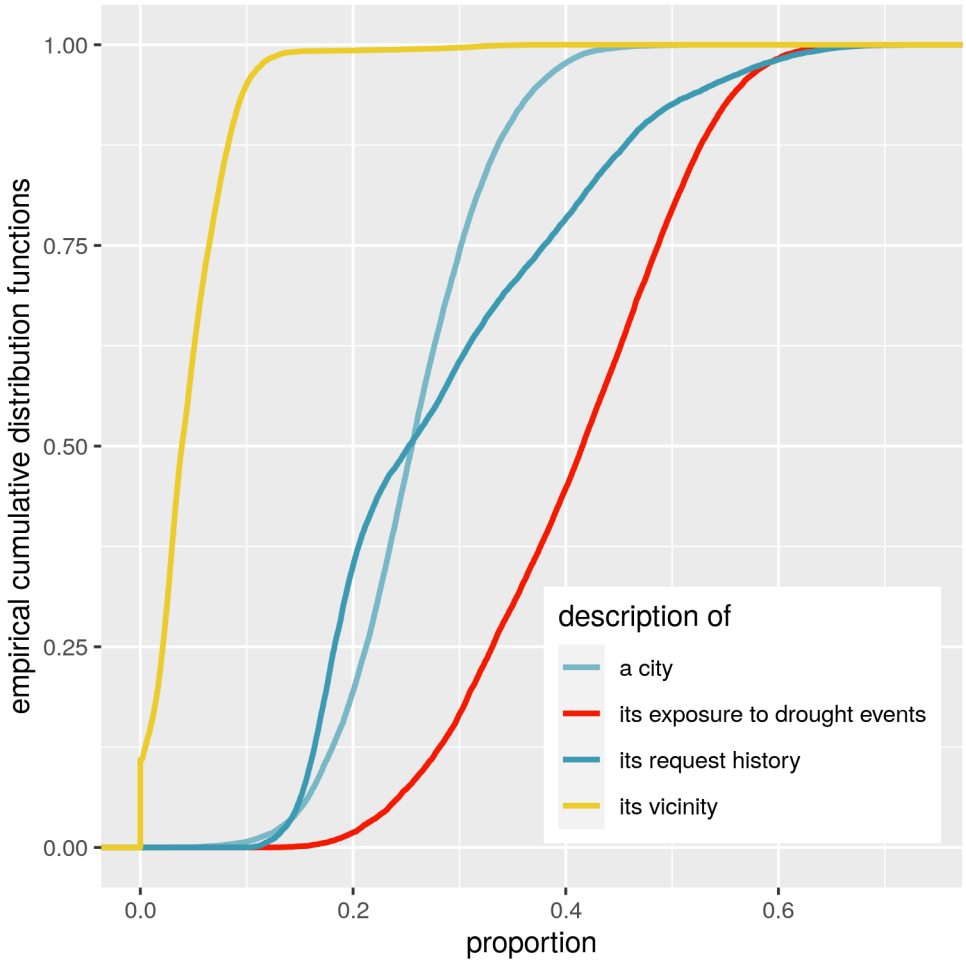

Figure 5.3 – Cumulative distribution functions of the sets $\{\mathsf{cst}_{m,n} \times a_k \|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 : m, n \in [\![128]\!]\}$ $(k = 1, 2, 3, 4)$ where $\widetilde{x}_1, \ldots, \widetilde{x}_{128}$ and $\widetilde{x}'_1, \ldots, \widetilde{x}'_{128}$ are derived from $x_1, \ldots, x_{128}$ and $x'_1, \ldots, x'_{128}$ which are independently sampled, uniformly without replacement, from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\}$ and $\{\xi_{\alpha,2,48} : \alpha \in \mathcal{A}\}$, where a is selected based on the HYPERBAND algorithm, and where each $\mathsf{cst}_{m,n}$ is such that $\mathsf{cst}_{m,n} \times \sum_{k=1}^4 a_k \|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 = 1$ for all $m, n \in [\![128]\!]$. The more a cumulative distribution function is shifted to the right the more a generic sum $\sum_{k=1}^4 a_k \|x_{[k]} - x'_{[k]}\|_2^2$ (for any $x, x' \in \mathcal{X}$, the left-hand side sum in (5.13)) is driven by the corresponding groups of covariates. See also Table 5.4.

which did not submit a request yet by week u, that is $\widehat{\zeta}^{\bullet,u}_{\alpha,2} := \widehat{\zeta}^{\bullet}_1(\xi_{\alpha,2,u})$ for every $u \in \mathcal{U}_2$ and $\alpha \in \mathcal{A}_2$ such that $\zeta_{\alpha,2,u} = 0$. Thirdly, for each algorithm in turn, we compute the overall MSE score

$$\frac{\sum_{u \in \mathcal{U}_2} \sum_{\alpha \in \mathcal{A}_2} (\widehat{\zeta}^{\bullet,u}_{\alpha,2} - \zeta_{\alpha,2})^2 \, \mathbf{1}\{\zeta_{\alpha,2,u} = 0\}}{\sum_{u \in \mathcal{U}_2} \sum_{\alpha \in \mathcal{A}_2} \mathbf{1}\{\zeta_{\alpha,2,u} = 0\}}.$$

The top-performing algorithm, GLM, is defined as the one with the smallest overall MSE score among all. We refer to it as the *discrete* super learner SL for year 2021 (we comment on the word "discrete" in the next paragraph). Lastly we retrain GLM, leveraging all data

relative to years 2019 and 2020, that is

$$\{(x_m, y_m) : m \in [\![M]\!]\}$$
$$= \{(\xi_{\alpha,t,u}, \zeta_{\alpha,t}) : t = 1, 2, \alpha \in \mathcal{A}_t, u \in \mathcal{U}_t \text{ st } \zeta_{\alpha,t,u} = 0 \text{ or } (\zeta_{\alpha,t,u^-}, \zeta_{\alpha,t,u}) = (0, 1)\},$$

yielding the function $\widehat{\zeta}^{\mathrm{SL}}_{1:2} : \mathcal{X} \to [0, 1]$.

Returning to the word "discrete" mentioned in the previous paragraph, it suggests that our focus lies in determining the top-performing algorithm rather than seeking the best combination of all the algorithms. This approach is justified due to our limited hindsight, relying solely on two years of data. To illustrate, consider a future scenario where we aim to forecast the requests of the government declaration of natural disaster for a drought event for year t beyond 2021 based on data from years 2019 to $(t-1)$. The sequential learning procedure outlined above would naturally extend, opening the possibility that another algorithm may outperform GLM as the best-performing algorithm.

5.6.3 Results

We compute the predicted probabilities of submitting a request relative to year 2021 for every week $u \in \mathcal{U}_3$ and all cities which did not submit a request yet by week u, that is $\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3} := \widehat{\zeta}^{\mathrm{SL}}_{1:2}(\xi_{\alpha,3,u})$ for every $u \in \mathcal{U}_3$ and $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Moreover, we run Algorithm 2 sequentially for each $u \in \mathcal{U}_3$, using the cost function (5.13) with $a \approx (16.75, 18.74, 30.57, 6.94, 0.34)$, $(\gamma, \alpha, \beta) = (10^{-2}, 10^{-3}, 10^{-4})$, $B = 128$, $T = 30,000$ and $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$ with $\tau = \|(\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3})_{\alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1$. This yields the predictions $\widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3}$ for every $u \in \mathcal{U}_3$ and $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Finally, we compute the predictions according to the hybrid procedure, that is, $\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3} := (\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3} \times \widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3})^{1/2}$ for every $u \in \mathcal{U}_3$ and $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Of note, it necessarily holds by design that

$$\|(\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3})_{\alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1 \leq \|(\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3})_{\alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1 \tag{5.17}$$

for every $u \in \mathcal{U}_3$. Indeed, for any $\theta, \theta' \in \mathbb{R}^N_+$ such that $\|\theta\|_1 \geq \|\theta'\|_1$, the Cauchy-Schwarz inequality yields

$$\|([\theta_n \theta'_n]^{1/2})_{n \in [\![N]\!]}\|_1 \leq (\|\theta\|_1 \times \|\theta'\|_1)^{1/2} \leq \|\theta\|_1.$$

Figure 5.4 represents the ecdfs of the predicted probabilities $\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ of submitting a request for the government declaration of natural disaster for a drought event for year 2021 output by the super learner, the OT-procedure and the hybrid procedure for a selection of weeks u: the 49th week of 2021 (December 6th to 12th, $u = \min \mathcal{U}_3 = 49$), the 7th, 17th and 26th weeks of 2022 (February 15th to 21st, $u = 59$; April 26th to May 2nd, $u = 69$; June 28th to July 4th, $u = \max \mathcal{U}_3 = 78$). For each week, the right-hand side and left-hand side panels respectively focus on cities that will and that will not submit a request eventually. As expected, the curves in the left-hand side panels dominate their counterparts in the right-hand side panels, illustrating the fact that the predicted probabilities are smaller (in law) for cities that will not submit a request eventually than for cities that will. The curves mainly differ around the origin. The left-hand side panels clearly showcase the ability of the OT-procedure to rightly assign a 0 probability to submit a request to cities that, indeed, will not submit one eventually: this concerns 49.5%, 51.2%, 50.7% and 56.4% of them for weeks 49, 59, 69 and 78 respectively. In contrast, the quantiles of order 49.5%, 51.2%, 50.7% and 56.4% of the super learner's predictions for these cities are 1.5%, 1.3%, 0.8% and 0.5% respectively. This notable ability comes at a price, as illustrated by the right-hand side panels showing that a 0-probability to submit a request is wrongly assigned to a

fraction of the cities that, in fact, will submit one eventually: this concerns 4.3%, 7.6%, 6.7% and 14.6% of them for weeks 49, 59, 69 and 78 respectively. In comparison, the quantiles of order 4.3%, 7.6%, 6.7% and 14.6% of the super learner's predictions for these cities are 1.7%, 1.9%, 0.9% and 0.9% respectively.

Figure 5.5 compares the predicted probabilities of submitting a request for the government declaration of natural disaster for a drought event for year 2021 output by the super learner and by the OT-procedure during the 49th week of 2021 ($u = \min \mathcal{U}_3 = 49$) and the 26th week of 2022 ($u = \max \mathcal{U}_3 = 78$). For each week, the right-hand side and left-hand side panels respectively focus on cities that will and that will not submit a request eventually. Points lying above the first bisecting line correspond to cities $\alpha \in \mathcal{A}_3$ for which $\widehat{\zeta}_{\alpha,3}^{\mathrm{OT},u} > \widehat{\zeta}_{\alpha,3}^{\mathrm{SL},u}$. Colored points represent quantiles of order 10%, 50% and 90%. Two patterns emerge. On the one hand, for $u = 49$ and $u = 78$ both, when concentrating on cities that will not submit a request eventually: *(a)* the 10%-quantile and median of $\{\widehat{\zeta}_{\alpha,3}^{\mathrm{OT},u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ are smaller than those of $\{\widehat{\zeta}_{\alpha,3}^{\mathrm{SL},u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ while *(b)* the 90%-quantile of the former set is larger than that of the latter. Finding *(a)* is in favor of the OT-procedure while finding *(b)* is in favor of the super learner. On the other hand, for $u = 49$ and $u = 78$ both, when centering on cities that will submit a request eventually: *(c)* the median of $\{\widehat{\zeta}_{\alpha,3}^{\mathrm{OT},u} : \alpha \in \mathcal{A}_3 \text{ st } (\zeta_{\alpha,3,u}, \zeta_{\alpha,3,u^-}) = (1,0)\}$ is larger than that of $\{\widehat{\zeta}_{\alpha,3}^{\mathrm{SL},u} : \alpha \in \mathcal{A}_3 \text{ st } (\zeta_{\alpha,3,u}, \zeta_{\alpha,3,u^-}) = (1,0)\}$ while *(d)* the 10%- and 90%-quantiles of the former set are smaller than that of the latter. Finding *(c)* is in favor of the OT-procedure while finding *(d)* is in favor of the super learner.

Figure 5.6 pays special attention to the medians, representing those of the predicted probabilities of submitting a request for the government declaration of natural disaster for a drought event for year 2021 as output by the super learner, the OT-procedure and the hybrid procedure as weeks go by, its right-hand side and left-hand side panels focusing on cities that will and that will not submit a request eventually. A clear pattern emerges: when centering on cities that will not submit a request eventually, the week-specific median of the predictions made by the super learner is consistently larger than that of the predictions made by our procedure which, in turn, is consistently larger than that of the predictions made by the hybrid procedure. Conversely, when concentrating on cities that will submit a request eventually, the week-specific median of the predictions made by the super learner is consistently smaller than that of predictions made by the OT-procedure which, in turn, is consistently larger than that of the predictions made by the hybrid procedure. From this perspective, the hybrid procedure outperforms the OT-procedure which, in turn, performs better than the super learner.

