



HAL
open science

Quelques contributions autour des modèles de mélanges et pour l'analyse de données transcriptomiques

Cathy Maugis Rabusseau

► **To cite this version:**

Cathy Maugis Rabusseau. Quelques contributions autour des modèles de mélanges et pour l'analyse de données transcriptomiques. Mathématiques [math]. Université Paul Sabatier, 2022. tel-03898079

HAL Id: tel-03898079

<https://theses.hal.science/tel-03898079>

Submitted on 14 Dec 2022

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

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

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Document de synthèse

présenté par

Cathy MAUGIS-RABUSSEAU

en vue de l'obtention de

l'Habilitation à Diriger des Recherches

Spécialité : Mathématiques Appliquées

Quelques contributions autour des modèles de mélanges et pour l'analyse de données transcriptomiques

Soutenu le 12 décembre 2022

devant le jury composée de :

Christophe BIERNACKI	Professeur (Université Lille 1)	Président du jury
Pierre LATOUCHE	Professeur (Université de Clermont Auvergne)	Rapporteur
Catherine MATIAS	Directrice de recherche (CNRS)	Rapporteuse
Pierre NEUVIAL	Directeur de recherche (CNRS)	Parrain
Franck PICARD	Directeur de recherche (CNRS)	Rapporteur
Nathalie VIALANEIX	Directrice de Recherche (INRAE)	Examinatrice

Remerciements

Je remercie tout d'abord Pierre Latouche, Catherine Matias et Franck Picard d'avoir accepté de rapporter mon manuscrit d'Habilitation à Diriger des Recherches. Merci d'avoir consacré de votre temps pour expertiser mes travaux, pour votre lecture attentive et pour vos retours. Je remercie également Christophe Biernacki et Nathalie Vialaneix de m'avoir fait l'honneur de participer à ce jury, et Pierre Neuvial d'avoir parrainé mon HDR.

Je remercie chaleureusement mes collaborateurs.trice.s pour les moments scientifiques (et pas seulement) que nous avons pu partager. Je remercie en particulier Béatrice de m'avoir accueilli à l'INSA et pour sa bienveillance; le super coureur Clément pour les très bons moments partagés durant nos années communes à l'INSA; mon américaine préférée Andréa avec qui j'ai toujours plaisir d'échanger et pour ses défis en R/Rmarkdown/Shiny (promis, j'ai bien progressé!). Merci également à Bertrand, Sébastien, Yohann, Antoine, François, Pierre, Juan, Marie-Laure, Gilles... Merci aux biologistes avec qui j'ai eu le plaisir d'interagir dans le cadre de projets interdisciplinaires. Merci également aux étudiant.e.s, stagiaires et doctorants, que j'ai pris plaisir à encadrer.

Je remercie mes collègues de l'IMT pour tous nos moments partagés: nos discussions scientifiques ou pas, nos nombreux cafés et déjeuners, ... Et je décerne une palme du mérite à mes collègues du GMM qui me supportent au quotidien! Merci à Jean-Yves et Pascal de m'avoir fait/ me faire confiance à les seconder à la direction du GMM; à notre super assistante Sandrine pour sa gentillesse, son écoute et son engagement pour le département; à Anthony pour me prêter une oreille attentive quand la coupe est trop pleine. Bien sûr, je ne peux pas oublier de remercier mon co-bureau Frédéric qui depuis plus de 10 ans fait du bureau GMM 116 le plus animé du département (je n'ai toujours pas son don des déclarations enflammées). Je remercie également Pierre pour sa gentillesse, son soutien, son humour et de me permettre de me cacher de temps en temps au bureau 208.

Enfin merci à mon mari Flavien et mes enfants Pauline et Baptiste, pour maintenir ma charge mentale élevée, me faire pédaler pour entretenir ma forme, pour partager les petits plaisirs de la vie et pour tout leur amour.