To conclude, we report in Table 5.5 the week-specific MSE scores

$$\frac{\sum_{\alpha \in \mathcal{A}_3} (\widehat{\zeta}_{\alpha,3}^{\bullet,u} - \zeta_{\alpha,3})^2 \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}}{\sum_{\alpha \in \mathcal{A}_3} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}} \tag{5.18}$$

(all $u \in \mathcal{U}_3$, the symbol $\bullet$ standing for SL, OT and HYB). The key insight from Table 5.5 is that the hybrid procedure exhibits superior performance, by consistently outperforming both the OT-procedure and the super learner. Interestingly we also observe that, for every procedure, (5.18) decreases as $u \in \mathcal{U}_3$ increases, suggesting that the challenge of forecasting which cities will eventually request the government declaration of natural disaster for a drought event becomes progressively less challenging as the weeks go by. The evolution of (5.18) for $u \in \mathcal{U}_3$ is represented in Figure 5.7, with those of the stock of requests already submitted ($u \mapsto \sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3,u}$, necessarily increasing) and of the sum of the predicted probabilities that the cities which have not yet submitted such a request will eventually do,

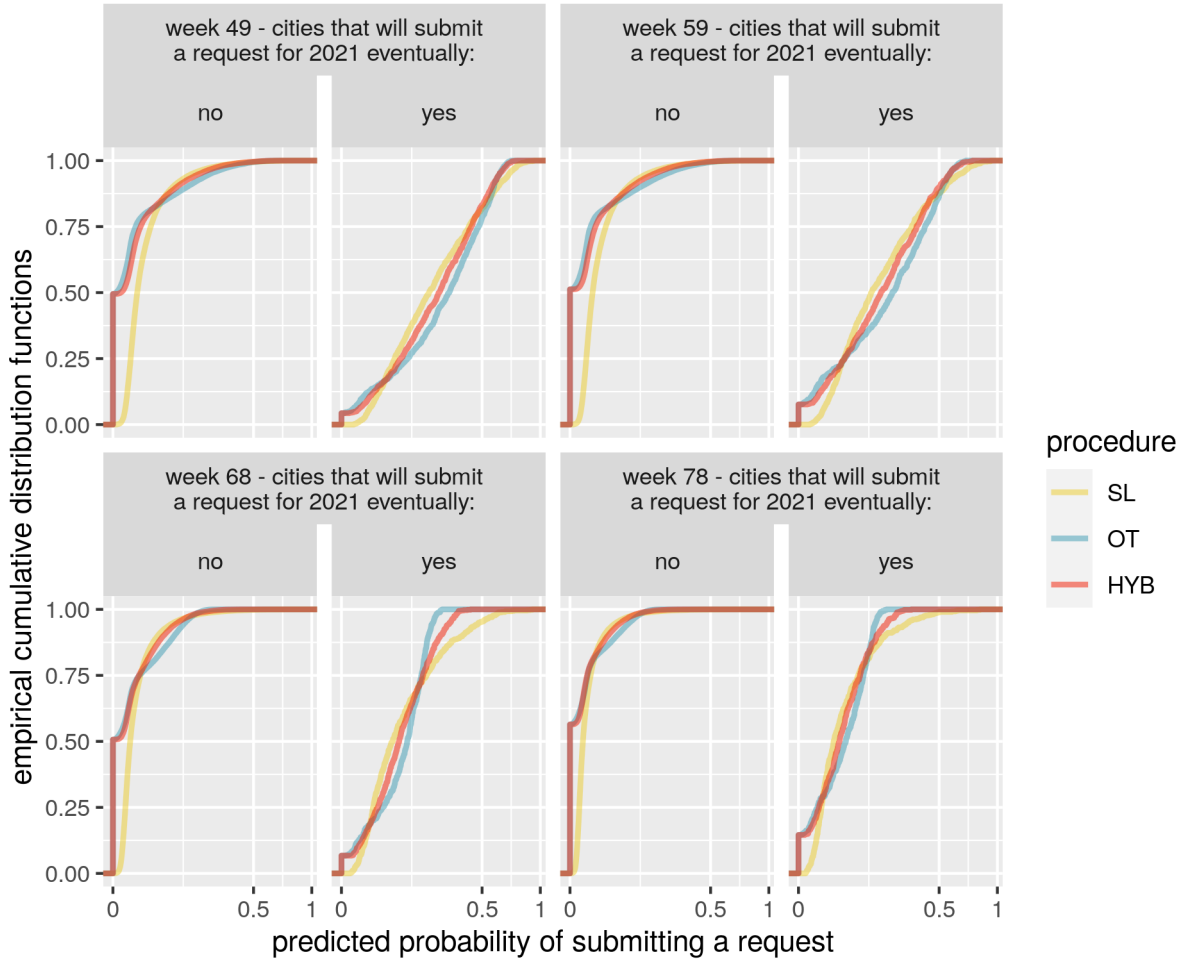

Figure 5.4 – This plot shows, when week u is one of the 49th week of 2021 (December 6th to 12th), the $(59-52)=7$th, $(69-52)=17$th and $(78-52)=26$th weeks of 2022 (February 15th to 21st, April 26th to May 2nd, June 28th to July 4th), the empirical cumulative distribution functions (ecdfs) of the predicted probabilities of submitting a request made by procedures SL, OT and HYB separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, the ecdfs of $\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3}=0\}$, left-hand side panels) and for those that will (that is, the ecdfs of $\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3}=1\}$, right-hand side panels). See also Figure 5.6 for a focus on medians.

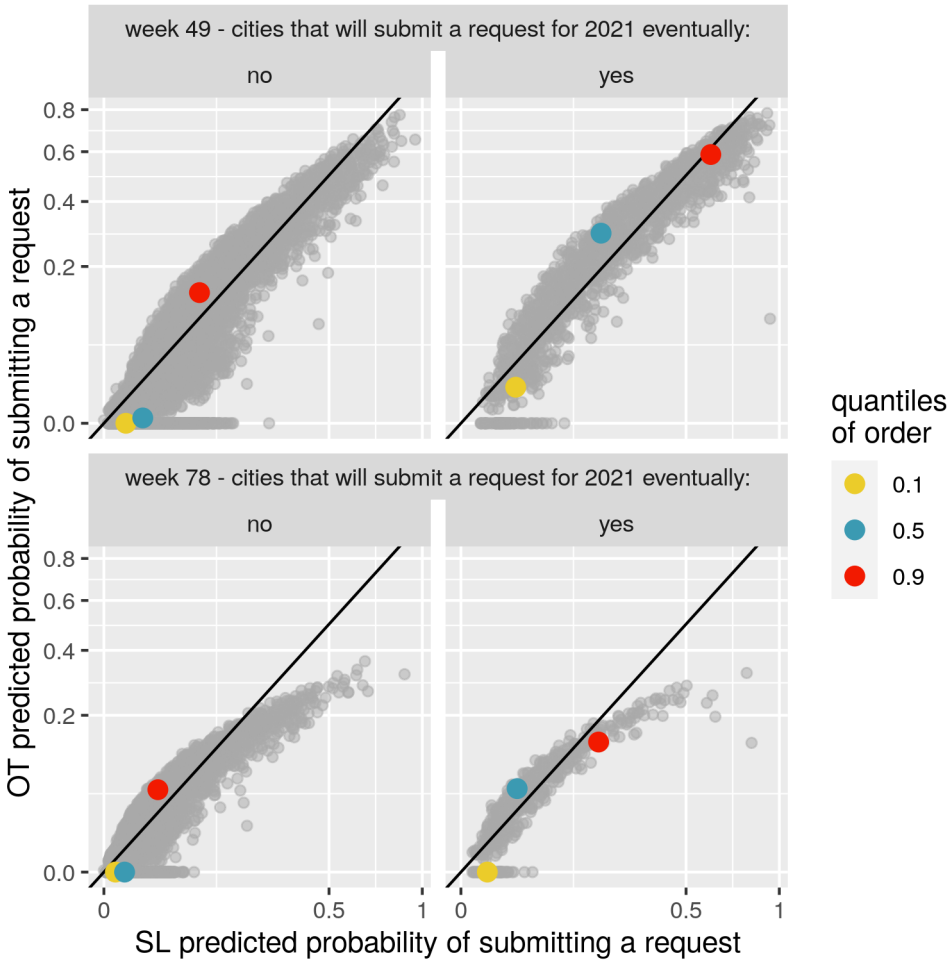

Figure 5.5 – This plot shows, for week u equal either to the 49th week of 2021 (December 6th to December 12th) or the $(78-52)=26$th week of 2022 (June 27th to July 3rd), the predicted probabilities of submitting a request made by procedures SL (x-axis) and OT (y-axis) separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, $\{(\widehat{\zeta}^{\mathsf{SL},u}_{\alpha,3}, \widehat{\zeta}^{\mathsf{OT},u}_{\alpha,3}) : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u}=0, \zeta_{\alpha,3}=0\}$, left-hand side panels) and for those that will (that is, $\{(\widehat{\zeta}^{\mathsf{SL},u}_{\alpha,3}, \widehat{\zeta}^{\mathsf{OT},u}_{\alpha,3}) : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u}=0, \zeta_{\alpha,3}=1\}$, right-hand side panels). In addition, three colored points represent in each panel the coordinate-specific quantiles of order 10%, 50% and 90%.

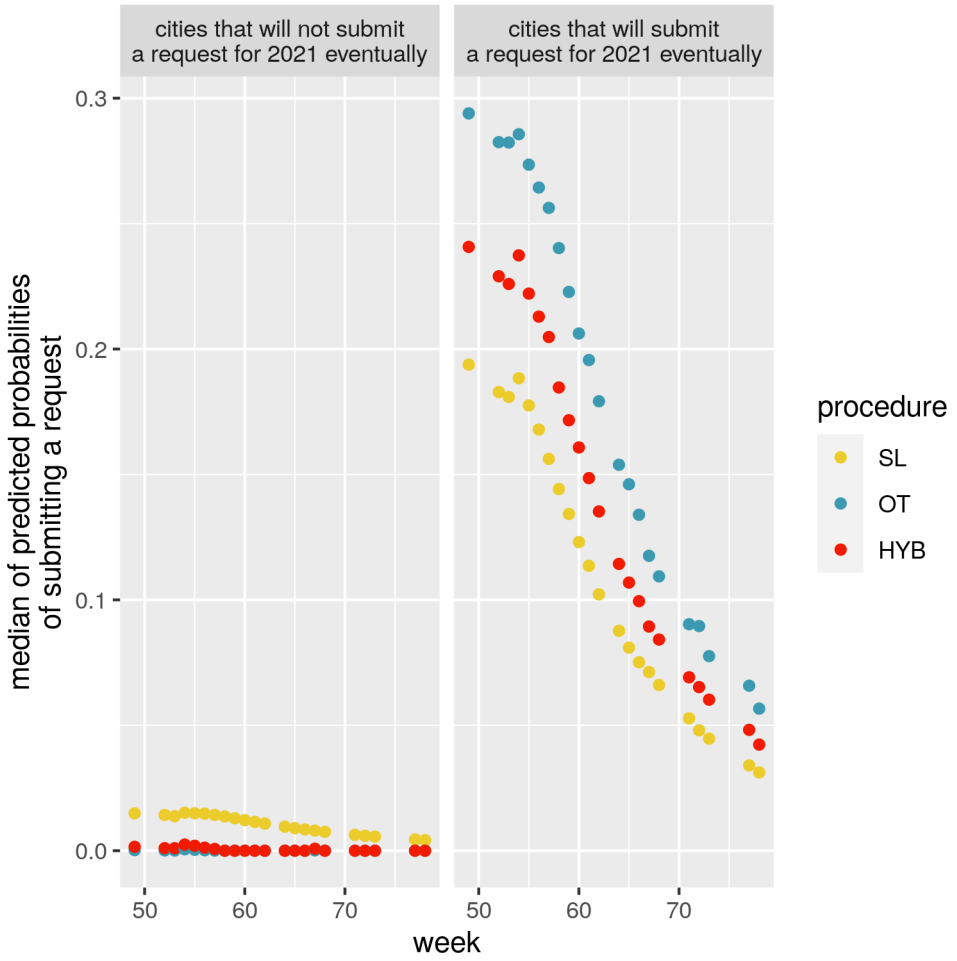

Figure 5.6 – This plot shows, as week u goes from the 49th week of 2021 (December 6th to December 12th) to the $(78-52)=26$th week of 2022 (June 27th to July 3rd), the evolutions of the medians of the predicted probabilities of submitting a request made by procedures SL, OT and HYB separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, of $u \mapsto \text{median}\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u}=0, \zeta_{\alpha,3}=0\}$, left-hand side panel) and for those that will (that is, of $u \mapsto \text{median}\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u}=0, \zeta_{\alpha,3}=1\}$, right-hand side panel). See also Figure 5.4 for more comprehensive descriptions through empirical cumulative distribution functions.

	MSE				MSE		
week u	SL	OT	HYB	week u	SL	OT	HYB
49	0.0341	0.0341	**0.0333**	62	0.0236	0.0241	**0.0231**
52	0.0336	0.0333	**0.0327**	64	0.0223	0.0228	**0.0219**
53	0.0332	0.0331	**0.0324**	65	0.0216	0.0221	**0.0212**
54	0.0317	0.0321	**0.0309**	66	0.0208	0.0214	**0.0205**
55	0.0307	0.0311	**0.0299**	67	0.0202	0.0203	**0.0198**
56	0.0294	0.0302	**0.0288**	68	0.0195	0.0195	**0.0190**
57	0.0281	0.0290	**0.0275**	71	0.0179	0.0180	**0.0176**
58	0.0268	0.0280	**0.0264**	72	0.0177	0.0177	**0.0174**
59	0.0258	0.0271	**0.0255**	73	0.0168	0.0168	**0.0165**
60	0.0248	0.0261	**0.0245**	77	0.0156	0.0156	**0.0154**
61	0.0242	0.0248	**0.0237**	78	0.0150	0.0150	**0.0148**

Table 5.5 – Evolution of MSE $u \mapsto n_{3,u}^{-1} \sum_{\alpha \in \mathcal{A}_3} (\widehat{\zeta}_{\alpha,3}^{\bullet,u} - \zeta_{\alpha,3})^2 \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ where $n_{3,u} := \sum_{\alpha \in \mathcal{A}_3} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ is the number of cities which have not submitted such a request yet at week $u \in \mathcal{U}_3$ and the symbol $\bullet$ stands for SL, OT, HYB. In each row, the smallest value stand out in bold characters. See also Figure 5.7.

according to the hybrid procedure ($u \mapsto \sum_{\alpha \in \mathcal{U}_3} \widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$). The quartiles and range of

$$\left\{ \sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3,u} + \sum_{\alpha \in \mathcal{U}_3} \widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\} : u \in \mathcal{U}_3 \right\} \tag{5.19}$$

(the heights of the bars in Figure 5.7) are 1572 (minimum), 1636 (first quartile), 1731 (median), 1853 (third quartile), 1908 (maximum), 336 (range) while its mean is 1740. In comparison, the quartiles and range of

$$\left\{ \sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3,u} + \sum_{\alpha \in \mathcal{U}_3} \widehat{\zeta}_{\alpha,3}^{\mathrm{SL},u} \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\} : u \in \mathcal{U}_3 \right\} \tag{5.20}$$

are 1662 (minimum), 1776 (first quartile), 1881 (median), 2051 (third quartile), 2133 (maximum), 471 (range), while its mean is 1905 – note that we could have substituted OT for SL in the above display. In view of (5.17), it was guaranteed that each of the quartile and mean associated to (5.19) would be smaller then its counterpart associated to (5.20). Both convex hulls of (5.19) and (5.20) contain the true value $\sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3} = 1696$, the former being more concentrated around it than the latter. This last observation stems from a comparison of the ranges of the sets and can be further substantiated by comparing the interquartile intervals, with that of (5.19) encompassing the true value, unlike that of (5.20).

5.6.4 On the importance of the variables used to make predictions

In this last subsection, we consider the influence that each covariate $\xi_{\alpha,3,u,s}$ (note the additional subscript s, indicating the sth covariate) in a generic $\xi_{\alpha,3,u}$ has on the prediction $\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u}$ that city $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$ will eventually submit a request for the government declaration of natural disaster for a drought event relative to year 2021 based on data available at week $u \in \mathcal{U}_3$. The question pertains to the definition and estimation of variable importance measures. The literature on this topic is rich, with notable contributions from studies such as (van der Laan, 2006; Hubbard et al., 2016; Williamson et al., 2021)

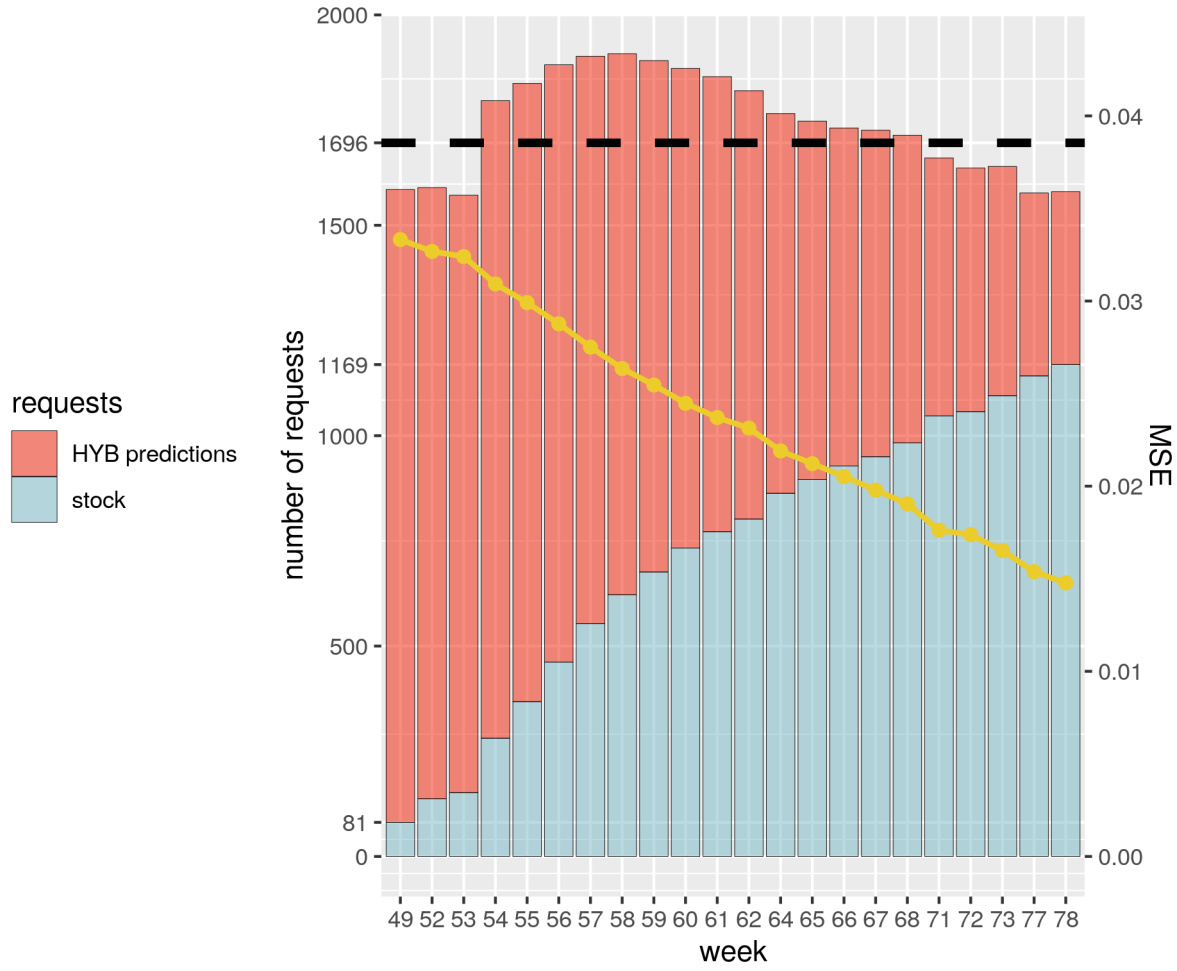

Figure 5.7 – This plot shows, as week u goes from the 49th week of 2021 (December 6th to 12th) to the $(78-52)=26$th week of 2022 (June 27th to July 3rd), the evolutions of the cardinality of the stock of requests already submitted for the government declaration of natural disaster for a drought event for year 2021 (that is, of $u \mapsto \sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,3,u}$, in blue) and of the sum of the predicted probabilities that the cities which have not yet submitted such a request will eventually do (that is, of $u \mapsto \sum_{\alpha \in \mathcal{A}_t} \widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} \mathbf{1}\{\zeta_{\alpha,3,u}=0\}$, in red). The actual eventual number of such requests (that is, $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,3}$, which equals 1696) is also represented (horizontal dashed line). In addition, the plot shows the evolution of MSE (that is, of $u \mapsto n_{3,u}^{-1} \sum_{\alpha \in \mathcal{A}_3} (\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} - \zeta_{\alpha,3})^2 \mathbf{1}\{\zeta_{\alpha,3,u}=0\}$ where $n_{3,u} := \sum_{\alpha \in \mathcal{A}_3} \mathbf{1}\{\zeta_{\alpha,3,u}=0\}$ is the number of cities which have not submitted such a request yet at week u, in yellow). See also Table 5.5.

on the one hand and (Lundberg and Lee, 2017, and references therein) on the other hand, offering valuable insights on how to tackle this question. However, applying these existing approaches to our specific scenario is impractical, mainly due to the interdependence of the data-structures specific to each $(\alpha, u) \in \mathcal{A}_3 \times \mathcal{U}_3$ and the fact that we are dealing with a relatively large number of covariates. As a result, we propose a simple approach tailored to the circumstances of the present situation. The approach is very similar to the one developed in (Ecoto and Chambaz, 2022, Section 4.4).

Set arbitrarily $s \in [\![67]\!]$ and $u \in \mathcal{U}_3$.

- If s is such that the covariate $\xi_{\alpha,3,u,s}$ corresponds to the overall number of French cities that submitted a request for year 2021 during week u or before, or to the ratio of the logarithm of that overall number to u (two elements of the description of a city's request history), then we cannot quantify the covariate's importance because all cities $\alpha \in \mathcal{U}_3$ share a common value.

- If s is such that $\xi_{\alpha,3,u,s}$ ($\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$) takes v values with $2 \leq v \leq 5$, then we let ρ_s be the correlation ratio computed based on $\{(\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3}, \xi_{\alpha,3,u,s}) : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$:

$$\rho^u_s := \left(\frac{\sum_{\nu=1}^{v} n_\nu (\bar{\zeta}_\nu - \bar{\zeta})^2}{\sum_{\alpha \in \mathcal{A}_3} (\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3} - \bar{\zeta})^2 \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}} \right)^{1/2}$$

  where $\bar{\zeta}_\nu$ is the average of the $\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3}$s such that $\xi_{\alpha,3,u,s} = \nu$ and $\bar{\zeta}$ is the average of all $\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3}$s.

- Otherwise, we treat the covariate $\xi_{\alpha,3,u,s}$ ($\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$) as a continuous variable and let ρ^u_s be the absolute value of the Spearman rank correlation coefficient (Hollander and Wolfe, 1999, Section 8.5) computed based on $\{(\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3}, \xi_{\alpha,3,u,s}) : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$.

Note that, in the second case, we could have defined ρ^u_s as Wilcoxon test's statistic (case $v = 2$) or the Kruskal-Wallis test's statistics (case $3 \leq v \leq 5$) (see Hollander and Wolfe, 1999, Sections 3.1 and 6.1). By guaranteeing that all ρ^u_ss naturally lie in $[0, 1]$, the present choice eases comparisons.

In all cases, the magnitude of ρ^u_s directly reflects the strength of the association between the sth covariate and the predictions made at week $u \in \mathcal{U}_3$. We resort to permutation tests to assess significance levels, with one million independent permutations drawn uniformly in each of the above cases. The maximum value obtained by permutation equals 3.16%.

Figure 5.8 shows the evolutions of $u \mapsto \rho^u_s$ for every eligible $s \in [\![67]\!]$, where the covariates are grouped based on the type of information they contribute. In each panel, values above the black horizontal lines (y-intercept at $(0, 3.16\%)$) are considered highly significant according to the permutation tests. From this perspective, most covariates play an effective role in the predictions. For the covariates related to a city's description, its exposure to drought events, or its request history, the curves appear relatively flat, indicating a steady strength of association with the predictions over time. In contrast, for the covariates describing a city's vicinity, the curves lying above the horizontal line show an increasing trend before levelling off. This suggests that the strength of association for each corresponding covariate gradually increases then stabilizes over time. In Table 5.6, we report the five variables which, in each group of covariates, feature the largest average variable importance ($\sum_{u \in \mathcal{U}_3} \rho^u_s / \operatorname{card} \mathcal{U}_3$).

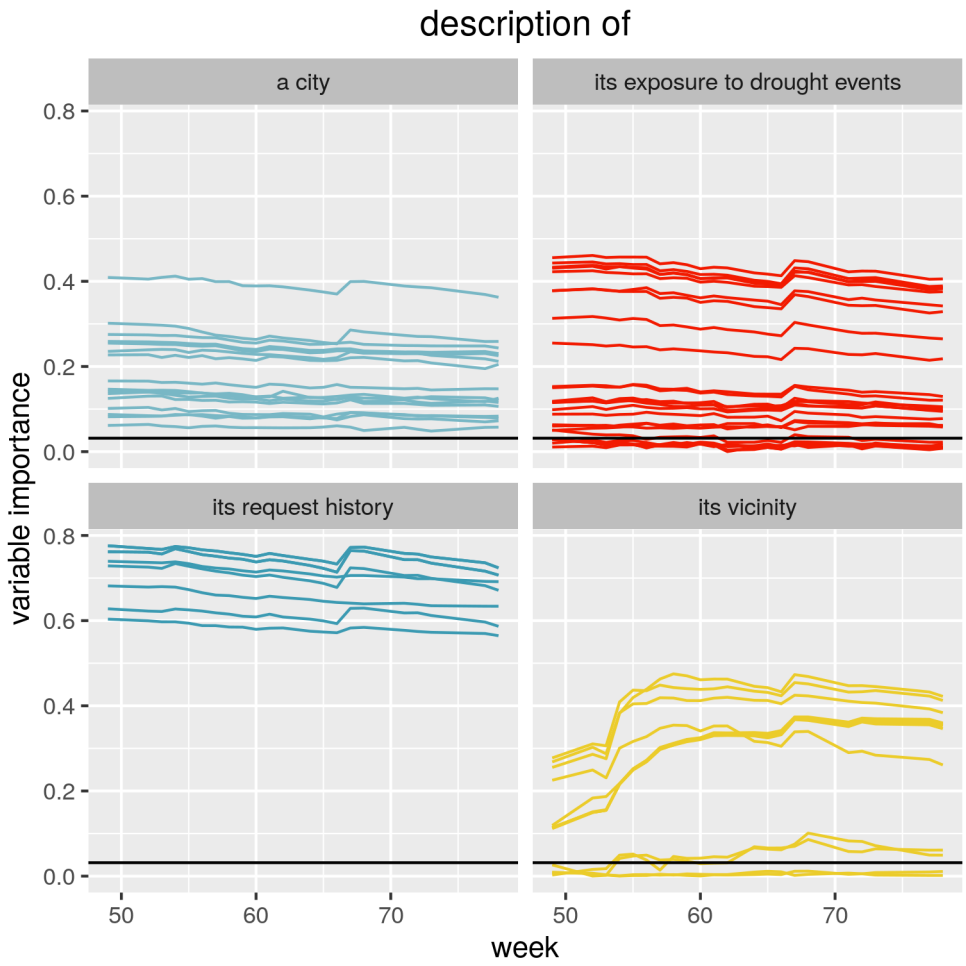

Figure 5.8 – This plot shows, as week u goes from the 49th week of 2021 (December 6th to 12th) to the $(78-52)=26$th week of 2022 (June 27th to July 3rd), the evolutions of the importance of each variable used to make predictions, as defined in Section 5.6.4. For every eligible $s \in [\![67]\!]$, the larger is ρ_s^u, the stronger is the association between the sth covariate $\xi_{\alpha,3,u,s}$ and the prediction $\widehat{\zeta}_{\alpha,3}^{\mathsf{HYB},u}$ across $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u}=0$. Values above the black horizontal lines are deemed highly significant based on permutation tests. See also Table 5.6.