Table des matières

CHAPITRE 1. Présentation générale	1
--	----------

I Quelques contributions théoriques autour des modèles de mélanges	5
---	----------

CHAPITRE 2. Détection de mélanges à deux composantes	7
---	----------

2.1. Contexte	8
2.2. Détection de mélanges unidimensionnels à deux composantes	10
2.2.1. Procédure de test fondée sur les statistiques d'ordre	10
2.2.2. Etude dans le régime dense	11
2.2.2.1. Borne inférieure non-asymptotique dans le cas gaussien	12
2.2.2.2. Comportement du test $\Psi_{\alpha}^{\text{ord}}$ dans le cas gaussien	12
2.2.2.3. Procédure de test fondée sur la variance	13
2.2.3. Etude dans le régime sparse	13
2.2.3.1. Cas des mélanges gaussiens	14
2.2.3.2. Cas des mélanges de lois de Laplace	14
2.2.3.3. Test sur la variance dans le cas sparse	15
2.3. Détection de mélanges gaussiens multidimensionnels à deux composantes	15
2.3.1. Frontière de détection pour la norme ℓ_2	15
2.3.1.1. Borne inférieure pour la norme ℓ_2 dans le régime dense	16
2.3.1.2. Procédure de test basée sur la moyenne empirique	16
2.3.1.3. Procédure de test basée sur le max des normes ℓ_2	17
2.3.1.4. Procédure de test basée sur les sous-échantillons	18
2.3.2. Frontière de détection pour la norme ℓ_{∞}	19
2.3.2.1. Borne inférieure pour la norme ℓ_{∞}	19
2.3.2.2. Procédures de test pour la norme ℓ_{∞}	20
2.4. Conclusions et perspectives	21

CHAPITRE 3. Estimation d'un mélange de contamination	23
---	-----------

3.1. Contexte et stratégie d'estimation	24
3.1.1. Cadre d'étude	24
3.1.2. Stratégie d'estimation pour la norme \mathbb{L}^2	25
3.1.3. Hypothèses sur ϕ	26
3.2. Vitesses d'estimation des paramètres (ε^*, μ^*)	26
3.3. Bornes inférieures	28
3.4. Lien entre Wasserstein et $\ \cdot\ _2$	29

CHAPITRE 4. Estimation adaptative de densités par mélanges gaussiens **31**

4.1.	Contexte	32
4.2.	Les classes de densités $\mathcal{H}(\beta, \mathcal{P})$	34
4.3.	Résultat d'approximation	35
4.4.	Borne inférieure	36
4.5.	Estimation adaptative de densités	37
4.6.	Conclusion	38

CHAPITRE 5. Estimation de mélanges par super-résolution **39**

5.1.	Contexte	40
5.2.	Hypothèses et premiers résultats	41
5.2.1.	Hypothèses et notations	41
5.2.2.	Estimation exacte de G^* à partir de f^*	42
5.3.	Estimation via le Beurling-LASSO (BLASSO)	43
5.3.1.	Approche par noyau	43
5.3.2.	Estimation par programme convexe	45
5.4.	Performances statistiques de \hat{G}_n	46
5.4.1.	Construction du certificat dual	46
5.4.2.	Divergence de Bregman $D_{\mathcal{P}_m}(\hat{G}_n, G^*)$	47
5.4.3.	Détection des régions proche et éloignée	48
5.4.4.	Stabilité du support pour n assez grand	49
5.5.	Application à quelques mélanges usuels	50
5.5.1.	Noyau sinus cardinal et coupure des fréquences	50
5.5.2.	Mélanges de densités "super-smooth" dont le cas gaussien	51
5.5.3.	Mélanges de densités "ordinary smooth"	52
5.6.	Conclusion et perspectives	53

II Contributions pour l'analyse de données transcriptomiques **55**

CHAPITRE 6. Clustering de gènes co-exprimés à partir de données bulk RNA-seq **57**

6.1.	Contexte et objectif	58
6.1.1.	Principe des données bulk RNA-seq	58
6.1.2.	Classes de gènes co-exprimés	58
6.1.3.	La classification par modèles de mélange	59
6.2.	Classification à partir des comptages	61
6.2.1.	Modèles de mélange de Poisson	61
6.2.2.	Illustration	63
6.2.3.	Discussion	64
6.3.	Classification à partir des profils transformés	64
6.3.1.	Mélanges gaussiens sur profils transformés	64
6.3.1.1.	Mélanges gaussiens	66
6.3.1.2.	Transformation des profils	66
6.3.1.3.	Adaptation des critères de sélection de modèle	67
6.3.1.4.	Illustration	67
6.3.2.	Classification par K -means de profils logCLR-transformés	68
6.3.2.1.	Motivations	68
6.3.2.2.	Procédure de classification	70
6.3.2.3.	Illustration	72
6.4.	Conclusions et perspectives	74

CHAPITRE 7. Quelques challenges pour l'analyse de données single-cell RNA-seq	75
7.1. Contexte	75
7.2. Classification non supervisée de cellules	77
7.3. Détection des gènes marqueurs	80
7.4. Analyse différentielle	83

Présentation générale

Parcours personnel

Suite à mes études à l'Université Paris-Sud 11 à Orsay, j'ai poursuivi en thèse sous la direction de Gilles Celeux et Marie-Laure Martin-Magniette. J'ai soutenu ma thèse en novembre 2008 intitulée "Variable selection for model-based clustering. Application for transcriptome data analysis". J'ai rejoint ensuite l'INSA Toulouse et l'Institut de Mathématiques de Toulouse en septembre 2009 en qualité de Maître de Conférences.

En résumant dans les grandes lignes, mes travaux de recherche sont issus de différentes collaborations et se partagent entre des contributions en statistiques mathématiques autour des modèles de mélanges ; et des contributions interdisciplinaires autour de l'analyse de données biologiques, sans frontière stricte entre les deux. Durant ma thèse, je me suis intéressée à la problématique de la classification non supervisée de gènes à partir de données d'expression issues de puces à ADN. Des développements méthodologiques autour de la classification par mélanges gaussiens avec sélection de variables ont été proposés ([MCM09a], [MCM09b], [MCM11], [M09], [MMP12], [CMMR14]). Ce choix de modèles a conduit également à l'étude de propriétés théoriques des mélanges ([MM11a], [MM11b], [BMM12]). Après ma thèse, plusieurs collaborations ont abouti à des contributions en statistique mathématique dans des contextes variés (tests, estimation de densité et de paramètres) autour des modèles de mélange ([MM13], [LMM16], [LMM18], [GKMM20], [DGMM21]) et d'autres autour de la problématique de la sélection de variables ([CMS9], [MM12]). Du côté de l'analyse de données biologiques (principalement transcriptomiques), des collaborations ont mené à des développements méthodologiques pour la classification de gènes co-exprimés afin de s'adapter à l'évolution des technologies, avec le passage des puces à ADN au séquençage à haut-débit ([RMMC15], [PMM16], [RM18], [R18], [GMR19], [GMR20], [A21]), et d'autres sont en cours autour de l'analyse de données de séquençage en cellule unique.

De mars 2014 à août 2018, j'ai coordonné l'ANR-JCJC MixStatSeq pour "Mixture-based procedures for statistical analysis of RNA-seq data" dont une majorité des travaux résultants sont présentés dans ce manuscrit. J'ai également localement coordonné des projets interdisciplinaires : le BQR-2012 "RNA-Seq et Stat" et le projet Toulouse Tech Interlabs 2018 (TTIL) intitulé "Single-Cell". Je suis actuellement membre du projet DDisc¹ (Double-Dipping in single-cell RNAseq) porté par Pierre Neuvial et le projet interdisciplinaire INSA DEFIANT

¹<https://www.math.univ-toulouse.fr/~pneuvial/ddisc>

(Design of effective nanoparticle-based antimicrobials) porté par Juan Cortés. Dans chacun de ces projets, je co-encadre une thèse débutée en octobre 2021.

La liste de mes travaux de recherche est disponible en page 85 et la liste des étudiant-e-s (co)-encadré-e-s en stage de Master, doctorat ou post-doctorat est donnée en page 89.