description of	variable	avg. importance
a city	proportion of houses* in the 2nd clay-shrinkage-swelling hazard category	0.392
	climatic zone	0.275
	insured sum	0.259
	number of houses*	0.244
	population	0.239
its exposure to drought events	average SWI over Q1, Q2, Q3†	0.436
	overall average SWI	0.420
	average SWI over Q2, Q3	0.412
	minimum SWI over Q2	0.412
	global minimum SWI	0.402
its request history	number of requests submitted during the 5 previous years	0.757
	number of requests submitted since 1990	0.744
	number of requests denied during the 2 previous years	0.715
	number of requests granted during the 2 previous years	0.708
	indicator of request denied the previous year	0.654
its vicinity	number of claims in the same department	0.423
	proportion of cities in the same department that submitted a request for year 2023 before week u	0.416
	proportion of cities in the same department that submitted a request for the first time during the 5 previous years	0.392
	ratio of the number of claims in the same department to the number of cities in the department	0.308
	number of neighboring cities that submitted a request for year 2023 before week u	0.305

* within the city's limits

† Q1, Q2, Q3, Q4 are the 1st to 4th quarters

Table 5.6 – The five variables used to make predictions with the highest average importance ($\sum_{u\in\mathcal{U}_3} \rho_s^u / \operatorname{card}\mathcal{U}_3$, see definition in Section 5.6.4) in each group of covariates. For every eligible $s \in [\![67]\!]$, the larger is ρ_s^u, the stronger is the association between the sth covariate $\xi_{\alpha,3,u,s}$ and the prediction $\widehat{\zeta}_{\alpha,3}^{\mathsf{HYB},u}$ across $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. See also Figure 5.8.

5.7 Appendix: checking the iPiano assumptions

The iPiano assumptions consist in

1. f being C^1-smooth with a Lipschitz continuous gradient on $\operatorname{dom} g_\tau$, see Section 5.7.1;
2. for any $\delta > 0$, $H_\delta : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ given by $H_\delta(\theta, \theta') := f(\theta') + g_\tau(\theta') + \delta\|\theta - \theta'\|_2^2$ having the Kurdyka-Lojasiewicz property at a cluster point $(\theta^\star, \theta^\star)$ of the sequence $(\theta^k)_{k\geq 1}$, see Section 5.7.2.

5.7.1 The function f is C^1-smooth and its gradient is Lipschitz continuous on $\operatorname{dom}\ g_\tau$

5.7.1.a Preliminaries

ON MATRIX NORMS. For self-containedness, let us recall several definitions and results concerning matrix norms. For any matrix $A \in \mathbb{R}^{d\times d'}$, the Frobenius and maximum norms of A are given by $\|A\|_F := \left(\sum_{i\in[\![d]\!],j\in[\![d']\!]} A_{i,j}^2\right)^{1/2}$ and $\|A\|_{\max} := \max\{|A_{i,j}| : i \in [\![d]\!], j \in [\![d']\!]\}$. For any vector $x \in \mathbb{R}^d$, the variation seminorm of x is defined as $\|x\|_{\mathrm{var}} := \max\{x_i : i \in [\![d]\!]\} - \min\{x_i : i \in [\![d]\!]\}$. We will use the following classical inequalities and equality:

$$\forall A \in \mathbb{R}^{d\times d'}, \forall B \in \mathbb{R}^{d'\times d''}, \|AB\|_F \leq \|A\|_F\|B\|_F; \tag{5.21}$$

$$\forall A \in \mathbb{R}^{d\times d'}, \forall x \in \mathbb{R}^{d'}, \|Ax\|_2 \leq \|A\|_F\|x\|_2; \tag{5.22}$$

$$\forall x \in \mathbb{R}^d, \|\operatorname{diag}(x)\|_F = \|x\|_2; \tag{5.23}$$

$$\forall x \in \mathbb{R}^d, \|x\|_{\mathrm{var}} \leq 2\|x\|_\infty; \tag{5.24}$$

$$\forall x \in \{0\} \times \mathbb{R}^{d-1}, \|x\|_\infty \leq \|x\|_{\mathrm{var}}. \tag{5.25}$$

ON THE HILBERT PROJECTIVE METRIC. The Hilbert projective metric on $(\mathbb{R}_+^*)^d$ is defined by

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') := \log\max\left\{\tfrac{x_i x'_j}{x'_i x_j} : i, j \in [\![d]\!]\right\}.$$

We will use the following properties (Birkhoff, 1957):

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = \|\log(x) - \log(x')\|_{\mathrm{var}}; \tag{5.26}$$

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = d_{\mathcal{H}}(x/x', \mathbf{1}_d) = d_{\mathcal{H}}(\mathbf{1}_d/x', \mathbf{1}_d/x); \tag{5.27}$$

$$\forall K \in (\mathbb{R}_+^*)^{d\times d'}, \forall x, x' \in (\mathbb{R}_+^*)^{d'}, d_{\mathcal{H}}(Kx, Kx') \leq \lambda(K) d_{\mathcal{H}}(x, x'), \tag{5.28}$$

where $\lambda(K) := \frac{\sqrt{\eta(K)}-1}{\sqrt{\eta(K)}+1} < 1$ with $\eta(K) := \max\left\{\frac{K_{i,k}K_{j,\ell}}{K_{j,k}K_{i,\ell}} : i, j \in [\![d]\!], k, \ell \in [\![d']\!]\right\}$.

We end this section with a lemma.

Lemma 5.1. *Let $x, x' \in (\mathbb{R}_+^*)^d$ be such that $0 < t \leq \min\{x_j, x'_j : j \in [\![d]\!]\} \leq \max\{x_j, x'_j : j \in [\![d]\!]\} \leq T$. It holds that $\frac{1}{2}td_{\mathcal{H}}(x, x') \leq \|x - x'\|_2$. Moreover, if $x_1 = x'_1 = 1$, then it also holds that $\|x - x'\|_2 \leq \sqrt{d}Td_{\mathcal{H}}(x, x')$.*

Proof. Set $x, x' \in (\mathbb{R}_+^*)^d$ as in the statement of the lemma, and denote $\ell := \log(x), \ell' := \log(x')$ (the logarithms are elementwise). Set arbitrarily $i \in [\![d]\!]$. We can assume without

loss of generality that $x_i \geq x'_i$ (or, equivalently, $\ell_i \geq \ell'_i$). Therefore if $x_1 = x'_1 = 1$ (or, equivalently, $\ell_1 = \ell'_1 = 0$), then

$$\begin{aligned}
|x_i - x'_i| &= \max(x_i, x'_i) \times |1 - e^{-|\ell_i - \ell'_i|}| \\
&\leq T \times |\ell_i - \ell'_i| \quad \text{because } |1 - e^{-|q|}| \leq |q| \text{ for all } q \in \mathbb{R} \\
&\leq T \times \|\ell - \ell'\|_\infty \\
&\leq T \times \|\ell - \ell'\|_{\mathrm{var}} \quad \text{by (5.25) since } \ell_1 = \ell'_1 = 0 \\
&= T d_{\mathcal{H}}(x, x') \quad \text{by (5.26).}
\end{aligned}$$

Consequently, $\|x - x'\|_2 \leq \sqrt{d}\|x - x'\|_\infty \leq \sqrt{d} T d_{\mathcal{H}}(x, x')$. Furthermore,

$$\begin{aligned}
|x_i - x'_i| &= \min(x_i, x'_i) \times |e^{|\ell_i - \ell'_i|} - 1| \\
&\geq t \times |\ell_i - \ell'_i| \quad \text{because } |e^{|q|} - 1| \geq |q| \text{ for all } q \in \mathbb{R}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\|x - x'\|_2 \geq \|x - x'\|_\infty \geq t\|\ell - \ell'\|_\infty &\geq \tfrac{1}{2} t \|\ell - \ell'\|_{\mathrm{var}} \quad \text{by (5.24)} \\
&= \tfrac{1}{2} t d_{\mathcal{H}}(x, x') \quad \text{by (5.26).}
\end{aligned}$$

This completes the proof. □

5.7.1.b The function f is differentiable

To prove that f is differentiable, we rely on the following classical result (Danskin, 1966):

Theorem 5.1 (Danskin's theorem, Proposition B.25 in Bertsekas (1999))**.** *Let $\mathcal{C} \subset \mathbb{R}^{d'}$ be a compact set and $\phi : \mathbb{R}^d \times \mathcal{C} \to \mathbb{R}$ be a continuous function such that $\phi(\cdot, y)$ is convex for every $y \in \mathcal{C}$. The function $\psi : \mathbb{R}^d \to \mathbb{R}$ given by $\psi(x) := \max_{y \in \mathcal{C}} \phi(x, y)$ is convex. Moreover, if there exists a unique $\hat{y}$ maximizing $\phi(x, \cdot)$ and if $\phi(\cdot, \hat{y})$ is differentiable, then ψ is differentiable at x and $\nabla\psi(x) = \nabla\phi(\cdot, \hat{y})|_x$.*

Let $\mathcal{C} = \Pi_{R,R'}$ (a compact set) and $\phi : \mathbb{R}^{R \times R'} \times \Pi_{R,R'} \to \mathbb{R}$ be given by $\phi(C, P) := -[\langle P, C \rangle - \gamma E(P)]$. The function ϕ is continuous and $\phi(\cdot, P)$ is convex for every $P \in \Pi_{R,R'}$. Therefore, by the above theorem, the function $\psi : \mathbb{R}^{R \times R'} \to \mathbb{R}$ given by $\psi(C) := \max_{P \in \Pi_{R,R'}} \phi(C, P) = -\mathcal{W}_\gamma(C)$ is convex. Moreover, for every $C \in \mathbb{R}^{R \times R'}$, there exists a unique $\widehat{P}_C$ such that $\psi(C) = \phi(C, \widehat{P}_C)$ (Cuturi and Doucet, 2014, Proposition 4.3) and $\phi(\cdot, \widehat{P}_C)$ is affine hence differentiable. Therefore, $C \mapsto \mathcal{W}_\gamma(C)$ is differentiable at every $C \in \mathbb{R}^{R \times R'}$ with a gradient given by $\nabla \mathcal{W}_\gamma(C) = \widehat{P}_C$.

We use now that $f = f_a - \frac{1}{2} f_b + \text{constant}$ with $f_a, f_b : \mathbb{R}^N \to \mathbb{R}$ given by

$$f_a(\theta) := \mathcal{W}_\gamma\left(C(\mathbf{z}, \mathbf{z}'(\theta))\right) \quad \text{and} \quad f_b(\theta) := \mathcal{W}_\gamma\left(C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))\right)$$

where the cost matrices $C(\mathbf{z}, \mathbf{z}'(\theta))$ and $C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))$ are such that $(C(\mathbf{z}, \mathbf{z}'(\theta)))_{m,n} := \mathrm{dis}(x_m, x'_n)^2 + (y_m - \theta_n)^2$ and $(C(\mathbf{z}'(\theta), \mathbf{z}'(\theta)))_{n,n'} := \mathrm{dis}(x'_n, x'_{n'})^2 + (\theta_n - \theta_{n'})^2$. In view of the previous paragraph, and by the chain rule, f_a and f_b are thus differentiable at every $\theta \in \mathbb{R}^N$ with gradients

$$\nabla f_a(\theta) = 2(\tfrac{1}{N}\theta - \widehat{P}_\theta^\top y) \quad \text{and} \quad \nabla f_b(\theta) = 2(\tfrac{2}{N}\theta - (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta)$$

($\widehat{P}_\theta$ and $\widehat{Q}_\theta$ are defined in (5.9) and (5.10)). Therefore f is differentiable at every $\theta \in \mathbb{R}^N$ and (5.8) follows straightforwardly.

5.7.1.c $\widehat{P}_\theta$ and $\widehat{Q}_\theta$ are Lipschitz continuous (as functions of θ)

The fact that $\theta \mapsto \widehat{P}_\theta$ and $\theta \mapsto \widehat{Q}_\theta$ are Lipschitz continuous on $\operatorname{dom} g_\tau$ is a consequence of the following lemma.

Lemma 5.2. *Let $\theta \mapsto C(\theta)$ be a bounded and Lipschitz continuous function from $[0,1]^{R'}$ to $\mathbb{R}_+^{R\times R'}$. For each $\theta \in [0,1]^{R'}$, let $\widehat{P}(\theta)$ be the minimizer in (5.3) with $C(\theta)$ substituted for C. Then $\theta \mapsto \widehat{P}(\theta)$ is Lipschitz continuous from $[0,1]^{R'}$ to $\mathbb{R}_+^{R\times R'}$.*

Indeed, $\theta \mapsto C(\mathbf{z}, \mathbf{z}'(\theta))$ and $\theta \mapsto C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))$ (defined in Section 5.7.1.b) are obviously bounded and Lipschitz continuous.