Organisation du manuscrit

Mes principaux travaux de recherche effectués au sein de l’Institut de Mathématiques de Toulouse sont présentés dans ce manuscrit. J’ai pris le parti de rendre le contenu de chaque chapitre globalement indépendant des autres afin de permettre aux lectrices/lecteurs de feuilleter selon leurs intérêts. Le contexte scientifique et les principaux résultats sont présentés ; les articles originaux associés sont rappelés en début de chapitre et sont accessibles depuis ma page web professionnelle ² pour plus de détails, pour les preuves et les détails des simulations et applications omis dans ce manuscrit.

La première partie est dédiée aux travaux en statistique mathématique autour des modèles de mélange dans différents contextes. Le chapitre 2 résume des travaux en collaboration avec Béatrice Laurent et Clément Marteau autour de la détection de mélanges à deux composantes. Nous avons proposé des procédures de test et étudié les conditions de séparation optimales dans le cas unidimensionnel et pour le cas de la contamination en multidimensionnel. Dans le chapitre 3, l’estimation des paramètres d’un mélange de contamination à deux composantes est adressée. Avec Sébastien Gadat, Jonas Kahn et Clément Marteau, nous avons établi des vitesses de convergence optimale pour ce problème d’estimation des paramètres. Le chapitre 4 est dédié à l’estimation adaptative de densités par des mélanges gaussiens, ce travail faisant suite à une collaboration débutée avec Bertrand Michel en thèse. Enfin cette première partie se conclut par le chapitre 5 où, avec Yohann De Castro, Sébastien Gadat et Clément Marteau, nous nous sommes intéressés aux performances d’un estimateur induit par la méthode Beurling-LASSO pour l’estimation d’une mesure de mélange.

La deuxième partie est dédiée aux contributions pour l’analyse de données transcriptomiques. Le chapitre 6 résume plusieurs travaux autour de la question de la classification non supervisée de gènes co-exprimés à partir de données de séquençage à haut-débit (bulk RNA-seq). Les trois articles résumés sont le fruit d’une collaboration de plusieurs années avec Andréa Rau, débutant lors du post-doctorat d’Andréa et jusqu’au co-encadrement du post-doctorat d’Antoine Godichon-Baggioni financé par le projet ANR-JCJC MixStatSeq. Dans le chapitre 7, quelques challenges autour de l’analyse des données single-cell RNA-seq sont présentés. Ce sont des perspectives de travail issues de plusieurs encadrements de stage interdisciplinaires, du projet TTIL SingleCell et du projet DDisc autour de ces nouvelles données de séquençement à haut-débit en cellule unique.

Dans tout le manuscrit, les citations entre [.] réfèrent à mes publications dont la liste est donnée en page 85.

²<https://www.math.univ-toulouse.fr/~maugis/>

Notations

Les notations suivantes sont utilisées dans les différents chapitres de ce manuscrit :

\mathbb{R}	ensemble des nombres réels
\mathbb{N}	ensemble des nombres entiers naturels
\mathbb{Z}	ensemble des nombres entiers relatifs
\mathbb{P}	Notation générique d'une probabilité
\mathbb{E}	Notation générique d'une espérance

$\underline{X} = (X_1, \dots, X_n)$ un n -échantillon

Pour deux suites de réels $(a_n)_{n \in \mathbb{N}}$ et $(b_n)_{n \in \mathbb{N}}$,

$a_n \lesssim b_n$ $\exists C > 0$ indépendante de n telle que $a_n \leq Cb_n$ pour tout $n \in \mathbb{N}$

$a_n \underset{n \rightarrow +\infty}{\sim} b_n$ $\lim_{n \rightarrow +\infty} \frac{a_n}{b_n} = 1$

$a_n \ll b_n$ $\lim_{n \rightarrow +\infty} \frac{a_n}{b_n} = 0$.

Première partie

Quelques contributions théoriques
autour des modèles de mélanges

Détection de mélanges à deux composantes

Résumé Dans ce chapitre sont résumés les travaux publiés dans [LMM16] et [LMM18]. Soit un n -échantillon X_1, \dots, X_n de \mathbb{R}^d de densité inconnue f . Dans le cas unidimensionnel, on s'intéresse au problème de détection d'un mélange à deux composantes : se donnant une densité paire ϕ , l'objectif est de déterminer si f est de la forme $\phi(\cdot - \mu_0)$ avec μ_0 inconnu, ou $(1 - \varepsilon)\phi(\cdot - \mu_1) + \varepsilon\phi(\cdot - \mu_2)$. Dans le cas multidimensionnel, nous avons abordé le problème simplifié de la contamination ($\mu_0 = \mu_1 = 0$ fixés) pour un mélange gaussien. Des conditions de séparation optimales sont établies et plusieurs procédures de test sont proposées pour répondre aux problèmes abordés dans différents régimes.

[LMM16] Laurent B., Marteau C. and Maugis-Rabusseau C. (2016). Non asymptotic detection of two component mixtures with unknown means. *Bernoulli*, 22, 242-274.

[LMM18] Laurent B., Marteau C. and Maugis-Rabusseau C. (2018) Multidimensional two-component Gaussian mixtures detection. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 54 (2), 842-865.

Sommaire

2.1. Contexte	8
2.2. Détection de mélanges unidimensionnels à deux composantes . . .	10
2.2.1. Procédure de test fondée sur les statistiques d'ordre	10
2.2.2. Etude dans le régime dense	11
2.2.3. Etude dans le régime sparse	13
2.3. Détection de mélanges gaussiens multidimensionnels à deux composantes	15
2.3.1. Frontière de détection pour la norme ℓ_2	15
2.3.2. Frontière de détection pour la norme ℓ_∞	19
2.4. Conclusions et perspectives	21

2.1 Contexte

Dans ce chapitre, on s'intéresse au problème de la détection d'une loi de mélange à deux composantes. On considère un n -échantillon i.i.d $\underline{X} = (X_1, \dots, X_n)$, $X_i \in \mathbb{R}^d$ de loi F inconnue. On suppose que F admet une densité f par rapport à la mesure de Lebesgue sur \mathbb{R}^d . On souhaite ici tester si l'échantillon provient d'une population homogène ou hétérogène au sens suivant :

$$\begin{aligned} \mathcal{H}_0 : f \in \mathcal{F}_0 \text{ contre } \mathcal{H}_1 : f \in \mathcal{F}_1 \\ \text{avec} \\ \mathcal{F}_0 = \{x \in \mathbb{R}^d \mapsto \phi(x - \mu_0), \mu_0 \in \mathbb{R}^d\} \\ \mathcal{F}_1 = \{x \in \mathbb{R}^d \mapsto (1 - \varepsilon)\phi(x - \mu_1) + \varepsilon\phi(x - \mu_2); \varepsilon \in]0, 1[, (\mu_1, \mu_2) \in (\mathbb{R}^d)^2\} \end{aligned} \quad (2.1)$$

où $\phi(\cdot)$ est une densité dont on précisera les hypothèses par la suite.

On parle de *modèle de contamination* dans le cas particulier où $\mu_0 = \mu_1$ connu (et même nul sans perte de généralité). On se retrouve alors à tester

$$\begin{aligned} \mathcal{H}_0 : f = \phi \text{ contre } \mathcal{H}_1 : f \in \mathcal{F}_1 \\ \text{avec } \mathcal{F}_1 = \{x \in \mathbb{R}^d \mapsto (1 - \varepsilon)\phi(x) + \varepsilon\phi(x - \mu_2); \varepsilon \in]0, 1[, \mu_2 \in \mathbb{R}^d\}. \end{aligned} \quad (2.2)$$

Dans les problèmes de test impliquant des mélanges finis unidimensionnels, les propriétés du test du rapport de vraisemblance ont été largement étudiées. On peut citer en autres Azaïs *et al.* (2009); Chernoff et Lander (1995); Dacunha-Castelle et Gassiat (1999); Garel (2007). Dans ces articles, le principal objectif est de déterminer le comportement asymptotique du rapport de vraisemblance sous l'hypothèse alternative afin d'étudier la puissance du test associé. Des méthodes alternatives ont également été envisagées : test du rapport de vraisemblance modifié (Chen *et al.*, 2001), estimation de la distance L^2 entre les densités associées aux hypothèses nulle et alternative (Charnigo et Sun, 2004), approche EM (Chen et Li, 2009) ou des tests basés sur la fonction caractéristique empirique (Klar et Meintanis, 2005).