Let us prove Lemma 5.2. By (Cuturi and Doucet, 2014, Proposition 4.3), for every $\theta \in \mathbb{R}^{R'}$,

$$\widehat{P}(\theta) = \operatorname{diag}(\widehat{u}(\theta)) K(\theta) \operatorname{diag}(\widehat{v}(\theta)),$$

where $\widehat{u} : \mathbb{R}^{R'} \to (\mathbb{R}_+^*)^R$, $\widehat{v} : \mathbb{R}^{R'} \to (\mathbb{R}_+^*)^{R'}$ and the Gibbs kernel functions $K : \mathbb{R}^{R'} \to \mathbb{R}^{R\times R'}$, given by

$$K(\theta) := \left(\exp\left[-\left(C(\theta)\right)_{r,r'}/\gamma\right]\right)_{r\in[\![R]\!], r'\in[\![R']\!]}$$

satisfy the mass conservation constraints inherent to $\Pi_{R,R'}$:

$$\operatorname{diag}(\widehat{u}(\theta)) K(\theta) \operatorname{diag}(\widehat{v}(\theta))\, \mathbf{1}_{R'} = \tfrac{1}{R}\, \mathbf{1}_R \tag{5.29}$$

$$\operatorname{diag}(\widehat{v}(\theta)) K(\theta)^\top \operatorname{diag}(\widehat{u}(\theta))\, \mathbf{1}_{R} = \tfrac{1}{R'}\, \mathbf{1}_{R'}, \tag{5.30}$$

Equivalently, using the entrywise division of vectors,

$$\widehat{u}(\theta) = \frac{\frac{1}{R}\,\mathbf{1}_R}{K(\theta)\widehat{v}(\theta)}, \quad \widehat{v}(\theta) = \frac{\frac{1}{R'}\,\mathbf{1}_{R'}}{K(\theta)^\top \widehat{u}(\theta)}. \tag{5.31}$$

Note that $(\rho\widehat{u}(\theta), \widehat{v}(\theta)/\rho)$ also satisfy (5.29) and (5.30) for any $\rho > 0$. Thus, without loss of generality, we can impose from now on that, for all $\theta \in \operatorname{dom} g_\tau$, the first element $\widehat{u}_1(\theta)$ of $\widehat{u}(\theta)$ equals 1 (this affects both $\widehat{u}(\theta)$ and $\widehat{v}(\theta)$).

We now consider the following steps.

- The Gibbs kernel function K is Lipschitz continuous on $\operatorname{dom} g_\tau$ with Lipschitz constant $L_K := k_u^2 L_C^2/\gamma^2$ where $k_u := \max\{(K(\theta))_{r,r'} : r \in [\![R]\!], r' \in [\![R']\!]\}$ and L_C is the Lipschitz constant of $\theta \mapsto C(\theta)$.
  Proof: The function $\theta \mapsto C(\theta)$ is bounded, so $\theta \mapsto K(\theta)$ is bounded as well. For all $\theta, \theta' \in [0,1]^{R'}$, $r \in [\![R]\!]$ and $r' \in [\![R']\!]$, it holds that

$$\begin{aligned}
&|(K(\theta))_{r,r'} - (K(\theta'))_{r,r'}| \\
&= \max\{e^{-(C(\theta))_{r,r'}/\gamma}, e^{-(C(\theta'))_{r,r'}/\gamma}\} \times |1 - \exp(-|(C(\theta))_{r,r'} - (C(\theta'))_{r,r'}|/\gamma)| \\
&\leq \frac{k_u}{\gamma} \times |(C(\theta))_{r,r'} - (C(\theta'))_{r,r'}|.
\end{aligned}$$

Therefore,

$$\begin{aligned}\|K(\theta)-K(\theta')\|_F^2 &= \sum_{r\in[\![R]\!],r'\in[\![R']\!]}[(K(\theta))_{r,r'}-(K(\theta'))_{r,r'}]^2\\ &\le \frac{k_u^2}{\gamma^2}\sum_{r\in[\![R]\!],r'\in[\![R']\!]}[(C(\theta))_{r,r'}-(C(\theta'))_{r,r'}]^2\\ &\le \frac{k_u^2L_C^2}{\gamma^2}\|\theta-\theta'\|_2^2.\end{aligned}$$

- Denote $k_\ell := \min\{(K(\theta))_{r,r'} : r\in[\![R]\!], r'\in[\![R']\!]\}$. For every $\theta\in\operatorname{dom} g_\tau$,

$$\lambda(K(\theta)) \le \Lambda := (k_u-k_\ell)/(k_u+k_\ell) < 1. \tag{5.32}$$

Proof: Because $k_\ell \le (K(\theta))_{r,r'} \le k_u$ for all $\theta\in\operatorname{dom} g_\tau$, $r\in[\![R]\!], r'\in[\![R']\!]$, it holds that $(K(\theta))_{i,k}(K(\theta))_{j,\ell}/((K(\theta))_{j,k}(K(\theta))_{i,\ell}) \le k_u^2/k_\ell^2$ for all $i,j\in[\![R]\!], k,\ell\in[\![R']\!]$. Consequently, $\eta(K(\theta)) \le k_u^2/k_\ell^2$ hence $\lambda(K(\theta)) = (\sqrt{\eta(K)}-1)/(\sqrt{\eta(K)}+1) \le (k_u-k_\ell)/(k_u+k_\ell)$.

- For every $\theta\in\operatorname{dom} g_\tau$, $\widehat{u}(\theta)$ and $\widehat{v}(\theta)$ are uniformly bounded: for all $r\in[\![R]\!]$, $r'\in[\![R']\!]$,

$$\frac{k_\ell}{k_uR'} \le \widehat{u}_r(\theta) \le \frac{k_uR}{k_\ell}, \tag{5.33}$$

$$\frac{k_\ell}{k_u^2R'R^2} \le \widehat{v}_{r'}(\theta) \le \frac{1}{k_\ell R}. \tag{5.34}$$

Proof: Set arbitrarily $\theta\in\operatorname{dom} g_\tau$. In view of (5.29) (first row), since $\widehat{u}_1(\theta)=1$, we have

$$k_\ell\|\widehat{v}(\theta)\|_\infty \le \frac{1}{R} = \sum_{r'\in[\![R']\!]}(K(\theta))_{1r'}\widehat{v}_{r'}(\theta) \le k_uR'\|\widehat{v}(\theta)\|_\infty. \tag{5.35}$$

Set $r_0'\in\arg\max\{\widehat{v}_i(\theta) : i\in[\![R']\!]\}$. In view of (5.30) ($r'$th row), we have

$$\frac{1}{R'} = \widehat{v}_{r_0'}(\theta)\sum_{r\in[\![R]\!]}(K(\theta))_{rr_0'}\widehat{u}_r(\theta) \ge k_\ell\|\widehat{v}(\theta)\|_\infty\|\widehat{u}(\theta)\|_\infty.$$

Hence, by (5.35),

$$\|\widehat{u}(\theta)\|_\infty \le \frac{1}{k_\ell R'\|\widehat{v}(\theta)\|_\infty} \le \frac{k_uRR'}{k_\ell R'} = \frac{k_uR}{k_\ell}. \tag{5.36}$$

Furthermore, for any $r'\in[\![R']\!]$, in view of (5.30) (r'th row) and (5.36),

$$\frac{1}{R'} = \widehat{v}_{r'}(\theta)\sum_{r\in[\![R]\!]}(K(\theta))_{rr'}\widehat{u}_r(\theta) \le Rk_u\|\widehat{u}(\theta)\|_\infty\widehat{v}_{r'}(\theta) \le \frac{k_u^2R^2}{k_\ell}\widehat{v}_{r'}(\theta). \tag{5.37}$$

The inequalities (5.35) and (5.37) readily imply (5.34). Likewise, for any $r\in[\![R]\!]$, in view of (5.29) (rth row),

$$\frac{1}{R} = \widehat{u}_r(\theta)\sum_{r'\in[\![R']\!]}(K(\theta))_{rr'}\widehat{v}_{r'}(\theta) \le R'k_u\|\widehat{v}(\theta)\|_\infty\widehat{u}_r(\theta) \le \frac{k_uR'}{k_\ell R}\widehat{u}_r(\theta). \tag{5.38}$$

The inequalities (5.36)and (5.38) readily imply (5.33).

- The function $\theta \mapsto \widehat{u}(\theta)$ is Lipschitz continuous on $\operatorname{dom} g_\tau$ with Lipschitz constant

$$L_{\widehat{u}} := \frac{2k_u^3 R^2 \sqrt{R'} L_K}{(1-\Lambda^2)k_\ell^4}(\sqrt{R} + \Lambda\sqrt{R'}).$$

Proof. Set arbitrarily $\theta, \theta' \in \operatorname{dom} g_\tau$. Inequalities (5.33) and (5.34) imply that

$$\begin{aligned}\min\{(K(\theta)\widehat{v}(\theta'))_r : r \in [\![R]\!]\} &\geq k_\ell^2/(k_u^2 R^2),\\ \min\{(K(\theta)^\top \widehat{u}(\theta'))_{r'} : r' \in [\![R']\!]\} &\geq k_\ell^2 R/(k_u R').\end{aligned}$$

In view of Lemma 5.1 (first inequality), (5.22) (second inequality), (5.34) and the fact that K is L_K-Lipschitz (third inequality), we obtain

$$\begin{aligned} d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)) &\leq \frac{2k_u^2 R^2}{k_\ell^2}\|K(\theta)\widehat{v}(\theta) - K(\theta')\widehat{v}(\theta)\|_2\\ &\leq \frac{2k_u^2 R^2}{k_\ell^2}\|K(\theta) - K(\theta')\|_F\|\widehat{v}(\theta)\|_2\\ &\leq \frac{2k_u^2 R\sqrt{R'} L_K}{k_\ell^3}\|\theta - \theta'\|_2. \end{aligned} \tag{5.39}$$

Likewise, using (5.33) instead of (5.34)

$$\begin{aligned} d_{\mathcal{H}}(K(\theta)^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta)) &\leq \frac{2k_u R'}{k_\ell^2 R}\|K(\theta_1)^\top\widehat{u}(\theta_1) - K(\theta_2)^\top\widehat{u}(\theta_1)\|_2\\ &\leq \frac{2k_u R'}{k_\ell^2 R}\|K(\theta)^\top - K(\theta')^\top\|_F\|\widehat{u}(\theta)\|_2\\ &\leq \frac{2k_u^2\sqrt{R}R' L_K}{k_\ell^3}\|\theta - \theta'\|_2. \end{aligned} \tag{5.40}$$

We can now bound the Hilbert projective metric between $\widehat{v}(\theta)$ and $\widehat{v}(\theta')$: by invoking in turn (5.31), (5.27), the triangle inequality, (5.28) and both (5.40) and (5.32), we get

$$\begin{aligned} d_{\mathcal{H}}(\widehat{v}(\theta), \widehat{v}(\theta')) &= d_{\mathcal{H}}\left(\frac{\mathbf{1}_{R'}/R'}{K(\theta)^\top\widehat{u}(\theta)}, \frac{\mathbf{1}_{R'}/R'}{K(\theta')^\top\widehat{u}(\theta')}\right)\\ &= d_{\mathcal{H}}\left(K(\theta)^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta')\right)\\ &\leq d_{\mathcal{H}}\left(K(\theta)^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta)\right) + d_{\mathcal{H}}\left(K(\theta')^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta')\right)\\ &\leq d_{\mathcal{H}}\left(K(\theta)^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta)\right) + \lambda\left(K(\theta')\right) d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta'))\\ &\leq \frac{2k_u^2\sqrt{R}R' L_K}{k_\ell^3}\|\theta - \theta'\|_2 + \Lambda d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')). \end{aligned} \tag{5.41}$$

Likewise, by invoking in turn (5.31), (5.27), the triangle inequality, (5.28) and both

(5.40) and (5.41), we get

$$\begin{aligned}
d_{\mathcal{H}}(\widehat{u}(\theta),\widehat{u}(\theta')) &= d_{\mathcal{H}}\left(\frac{\mathbf{1}_R/R}{K(\theta)\widehat{v}(\theta)},\frac{\mathbf{1}_R/R}{K(\theta')\widehat{v}(\theta')}\right)\\
&= d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta),K(\theta')\widehat{v}(\theta')\right)\\
&\leq d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta),K(\theta')\widehat{v}(\theta)\right)+d_{\mathcal{H}}\left(K(\theta')\widehat{v}(\theta),K(\theta')\widehat{v}(\theta')\right)\\
&\leq d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta),K(\theta')\widehat{v}(\theta))+\lambda(K(\theta'))d_{\mathcal{H}}(\widehat{v}(\theta),\widehat{v}(\theta'))\\
&\leq d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta),K(\theta')\widehat{v}(\theta))\\
&\quad+\Lambda\left(\frac{2k_u^2\sqrt{R}R'L_K}{k_\ell^3}\|\theta-\theta'\|_2+\Lambda d_{\mathcal{H}}(\widehat{u}(\theta),\widehat{u}(\theta'))\right).
\end{aligned}$$

The above inequality and (5.39) then yield

$$\begin{aligned}
d_{\mathcal{H}}(\widehat{u}(\theta),\widehat{u}(\theta')) &\leq \frac{1}{1-\Lambda^2}\left(d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta),K(\theta')\widehat{v}(\theta))+\Lambda\frac{2k_u^2\sqrt{R}R'L_K}{k_\ell^3}\|\theta-\theta'\|_2\right)\\
&\leq \frac{2k_u^2\sqrt{RR'}L_K}{(1-\Lambda^2)k_\ell^3}(\sqrt{R}+\Lambda\sqrt{R'})\|\theta-\theta'\|_2.
\end{aligned}$$

Therefore, by Lemma 5.1 and (5.33), $\|\widehat{u}(\theta)-\widehat{u}(\theta')\|_2 \leq L_{\hat{u}}\|\theta-\theta'\|_2$, which completes the proof.