Le principal défi lié au problème (2.1) est d'établir des conditions (optimales) sur $(\varepsilon, \mu_1, \mu_2)$ pour lesquelles une erreur de seconde espèce β peut être atteinte. La première étude en ce sens est due à Ingster (1999), dans le cas particulier de la contamination avec $\phi(\cdot)$ la densité gaussienne standard. Des résultats similaires ont été obtenus dans Donoho et Jin (2004), où la procédure "*Higher Criticism*" (*HC*) est présentée. Cet algorithme est facile à mettre en œuvre et fournit une puissance similaire à celle du test de rapport de vraisemblance habituel. Plusieurs extensions de cette contribution ont été proposées dans un contexte élargi : par exemple Cai *et al.* (2007) pour une étude incluant les ensembles de confiance et le cadre *dense* ($\varepsilon \gg 1/\sqrt{n}$ quand $n \rightarrow +\infty$); Cai *et al.* (2011) pour les mélanges hétérogènes et hétéroscédastiques; Cai et Wu (2014) où des distributions générales et des conditions de séparation ont été étudiées; Jager et Wellner (2007) pour des tests basés sur les divergences de Renyi généralisant le *Higher Criticism*. En particulier, les régions de détection asymptotique ont été soigneusement étudiées dans le cas gaussien dans les deux régimes asymptotiques suivants :

- le régime *sparse* :

$$\varepsilon \underset{n \rightarrow +\infty}{\sim} n^{-\delta} \text{ et } \mu_2 - \mu_1 \underset{n \rightarrow +\infty}{\sim} \sqrt{2r \log(n)} \text{ avec } \delta \in]\frac{1}{2}, 1[\text{ et } r \in]0, 1[.$$

Dans ce cas, la séparation est asymptotiquement possible si

$$\begin{cases} r > \delta - \frac{1}{2} & \text{quand } \frac{1}{2} < \delta \leq \frac{3}{4} \\ r > (1 - \sqrt{1 - \delta})^2 & \text{quand } \frac{3}{4} < \delta < 1 \end{cases} ;$$

- le régime dense :

$$\varepsilon \underset{n \rightarrow +\infty}{\sim} n^{-\delta} \text{ et } \mu_2 - \mu_1 \underset{n \rightarrow +\infty}{\sim} n^{-r} \text{ avec } \delta \in]0, \frac{1}{2}[\text{ et } r \in]0, \frac{1}{2}[.$$

Dans ce cas, la séparation est asymptotiquement possible si $r < \frac{1}{2} - \delta$.

Dans les contributions mentionnées ci-dessus, seules des distributions unidimensionnelles sont considérées. Dans un cadre différent (détection de signaux), les problèmes multidimensionnels ont été étudiés par Arias-Castro *et al.* (2011); Butucea et Ingster (2013) entre autres. Dans Verzelen et Arias-Castro (2017), le problème du test de normalité dans un cadre multidimensionnel est abordé avec des alternatives de mélanges gaussiens à deux composantes où les proportions sont fixes et les différences de moyennes sont sparses. Addario-Berry *et al.* (2010) déterminent des conditions de séparation non-asymptotiques pour la contamination d'un vecteur gaussien standard par des composantes de moyenne non nulle lorsque les alternatives ont des structures combinatoires et géométriques particulières.

Dans ce chapitre, la Section 2.2 est dédiée à la présentation des résultats publiés dans [LMM16] pour le problème de test général (2.1) dans le cas unidimensionnel ($d = 1$). Les travaux publiés dans [LMM18] pour le problème de contamination (2.2) dans le cas multidimensionnel sont présentés en Section 2.3. Dans ces deux contributions, des procédures de test sont proposées avec une étude non-asymptotique de leur performance. Dans la suite, on utilisera les notations suivantes :

- pour toute densité f sur \mathbb{R}^d , on note respectivement \mathbb{P}_f et \mathbb{E}_f la probabilité et l'espérance sous l'hypothèse que la densité commune de l'échantillon $\underline{X} = (X_1, \dots, X_n)$ est f . Dans le cas où l'hypothèse nulle est simple ($\mathcal{F}_0 = \{\phi\}$), on utilisera les notations \mathbb{P}_0 et \mathbb{E}_0 ;
- dans le cas où on se restreint aux mélanges gaussiens, la densité ϕ sera notée ϕ_G pour la densité de la gaussienne standard et l'ensemble d'alternatives sera noté $\mathcal{F}_{1,G}$;
- une procédure de test Ψ_α de niveau $\alpha \in]0, 1[$ est une fonction mesurable de (X_1, \dots, X_n) à valeur dans $\{0, 1\}$ telle que l'hypothèse nulle est rejetée si $\Psi_\alpha = 1$ et

$$\sup_{f \in \mathcal{F}_0} \mathbb{P}_f(\Psi_\alpha = 1) \leq \alpha.$$

L'objectif est de déterminer des conditions explicites non-asymptotiques sur les paramètres $(\varepsilon, \mu_1, \mu_2)$ qui garantissent une puissance donnée pour la procédure de test Ψ_α sachant que f appartient à un ensemble d'alternatives \mathcal{F}_1 . Plus précisément, pour un mélange $f \in \mathcal{F}_1$ tel que $d(\varepsilon, \mu_1, \mu_2) \geq \rho$, où $d(\varepsilon, \mu_1, \mu_2)$ représente une "distance" (à préciser) à l'hypothèse nulle, on cherche la valeur minimale de ρ qui nous assure de bien détecter l'alternative. Et sous cette condition, on veut s'assurer une garantie sur la puissance du test considéré Ψ_α . Pour cela, on reprend la définition de borne inférieure non-asymptotique pour un problème de test introduit par Baraud (2002). Pour $\beta \in]0, 1[$, on définit le niveau de séparation uniforme

$\rho(\Psi_\alpha, \mathcal{F}_1, \beta)$ d'une procédure de test Ψ_α sur une classe d'alternative \mathcal{F}_1 comme la plus petite valeur positive de ρ telle que le test a une erreur de seconde espèce au plus égale à β pour tout f dans \mathcal{F}_1 tel que $d(\varepsilon, \mu_1, \mu_2) \geq \rho$:

$$\rho(\Psi_\alpha, \mathcal{F}_1, \beta) = \inf \left\{ \rho > 0 \mid \sup_{f \in \mathcal{F}_1, d(\varepsilon, \mu_1, \mu_2) \geq \rho} \mathbb{P}_f(\Psi_\alpha = 0) \leq \beta \right\}.$$

Alors le niveau de séparation (α, β) -minimax sur \mathcal{F}_1 est défini par

$$\underline{\rho}(\mathcal{F}_1, \alpha, \beta) = \inf_{\Psi_\alpha} \rho(\Psi_\alpha, \mathcal{F}_1, \beta), \quad (2.3)$$

où l'infimum est pris sur tous les tests Ψ_α de niveau α .

2.2 Détection de mélanges unidimensionnels à deux composantes

Dans cette section, on se place dans le cadre général (2.1) unidimensionnel ($d = 1$) où \mathcal{F}_0 est un ensemble de densités translatées, le paramètre de moyenne μ_0 sous \mathcal{H}_0 est donc inconnu. On suppose que la densité $\phi(\cdot)$ est une fonction paire connue. L'alternative \mathcal{F}_1 contient les mélanges de deux densités de \mathcal{F}_0 , on suppose que $\mu_1 < \mu_2$ dans la suite.

2.2.1 Procédure de test fondée sur les statistiques d'ordre

Dans [LMM16], nous avons proposé une procédure de test basée sur les espacements des statistiques d'ordre $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. L'idée principale sous-jacente est que les écarts entre statistiques d'ordre $X_{(l)} - X_{(k)}$ sont libres par rapport à la moyenne μ_0 inconnue sous \mathcal{H}_0 . De plus, la distribution des variables $X_{(l)} - X_{(k)}$ est connue sous \mathcal{H}_0 et a un comportement différent sous \mathcal{H}_1 , à condition que k et l soient bien choisis.