- The function $\theta \mapsto \widehat{v}(\theta)$ is Lipschitz continuous on $\operatorname{dom} g_\tau$ with Lipschitz constant

$$L_{\widehat{v}} := \frac{k_u L_K}{k_\ell^3\sqrt{R}}+\frac{k_u\sqrt{R'}L_{\hat{u}}}{k_\ell^2 R^{3/2}}.$$

Proof: Set arbitrarily $\theta,\theta' \in \operatorname{dom} g_\tau$. By (5.31) and (5.34),

$$\begin{aligned}
\|\widehat{v}(\theta)-\widehat{v}(\theta')\|_2 &= \left\|\frac{\mathbf{1}_{R'}/R'}{K(\theta)^\top\widehat{u}(\theta)}-\frac{\mathbf{1}_{R'}/R'}{K(\theta')^\top\widehat{u}(\theta')}\right\|_2\\
&\leq \frac{\|K(\theta)^\top\widehat{u}(\theta)-K(\theta')^\top\widehat{u}(\theta')\|_2}{R'\min_{r'\in[\![R']\!]}\{(K(\theta_1)^\top\widehat{u}(\theta_1))_{r'}\}\min_{r'\in[\![R']\!]}\{(K(\theta')^\top\widehat{u}(\theta'))_{r'}\}}\\
&= \frac{\|K(\theta)^\top\widehat{u}(\theta)-K(\theta')^\top\widehat{u}(\theta')\|_2}{\min_{r'\in[\![R']\!]}\{\widehat{v}_{r'}(\theta)^{-1}\}\min_{r'\in[\![R']\!]}\{\widehat{v}_{r'}(\theta')^{-1}\}}\\
&\leq \frac{1}{k_\ell^2 R^2}\|K(\theta)^\top\widehat{u}(\theta)-K(\theta')^\top\widehat{u}(\theta')\|_2.
\end{aligned}$$

Moreover, using in turn the triangle inequality, (5.22) then the fact that K and $\widehat{u}$ are Lipschitz continuous and bounded on $\operatorname{dom} g_\tau$, we get

$$\begin{aligned}
\|K(\theta)^\top\widehat{u}(\theta)-K(\theta')^\top\widehat{u}(\theta')\|_2 &\leq \|K(\theta)^\top\widehat{u}(\theta)-K(\theta')^\top\widehat{u}(\theta)\|_2+\|K(\theta')^\top\widehat{u}(\theta)-K(\theta')^\top\widehat{u}(\theta')\|_2\\
&\leq \|K(\theta)-K(\theta')\|_F\|\widehat{u}(\theta)\|_2+\|K(\theta')\|_F\|\widehat{u}(\theta)-\widehat{u}(\theta')\|_2\\
&\leq \left(\frac{k_u R^{3/2}L_K}{k_\ell}+\sqrt{RR'}k_u L_{\hat{u}}\right)\|\theta-\theta'\|_2.
\end{aligned}$$

Therefore, $\|\widehat{v}(\theta)-\widehat{v}(\theta')\|_2 \leq L_{\hat{v}}\|\theta-\theta'\|_2$, which completes the proof.

- The function $\hat{P}(\theta)$ is Lipschitz continuous on $\operatorname{dom} g_\tau$.
  Proof: We have proved that $\theta \mapsto \widehat{u}$, $\theta \mapsto K(\theta)$ and $\theta \mapsto \widehat{v}(\theta)$ are bounded and Lipschitz continuous on $\operatorname{dom} g_\tau$. Consequently, so is $\theta \mapsto \hat{P}(\theta) = \operatorname{diag}(\widehat{u}(\theta))K(\theta)\operatorname{diag}(\widehat{v}(\theta))$.

This completes the proof of Lemma 5.2, hence that of the fact that $\theta \mapsto \widehat{P}_\theta$ and $\theta \mapsto \widehat{Q}_\theta$ are Lipschitz continuous on $\operatorname{dom} g_\tau$.

5.7.1.d The gradient of f is Lipschitz continuous

Set arbitrarily $\theta, \theta' \in \operatorname{dom} g_\tau \subset [0,1]^N$. We begin by noting that, by the triangle inequality and (5.22),

$$\begin{aligned}\tfrac{1}{2}\|\nabla f(\theta) - \nabla f(\theta')\|_2 &\leq \|y\|_2 \times \|\widehat{P}_\theta - \widehat{P}_{\theta'}\|_F + \|\theta\|_2 \times \|\widehat{Q}_\theta - \widehat{Q}_{\theta'}\|_F + \|\widehat{Q}_{\theta'}\|_F \times \|\theta - \theta'\|_2 \\ &\leq \|y\|_2 \times \|\widehat{P}_\theta - \widehat{P}_{\theta'}\|_F + \sqrt{N} \times \|\widehat{Q}_\theta - \widehat{Q}_{\theta'}\|_F + \|\theta - \theta'\|_2.\end{aligned}$$

We then readily conclude because we showed in Section 5.7.1.c that $\theta \mapsto \widehat{P}_\theta$ and $\theta \mapsto \widehat{Q}_\theta$ are Lipschitz continuous on $\operatorname{dom} g_\tau$.

5.7.2 The function H_δ satisfies the Kurdyka-Lojasiewicz property

5.7.2.a The Kurdyka-Lojasiewicz property

Let us first recall what is the Kurdyka-Lojasiewicz property. Let $\ell : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous function. For any $-\infty < \eta_1 < \eta_2 \leq +\infty$, the bracket $[\eta_1 < \ell < \eta_2]$ is the set $\{x \in \mathbb{R}^d : \eta_1 < \ell(x) < \eta_2\}$. We refer the reader to (Attouch et al., 2010, Section 2) for elementary facts of nonsmooth analysis, including the definition of $\partial\ell$, the limiting-subdifferential of ℓ (Rockafellar and Wets, 1998).

Definition 5.1 (Kurdyka-Lojasiewicz property, definition 3.1 in Attouch et al. (2010))**.** *The function ℓ is said to have the Kurdyka-Lojasiewicz property at $\bar{x} \in \operatorname{dom} \partial\ell$ if there exists $\eta \in (0, +\infty]$, a neighborhoood U of $\bar{x}$ and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that:*

- *$\varphi(0) = 0$,*
- *φ is C^1 on $(0, \eta)$,*
- *for all $s \in (0, \eta)$, $\varphi'(s) > 0$,*
- *and for all $x \in U \cap [\ell(\bar{x}) < \ell < \ell(\bar{x}) + \eta]$, the Kurdyka-Lojasiewicz inequality holds:*

$$\varphi'(\ell(x) - \ell(\bar{x}))\operatorname{dist}(0, \partial\ell(x)) \geq 1. \tag{5.42}$$

Inequality (5.42) can be interpreted as follows: subject to the reparametrization of f through φ, we deal with a sharp function. To see this, consider the simple case where the finite-valued f is differentiable and $f(\bar{x}) = 0$, so that (5.42) rewrites as $\|\nabla\varphi \circ f(x)\| \geq 1$: the function φ transforms a singular region, characterized by arbitrarily small gradients, into a regular region where the gradients are bounded away from zero. Thus the transformation φ is aptly referred to as a "desingularizing function" for f. For further theoretical and geometrical insights, we refer to (Bolte et al., 2010).

To prove that H_δ satisfies the Kurdyka-Lojasiewicz property, we apply Theorem 4.1 in (Attouch et al., 2010). We state it below for the sake of completeness. The key notions necessary to understand the theorem are succinctly presented after the statement.

Theorem 5.2 (Theorem 4.1 in Attouch et al. (2010)). *Any proper lower semicontinuous function $\ell : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ which is definable in an o-minimal structure $\mathcal{O}$ over $\mathbb{R}$ has the Kurdyka-Lojasiewicz property at each point of* $\operatorname{dom} \partial \ell$. *Moreover the function φ appearing in* (5.42) *is definable in $\mathcal{O}$.*

ON *o*-MINIMAL STRUCTURES. An *o*-minimal structure over $\mathbb{R}$ can be viewed as an axiomatization of the quantitative properties of semialgebric sets. Semialgebric sets are finite unions and intersections of sets of the form $\{x \in \mathbb{R}^d : Q(x) = 0, R(x) < 0\}$ for some polynomial functions $Q, R : \mathbb{R}^d \to \mathbb{R}$. Algebraic sets are finite unions and intersections of sets of the form $\{x \in \mathbb{R}^d : Q(x) = 0\}$ for some polynomial function $Q : \mathbb{R}^d \to \mathbb{R}$.

Formally, a collection $\mathcal{O} = \{\mathcal{O}_n\}_{n \geq 0}$ is a structure over $\mathbb{R}$ if the following conditions are met:

(a) for each $n \geq 0$, $\mathcal{O}_n$ is a collection of subsets of $\mathbb{R}^n$;

(b) for each $n \geq 0$, all algebraic subsets of $\mathbb{R}^n$ are in $\mathcal{O}_n$;

(c) for each $n \geq 0$, $\mathcal{O}_n$ is a Boolean subalgebra, that is, $\emptyset \in \mathcal{O}_n$ and, for every $A, B \in \mathcal{O}_n$, $A \cup B$, $A \cap B$ and $\mathbb{R}^n \setminus A$ belong to $\mathcal{O}_n$;

(d) if $A \in \mathcal{O}_m$ and $B \in \mathcal{O}_n$, then $A \times B \in \mathcal{O}_{m+n}$;

(e) if $p : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is the projection on the first n coordinates and $A \in \mathcal{O}_{n+1}$, then $p(A) \in \mathcal{O}_n$.

It is *o*-minimal if, in addition,

(f) the elements of $\mathcal{O}_1$ are precisely the finite unions of intervals.

The smallest *o*-minimal structure over $\mathbb{R}$ containing the semialgebric sets is denoted $\mathbb{R}_{\text{alg}}$. It is the collection $\{\mathcal{O}_n\}_{n \geq 0}$ where each $\mathcal{O}_n$ is the class of semialgebric sets on $\mathbb{R}^n$ (Benedetti and Risler, 1990; Bochnak et al., 1998).

The smallest structure containing the semialgebraic sets and the graph of the exponential function $\exp : \mathbb{R} \to \mathbb{R}_+^*$ is denoted $\mathbb{R}_{\exp}$. It extends $\mathbb{R}_{\text{alg}}$ and it is *o*-minimal over $\mathbb{R}$ (Wilkie, 1996).

ON DEFINABLE SETS AND DEFINABLE FUNCTIONS. Given an *o*-minimal structure $\mathcal{O} = (\mathcal{O}_n)_{n \geq 0}$ over $\mathbb{R}$, the elements of each $\mathcal{O}_n$ are called the definable subsets of $\mathbb{R}^n$. A function $\varphi : A \to B$ between to definable sets is definable in $\mathcal{O}$ if its graph is definable in $\mathcal{O}$.

For instance, a polynomial function $Q : \mathbb{R}^d \to \mathbb{R}$ is definable in $\mathbb{R}_{\text{alg}}$, hence in $\mathbb{R}_{\exp}$ as well.

We use the following properties (Attouch et al., 2010) (from now on, we write "definable" in lieu of "definable in $\mathcal{O}$"):

(g) if $\varphi : A \to B$ is definable and if $A' \subset A$ is definable, then $\varphi|_{A'}$ is definable;

(h) if φ is definable, then $|\varphi|$ is definable;

(i) finite sums of definable function are definable;

(j) any indicator function $\mathbf{I}\{A\}$ (which equals 0 if the argument falls in A and $+\infty$ otherwise) of a definable set A is definable;

(k) generalized inverse functions of definable functions are definable;

(l) compositions of definable functions are definable;

(m) if ψ and C are definable, then $\mathbb{R}^n \ni x \mapsto \inf_{y\in C} \psi(x,y)$ and $\mathbb{R}^n \ni x \mapsto \sup_{y\in C} \psi(x,y)$ are definable.

5.7.2.b The function H_δ is definable in $\mathbb{R}_{\exp}$

Let us prove now that H_δ is definable in $\mathbb{R}_{\exp}$ – from now on, "definable" means definable in $\mathbb{R}_{\exp}$. We consider the following steps.

- The set $\Pi_{R,R'}$ is semialgebric hence definable.
  Proof: Introduce the sets $A_{r,r'} := \{P \in \mathbb{R}^{R\times R'} : P_{r,r'} \geq 0\}$, $B_r := \{P \in \mathbb{R}^{R\times R'} : \sum_{r'\in[\![R']\!]} P_{r,r'} = \frac{1}{R}\}$ and $C_{r'} := \{P \in \mathbb{R}^{R\times R'} : \sum_{r\in[\![R]\!]} P_{r,r'} = \frac{1}{R'}\}$ (for all $r \in [\![R]\!]$ and $r' \in [\![R']\!]$). Each of them is semialgebraic. Therefore their intersection, which equals $\Pi_{R,R'}$, is semialgebraic too, hence definable.

- Consider $F : \mathbb{R}^N \times \mathbb{R}^{M\times N} \times \mathbb{R}^{N\times N} \to \mathbb{R}$ given by

$$F(\theta,P,Q) := \sum_{m\in[\![M]\!],n\in[\![N]\!]} P_{m,n}\left(d(x_m,x'_n)^2 + (y_m-\theta_n)^2\right) - \tfrac{1}{2}\sum_{m\in[\![M]\!],n\in[\![N]\!]} Q_{n,n'}\left(d(x'_n,x'_{n'})^2 + (\theta_n-\theta_{n'})^2\right) + g_\tau(\theta).$$

  Proof: The function $(\theta,P) \mapsto F(\theta,P,Q) - g_\tau(\theta)$ is definable because it is polynomial. Moreover, g_τ is also definable.

  - *When* $g_\tau(\theta) = \tau\|\theta\|_1 + \mathbf{I}\{\theta \in [0,1]^N\}$: on the one hand, $\theta \mapsto \|\theta\|_1 = \sum_{n\in[\![N]\!]} |\theta_n|$ is definable as a finite sum of definable functions (properties (i) and (h); on the other hand, $\mathbf{I}\{[0,1]^N\}$ is definable because $[0,1]^N$ is definable (property (j)). Therefore, g_τ is definable (property (i)).
  - *When* $g_\tau(\theta) = \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0,1]^N\}$: on the one hand, the set $\{\theta \in \mathbb{R}^N : \|\theta\|_1 \leq \tau\}$ is definable because it can be written as

$$\bigcup_{\varepsilon\in\{\pm1\}^N} \left[\bigcap_{n\in[\![N]\!]} \{\theta \in \mathbb{R}^N : \varepsilon_n\theta_n \geq 0\} \cap \{\theta \in \mathbb{R}^N : \sum_{n\in[\![N]\!]} \varepsilon_n\theta_n - \tau \leq 0\}\right],$$

    which is semialgebraic since it is a finite union and intersection of semialgebraic sets; therefore, $\theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\}$ is definable (property (j)). On the other hand, we already proved that $\mathbf{I}\{[0,1]^N\}$ is definable, hence g_τ is definable (property (i)).

  It follows that F is definable (property (i)). Because the set $\mathbb{R}^N \times \Pi_{M,N} \times \Pi_{N,N}$ is definable, this implies that $F|_{\mathbb{R}^N\times\Pi_{M,N}\times\Pi_{N,N}}$ is definable (property (g)).

- The function $\gamma E : P \mapsto \gamma \times E(P)$ from $\Pi_{R,R'}$ to $\mathbb{R}$ is definable.
  Proof: The function $\log : \mathbb{R}_+^* \to \mathbb{R}$ is definable (property (k)). Consequently, $\varphi : \mathbb{R}_+^* \to \mathbb{R}^2$ given by $\varphi(x) := (\log(x), x)$ is definable because its graph can be written as

$$(\Gamma_{\log} \times \mathbb{R}) \cap \{(x,y,z) \in \mathbb{R}^3 : x - z = 0\}$$

  where the graph $\Gamma_{\log}$ of log is definable and the right-hand-side set is algebraic hence definable, revealing that the graph of φ is definable as the intersection of two definable

sets. Moreover, the polynomial function $Q : \mathbb{R}^2 \to \mathbb{R}$ given by $Q(x, y) := -\gamma x(y - 1)$ is definable. Therefore, $\phi := Q \circ \varphi : \mathbb{R}_+^* \to \mathbb{R}$, so that $\phi(x) = -\gamma x(\log(x) - 1)$, is definable (property (l)). Setting $\phi(0) := 0$ extends ϕ by continuity and yields a definable function $\phi : \mathbb{R}_+ \to \mathbb{R}$. It follows that $\gamma\mathcal{E} : (\mathbb{R}_+)^{R \times R'} \to \mathbb{R}$ given by $\gamma\mathcal{E}(P) := \sum_{r \in [\![R]\!], r' \in [\![R']\!]} \phi(P_{r,r'})$ is definable (property (i)), hence $\gamma E := \gamma\mathcal{E}|_{\Pi_{R,R'}}$ is definable too (property (g)).

- The function $(f + g_\tau) : \mathbb{R}^N \to \mathbb{R}$ is definable.
  Proof: This is a straightforward consequence of the fact that, for all $\theta \in \mathbb{R}^N$,

  $$(f + g_\tau)(\theta) := \min_{P \in \Pi_{M,N}} \max_{Q \in \Pi_{N,N}} \left\{ F|_{\mathbb{R}^N \times \Pi_{M,N} \times \Pi_{N,N}} + \gamma E(P) - \tfrac{1}{2}\gamma E(Q) \right\},$$

  where the sets $\Pi_{M,N}$ and $\Pi_{N,N}$ are definable (property (m)).

- The function H_δ is definable.
  Proof: Recall that $H_\delta : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is given by $H_\delta(\theta, \theta') := f(\theta') + g_\tau(\theta') + \delta\|\theta - \theta'\|_2^2$. The function $(\theta, \theta') \mapsto f(\theta') + g_\tau(\theta')$ between $\mathbb{R}^N \times \mathbb{R}^N$ and $\mathbb{R}$ is definable because its graph

  $$\{(\theta, \theta', f(\theta') + g_\tau(\theta')) : (\theta, \theta') \in \mathbb{R}^N \times \mathbb{R}^N\} = \mathbb{R}^N \times \Gamma_{f+g_\tau},$$

  where Γ_{f+g_τ} is the graph of $(f + g_\tau)$, is definable as the product of two definable sets. Moreover, the function $(\theta, \theta') \mapsto \delta\|\theta - \theta'\|_2^2$ between $\mathbb{R}^N \times \mathbb{R}^N$ and $\mathbb{R}$ is polynomial, hence definable. Therefore, H_δ is definable (property (i)).

5.7.2.c The function H_δ is proper and lower semicontinuous, hence satisfies the Kurdyka-Lojasiewicz property on the domain of ∂H_δ

The function H_δ never takes on the value $-\infty$ and $H_\delta(0)$ is finite, so H_δ is proper. Moreover, f is differentiable (see Section 5.7.1), g_τ is lower semicontinuous because it is either continuous (when $g_\tau(\cdot) = \tau\|\cdot\|_1$) or lower semicontinuous (when g_τ is the characteristic function of the closed $\|\cdot\|_1$-ball centered at 0 and with radius τ), and $(\theta, \theta') \mapsto \delta\|\theta - \theta'\|_2^2$ is continuous. Therefore, H_δ is proper and lower semicontinuous. By Theorem 5.2, H_δ satisfies the Kurdyka-Lojasiewicz property on the domain of ∂H_δ.

6

Conclusion and perspectives for further work

Optimal transport is a powerful tool to capture the similarity between two datasets and has found application in many diverse areas of machine learning including domain adaptation (Courty et al., 2017), generative modeling (Genevay et al., 2018). The contributions of this thesis can be organized in two main issues. In the first project, the contribution is to better understand micro-RNA (miRNA) regulation in the striatum of Huntington's disease (HD) model mice. In the second project, the contribution is to address a problem that involves predicting which cities will submit a request for the government declaration of natural disaster for a drought event. This is a sub-problem of forecasting the cost of drought events in France.

6.1 Conclusion

About Huntington's disease. We have developed two co-clustering algorithms (WTOT-SCC1 and WTOT-SCC2) and a matching algorithm (WTOT-matching) for the purpose of identifying groups of mRNAs and miRNAs that interact. The algorithms apply in any situation where it is of interest to cluster or match the elements of two data sets based on a parametric model Θ expressing what it means to interact for any two pairs of elements from the two data sets. The algorithms rely on optimal transport, spectral co-clustering and a matching procedure. In light of (Alvarez-Melis, 2019, Section 1.3, page 25), problem-specific knowledge is injected onto two of the three main components of the transportation problem: the representation spaces (via Θ) and the marginal constraints, leaving aside the cost function.

During the first stage, an optimal optimal transport plan P and mapping in Θ are learned from the data using the Sinkhorn-Knopp algorithm and a mini-batch gradient descent. During the second stage, P is exploited to derive either co-clusters or several sets of matched elements.

As in (Mégret et al., 2020), the motivation of our study is to shed light on the interaction

between mRNAs and miRNAs based on data collected in the striatum of HD model knock-in mice (Langfelder et al., 2016, 2018). Each data point takes the form of a multi-dimensional profile. The strong biological hypothesis is that if a miRNA induces the degradation of a target mRNA or blocks its translation into proteins, or both, then the profile of the former should be similar to minus the profile of the latter — this particular form of affine relationship drives the formulation of a loosened hypothesis and definition of model Θ. The fact that the algorithm learns from the data a best element in Θ provides more flexibility than in (Mégret et al., 2020).

The simulation study reveals on the one hand that WTOT-SCC2 works overall better than WTOT-SCC1, but that the co-clustering task can be very challenging in the presence of many irrelevant data points (data points that do not interact). On the other hand, it shows that the performances of WTOT-matching are satisfying.

An illustration on real data is given. The results are biologically relevant and illustrate how our algorithm strikes a good balance between two moderately and highly selective, competing algorithms. Our findings lead to reconsidering the formerly-expressed view on a limited role of miRNA regulation in the striatum of HD mice on a systems level (Mégret et al., 2020).

ABOUT THE ANTICIPATION OF THE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT. This study is motivated by the challenging task of forecasting which cities in France will submit a request for the government declaration of natural disaster for a drought event. While the problem can be addressed as a classification task using standard classification algorithms, we take a slightly different perspective and introduce an alternative procedure based on optimal transport theory (Peyré and Cuturi, 2019) and iPiano (Ochs et al., 2015), an inertial proximal algorithm for nonconvex optimization.

We build the OT-procedure upon two core ideas. Firstly, we aim to predict whether a city will submit a request by making an interpretable comparison of the city's covariates with those of other cities whose submission status may be already known. Secondly, recognizing that relatively few cities will submit requests, we seek to control the sparsity of our predictions and encourage 0-predictions, indicating cases where we predict that a city will not submit a request. Additionally, we develop a hybrid procedure that synergistically combines and utilizes both types of predictions, derived from classification algorithms and the OT-procedure.

We develop and program an algorithm that hinges on iPiano and a mini-batch procedure to cope with large data sets, see Algorithm 2. The convergence of the iPiano algorithm is established, using the notion of o-minimal structures from the field of tame geometry (Wilkie, 1996) to prove that a critical function related to (5.4) satisfies the Kurdyka-Lojasiewicz property (Attouch et al., 2010). Coded in `python/pytorch`, relying on the `GeomLoss` package (Feydy et al., 2019b) for its fast implementation of the Sinkhorn algorithm, the program will soon be made available.

We conduct a simulation study to illustrate the use of the OT-procedure and of the hybrid procedure in a simple context, laying the groundwork for the real-world application. The latter poses greater challenges than the former. Tangibly, these challenges arise because $\mathcal{X} \subset \mathbb{R}^d$ is a relatively high-dimensional space ($d = 67$) and because the sample sizes are large. Intangibly, the intricacies lie in the mechanisms that determine whether a request is submitted or not.

We rely on the HYPERBAND algorithm (Li et al., 2018) and on a simple grid search to define a relevant cost function and fine-tune the hyperparameters of Algorithm 2. An analysis of the cost function reveals that the more relevant groups of covariates are, in

decreasing order of importance, the covariates related to a city's exposure to drought events, its request history, its description and its vicinity.

For a total of 22 weeks spanning from the 49th week of 2021 (December 6th to 12th) to the 26th week of 2022 (June 28th to July 4th), intermittently, we predict whether or not the cities that have not yet submitted a request for the year 2021 will eventually do so. We employ the best of five standard classification algorithms, the OT-procedure and the hybrid procedure to make these predictions. Overall, the hybrid procedure yields enhanced forecasting accuracy, in particular while focusing on the estimation of the eventual number of requests. A simple analysis of the covariate's importance sheds light on the strength of association between each covariate and the predictions. It suggests that most covariates play an effective role in the predictions.

6.2 Perspectives for future work

ABOUT HUNTINGTON'S DISEASE. There are several directions for future work. First, we will develop a similar study to better understand miRNA regulation in the cortex of HD model mice. Second, we will evaluate the performances of our algorithms by simulation studies based on a simulation scheme learned from the real data so as to better mimic their law. Third, we will put our algorithms into the general context of co-clustering and matching of datasets and carry out more benchmark tests and comparisons.

ABOUT THE ANTICIPATION OF THE DECLARATION OF NATURAL DISASTER FOR A DROUGHT EVENT. We list potential avenues for future research. Firstly, the procedures discussed in the study may benefit from the use of an enhanced version of the city-level SWI. By considering the variation in the nature of the soil across different regions of France, this refined version could contribute to making more accurate predictions. Secondly, to make the hybrid procedure more acceptable to the experts at CCR, it would be interesting to complement the analysis of the covariates' importance. This additional analysis could offer further insights and explanations regarding the predictions. Thirdly, the current predictions obtained from the investigated procedures lack a measure of confidence. Developping a methodology to address this issue would be highly valuable. In conclusion, we acknowledge that the last two questions raised are very challenging, notably due to the complex interdependence within the data set.

Published Articles and Preprints

T. T. Y. Nguyen, W. Harchaoui, L. Mégret, C. Mendoza, O. Bouaziz, C. Neri, and A. Chambaz. Optimal transport-based machine learning to match specific patterns: application to the detection of molecular regulation patterns in omics data. Mar. 2023. URL https://hal.science/hal-03293786

L. Mégret, C. Mendoza, M. Arrieta Lobo, E. Brouillet, T. T. Y. Nguyen, O. Bouaziz, A. Chambaz, and C. Néri. Precision machine learning to understand micro-RNA regulation in neurodegenerative diseases. *Frontiers in Molecular Neuroscience*, 15, 2022. URL https://www.frontiersin.org/articles/10.3389/fnmol.2022.914830

Bibliography

M. Achour, S. Le Gras, C. Keime, F. Parmentier, F.-X. Lejeune, A.-L. Boutillier, C. Neri, I. Davidson, and K. Merienne. Neuronal identity genes regulated by super-enhancers are preferentially down-regulated in the striatum of Huntington's disease mice. *Human Molecular Genetics*, 24(12):3481–3496, 2015.

M. Ailem, F. Role, and M. Nadif. Graph modularity maximization as an effective method for co-clustering text data. *Knowledge-Based Systems*, 109:160–173, 2016.

D. Alvarez-Melis. *Optimal Transport in Structured Domains: Algorithms and Applications*. PhD thesis, Massachusetts Institute of Technology, 2019.

S. E. Andrew, Y. P. Goldberg, B. Kremer, H. Telenius, J. Theilmann, S. Adam, E. Starr, F. Squitieri, B. Lin, M. A. Kalchman, G. R. K., and M. R. Hayden. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature Genetics*, 4:398–403, 1993.

H. Assadollahi. *The impact of climatic events and drought on the shrinkage and swelling phenomenon of clayey soils interacting with constructions*. PhD thesis, Université de Strasbourg, June 2019. URL https://theses.hal.science/tel-02331567.

H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Lojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.

A. Baddeley and R. Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer Distance as a Solution to Biased Wasserstein Gradients, 2017.

B. Benayoun, E. Pollina, P. Singh, S. Mahmoudi, I. Harel, K. Casey, B. Dulken, A. Kundaje, and A. Brunet. Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. *Genome Research*, 29(4): 697–709, 2019.

R. Benedetti and J.-J. Risler. *Real algebraic and semi-algebraic sets*. Actualités Mathématiques. [Current Mathematical Topics]. Hermann, Paris, 1990.

I. Berkes and W. Philipp. An almost sure invariance principle for the empirical distribution function of mixing random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 41:115–137, 1977. doi: 10.1007/BF00538416.

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, 1999.

D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, 11:R90, 2010.

G. Birkhoff. Extensions of Jentzsch's theorem. *Trans. Amer. Math. Soc.*, 85, 1957.

J. Bochnak, M. Coste, and M.-F. Roy. *Real algebraic geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1998.

J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Lojasiewicz inequalities: Subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362:3319–3363, 2010.

V. Brault, C. Keribin, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25:1201–1216, 2014.

L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

CCR. Les catastrophes naturelles en France: bilan 1982-2020. Technical report, Caisse Centrale de Réassurance, 2021. URL https://side.developpement-durable.gouv.fr/ACCIDR/doc/SYRACUSE/795441.

CCR. Rapport d'activité 2021. Technical report, Caisse Centrale de Réassurance, 2022. URL https://www.ccr.fr/documents/35794/35839/CCR+RA+2021+web+all+24032022.pdf/84e4c7da-34b5-22e0-e048-06a0836b7392?t=1648135815072.

J. H. J. Cha. Transcriptional signatures in huntington's disease. *Progress in Neurobiology*, 83(4):228–248, 2007.

A. Charpentier, L. Barry, and M. R. James. Insurance against natural catastrophes: balancing actuarial fairness and social solidarity. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 47(1):50–78, January 2022a. URL https://ideas.repec.org/a/pal/gpprii/v47y2022i1d10.1057_s41288-021-00233-7.html.

A. Charpentier, M. James, and H. Ali. Predicting drought and subsidence risks in France. *Nat. Hazards Earth Syst. Sci.*, 22:2401–2418, 2022b. doi: 10.5194/nhess-22-2401-2022.

P. Chatelain and S. Loisel. Subsidence and household insurances in France: geolocated data and insurability. working paper or preprint, 2021. URL https://hal.science/hal-03791154.

E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):1–14, 2013.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1853–1865, 2017.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2292–2300, USA, 2013a. Curran Associates Inc.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013b.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693. PMLR, 22–24 Jun 2014.

J. M. Danskin. The theory of max − min, with applications. *SIAM J. Appl. Math.*, 14, 1966.

I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 269—274, New York, NY, USA, 2001. Association for Computing Machinery.

J. Ding, X. Li, and H. Hu. TarPmiR: a new approach for microRNA target site prediction. *BMC Bioinformatics*, 32:2768–2775, 2016.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 272–279, New York, NY, USA, 2008. Association for Computing Machinery.

G. Ecoto and A. Chambaz. Forecasting the cost of drought events in France by Super Learning. Technical report, submitted, Dec. 2022. URL https://hal.science/hal-03701743.

G. Ecoto, A. F. Bibaut, and A. Chambaz. One-step ahead sequential Super Learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters. Technical report, submitted, July 2021. URL https://hal.science/hal-03300559.

K. Fatras, Y. Zine, R. Flamary, R. Gribonval, and N. Courty. Learning with minibatch Wasserstein : asymptotic and gradient properties. In *The 23nd International Conference on Artificial Intelligence and Statistics*, volume volume 108 of *PMLR*, Palermo, Italy, 2020.

J. Feydy, T. Séjourné, F.-X. Vialard, S. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 2019a.

J. Feydy, T. Séjourné, F.-X. Vialard, S. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 16–18 Apr 2019b.

France Assureurs. Le risque sécheresse et son impact sur les habitations. 2022. URL https://www.franceassureurs.fr/wp-content/uploads/le-risque-secheresse-et-son-impact-sur-les-habitations-15-novembre-2022-web.pdf.

J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

G. Govaert and M. Nadif. Model-based co-clustering for continuous data. In *Machine Learning and Applications, Fourth International Conference on*, pages 175–180, Los Alamitos, CA, USA, dec 2010. IEEE Computer Society.

Y. Heng, J. Shi, and L. Carlone. Teaser: Fast and certifiable point cloud registration, 2020. arXiv:2001.07715.

A. Heranval, O. Lopez, and M. Thomas. Application of machine learning methods to predict drought cost in france. *European Actuarial Journal*, pages 1–23, 2022.

M. Hollander and D. A. Wolfe. *Nonparametric statistical methods*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons, Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

Y. Huang, Q. Zou, H. Song, F. Song, L. Wang, G. Zhang, and X.-J. Shen. A study of miRNA targets prediction and experimental validation. *Protein Cell*, 1:979–86, 11 2010.

A. E. Hubbard, S. Kherad-Pajouh, and M. J. van der Laan. Statistical inference for data adaptive target parameters. *Int. J. Biostat.*, 12(1):3–19, 2016.

IGN. GEOFLA. Technical report, Institut National de l'Information Géographique et Forestière, 2018. URL https://geoservices.ign.fr/sites/default/files/2021-07/DC_GEOFLA_2-2.pdf. version 2.2.

Insee. Recensement de la population 1999: tableaux analyses. Technical report, Institut national de la statistique et des études économiques, 2000.

L. Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk SSSR*, pages 227–229, 1942.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, M. C. D., G. G. W., and M. A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016.

C. Laclau, I. Redko, B. Matei, Y. Bennani, and V. Brault. Co-clustering through Optimal Transport. In *34th International Conference on Machine Learning*, volume 70, pages 1955–1964, Sydney, Australia, Aug. 2017.

P. Langfelder, J. Cantle, D. Chatzopoulou, N. Wang, F. Gao, I. Al-Ramahi, X. Lu, E. Ramos, K. Merz, Y. Zhao, S. Deverasetty, A. Tebbe, C. Schaab, D. Lavery, D. Howland, S. Kwak, J. Botas, J. Aaronson, J. Rosinski, and X. Yang. Integrated genomics and proteomics define Huntingtin CAGlength–dependent networks in mice. *Nature Neuroscience*, 19: 622–633, 02 2016.

P. Langfelder, F. Gao, N. Wang, D. Howland, S. Kwak, T. Vogt, J. Aaronson, J. Rosinski, G. Coppola, S. Horvath, and X. Yang. MicroRNA signatures of endogenous Huntingtin CAG repeat expansion in mice. *PloS One*, 13(1), 2018.

F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(4):853–877, 2015.

R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1): 15–20, 2005.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

T. Liu, A. W. Moore, and A. Gray. New algorithms for efficient high-dimensional nonparametric classification. *Journal of Machine Learning Research*, 7(41):1135–1158, 2006.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

I. Logar and J. C. J. M. van den Bergh. Methods for assessment of the costs of droughts. Technical report, Institute of environmental science and technology, Universitat Autònoma de Barcelona, 2011. WP5 final report.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, and et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell*, 72(6):971–983, 1993.

S. Maniatis, T. Äijö, S. Vickovic, C. Braine, K. Kang, A. Mollbrink, D. Fagegaltier, Ž. Andrusivová, S. Saarenpää, G. Saiz-Castro, M. Cuevas, A. Watters, J. Lundeberg, R. Bonneau, and H. Phatnani. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, 364(6435):89–93, 2019.

L. Mégret, S. Sasidharan Nair, J. Dancourt, J. Aaronson, J. Rosinski, and C. Neri. Combining feature selection and shape analysis uncovers precise rules for miRNA regulation in Huntington's disease mice. *BMC Bioinformatics*, 21(1):75, 2020.

MI. Procédure de reconnaissance de l'état de catastrophe naturelle - révision des critères permettant de caractériser l'intensité des épisodes de sécheresses-réhydrations des sols a l'origine des mouvement de terrains différentiels. Technical report, Ministère de l'intérieur, 2019. URL https://www.legifrance.gouv.fr/download/pdf/circ?id=44648. NOR: INTE1911312C.

G. Monge. *Mémoire sur la théorie des déblais et des remblais*. Histoire de l'Académie Royale des Sciences, 1781.

MTES. Le retrait-gonflement des argiles: comment prévenir les désordres dans l'habitat individuel. Technical report, Ministère de la transition écologique et solidaire, 2016. URL https://www.ecologie.gouv.fr/sites/default/files/dppr_secheresse_v5tbd.pdf.

L. Mégret, C. Mendoza, M. Arrieta Lobo, E. Brouillet, T. T. Y. Nguyen, O. Bouaziz, A. Chambaz, and C. Néri. Precision machine learning to understand micro-RNA regulation in neurodegenerative diseases. *Frontiers in Molecular Neuroscience*, 15, 2022. URL https://www.frontiersin.org/articles/10.3389/fnmol.2022.914830.

P. V. Nazarov and S. Kreis. Integrative approaches for analysis of mRNA and microRNA high-throughput data. *Computational and Structural Biotechnology Journal*, 19:1154–1162, 2021.

A. Neueder and G. P. Bates. A common gene expression signature in huntington's disease patient brain regions. 7:60, 2014.

T. T. Y. Nguyen, W. Harchaoui, L. Mégret, C. Mendoza, O. Bouaziz, C. Neri, and A. Chambaz. Optimal transport-based machine learning to match specific patterns: application to the detection of molecular regulation patterns in omics data. Mar. 2023. URL https://hal.science/hal-03293786.

P. Ochs, T. Brox, and T. Pock. iPiano: inertial proximal algorithm for strongly convex optimization. *J. Math. Imaging Vision*, 53(2):171–181, 2015.

I. G. Olmo, R. P. Olmo, A. N. A. Gonçalves, R. G. W. Pires, J. T. Marques, and F. M. Ribeiro. High-throughput sequencing of BACHD mice reveals upregulation of neuroprotective miRNAs at the pre-symptomatic stage of Huntington's disease. *ASN Neuro*, 13: 17590914211009857, 2021.

A. E. Pasquinelli. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature reviews. Genetics*, 13(4):271—282, March 2012.

O. Pele and M. Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467, 2009. doi: 10.1109/ICCV.2009.5459199.

S. Petry, R. Keraudren, B. Nateghi, A. Loiselle, K. Pircs, J. Jakobsson, C. Sephton, M. Langlois, I. St-Amour, and S. S. Hébert. Widespread alterations in microRNA biogenesis in human Huntington's disease putamen. *Acta Neuropathologica Communications*, 10(1): 1–11, 2022.

G. Peyré and M. Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends in Machine Learning Series. Now Publishers, 2019.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *ICML 2016*, Proc. 33rd International Conference on Machine Learning, New-York, United States, June 2016.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

E. Polley, E. LeDell, C. Kennedy, and M. J. van der Laan. *SuperLearner: Super Learner Prediction*, 2021. URL https://CRAN.R-project.org/package=SuperLearner. R package version 2.0-28.

E. C. Polley, S. Rose, and M. J. van der Laan. Super learning. In *Targeted learning*, Springer Ser. Statist., pages 43–66. Springer, New York, 2011.

B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

Z. J. Rutnam, T. N. Wight, and B. B. Yang. miRNAs regulate expression and function of extracellular matrix molecules. *Matrix Biology*, 32(2):74–85, 2013.

B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019. doi: 10.1137/16M1106018. URL https://doi.org/10.1137/16M1106018.

J. Spinoni, G. Naumann, J. Vogt, and P. Barbosa. European drought climatologies and trends based on a multi-indicator approach. *Global and Planetary Change*, 127:50–57, 2015. URL https://www.sciencedirect.com/science/article/pii/S0921818115000284.

J. Spinoni, G. Naumann, and J. V. Vogt. Pan-European seasonal trends and recent changes of drought frequency and severity. *Global and Planetary Change*, 148:113–130, 2017. URL https://www.sciencedirect.com/science/article/pii/S0921818116301801.

A. L. Teixeira, F. Dias, M. Gomes, M. Fernandes, and R. Medeiros. Circulating biomarkers in renal cell carcinoma: the link between microRNAs and extracellular vesicles, where are we now? *Journal of Kidney Cancer and VHL*, 1:84–98, 12 2014.

T. Therneau and B. Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. URL https://CRAN.R-project.org/package=rpart. R package version 4.1-15.

M. J. van der Laan. Statistical inference for variable importance. *Int. J. Biostat.*, 2:Art. 2, 33, 2006.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6:Art. 25, 23, 2007.

C. Villani and A. M. Society. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

F. O. Walker. Huntington's disease. *The Lancet*, 369(9557):218–228, 2007.

A. J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *J. Amer. Math. Soc*, 9: 1051–1094, 1996.

B. D. Williamson, P. B. Gilbert, M. Carone, and N. Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

Z. Xie, A. Bailey, M. Kuleshov, D. Clarke, J. Evangelista, S. Jenkins, and A. Lachmann. Gene set knowledge discovery with Enrichr. *Current Protocols*, 1(3):e90, 2021.

J. Zhao, H. Wang, L. Dong, S. Sun, and L. Li. miRNA-20b inhibits cerebral ischemia-induced inflammation through targeting NLRP3. *Int. J. Mol. Med.*, 43(3):1167–1178, 2019.