Soit $\alpha \in]0, 1[$ fixé. On suppose que $n \geq 2$ et on considère \mathcal{K}_n , un sous-espace de $\{1, 2, \dots, n/2\}$, défini par

$$\mathcal{K}_n = \{2^j, 0 \leq j \leq \lfloor \log_2(n/2) \rfloor\}.$$

Notre procédure de test est alors définie par

$$\Psi_\alpha^{\text{ord}} := \sup_{k \in \mathcal{K}_n} \left\{ \mathbb{1}_{X_{(n-k+1)} - X_{(k)} > q_{\alpha_n, k}} \right\}, \quad (2.4)$$

où, pour tout $u \in]0, 1[$, $q_{u, k}$ est le $(1 - u)$ -quantile de $X_{(n-k+1)} - X_{(k)}$ sous l'hypothèse nulle et

$$\alpha_n = \sup \left\{ u \in]0, 1[, \mathbb{P}_{H_0}(\exists k \in \mathcal{K}_n, X_{(n-k+1)} - X_{(k)} > q_{u, k}) \leq \alpha \right\}.$$

Par définition, Ψ_α^{ord} est un test de niveau α . En pratique, puisque la distribution de $X_{(n-k+1)} - X_{(k)}$ sous l'hypothèse nulle est indépendante de la valeur de μ_0 , $q_{\alpha_n, k}$ et α_n peuvent être approximés (via des simulations de Monte-Carlo par exemple). On peut aussi utiliser la majoration explicite des quantiles établie dans [LMM16]. En pratique, le choix de α_n , au lieu de la correction de Bonferroni $\alpha/|\mathcal{K}_n|$, permet une amélioration numérique des performances de Ψ_α^{ord} (voir Fromont et Laurent, 2006, pour une discussion approfondie sur ce sujet).

Dans [LMM16], nous avons établi des performances non-asymptotiques de notre procédure de test Ψ_α^{ord} dans le cadre général (ϕ est juste une densité paire). Soit $\bar{\Phi}(x) = 1 - \Phi(x)$, où Φ

est la fonction de répartition associée à la densité ϕ . Pour tout $\alpha \in]0, 1[$ et $k \in \{1, 2, \dots, n/2\}$, soit $t_{\alpha,k}$ un nombre réel positif défini par

$$\bar{\Phi}\left(\frac{t_{\alpha,k}}{2}\right) = \frac{k}{n} \left[1 - \sqrt{\frac{2 \log(\frac{4}{\alpha})}{k}} \right]$$

si $k > 2 \log(\frac{4}{\alpha})$, et $t_{\alpha,k} = +\infty$ sinon. Pour tout $\alpha \in]0, 1[$, $\rho > 0$, et $k \in \{1, 2, \dots, n/2\}$, on considère le sous-ensemble $\bar{\mathcal{S}}(\alpha, \rho, k)$ de \mathbb{R}^3 défini par

$$\bar{\mathcal{S}}(\alpha, \rho, k) = \left\{ \begin{array}{l} (\varepsilon, \mu_1, \mu_2) \in]0, 1[\times \mathbb{R}^2, \mu_1 < \mu_2; \exists c \in \mathbb{R} \text{ tel que :} \\ (1 - \varepsilon)\bar{\Phi}(t_{\alpha,k} - c + \varepsilon(\mu_2 - \mu_1)) + \varepsilon\bar{\Phi}(t_{\alpha,k} - c - (1 - \varepsilon)(\mu_2 - \mu_1)) > \rho \\ (1 - \varepsilon)\bar{\Phi}(c - \varepsilon(\mu_2 - \mu_1)) + \varepsilon\bar{\Phi}(c + (1 - \varepsilon)(\mu_2 - \mu_1)) > \rho \end{array} \right\}. \quad (2.5)$$

Quand $t_{\alpha,k} = +\infty$, on prend comme convention que $\bar{\mathcal{S}}(\alpha, \rho, k) = \emptyset$ pour tout $\rho > 0$.

Théorème 2.2.1. *Soit $\alpha \in]0, 1[$ et $\beta \in]0, 1 - \alpha[$. Soit la procédure de test Ψ_α^{ord} décrite dans (2.4). On suppose que $n \geq 8 \log(4/\alpha_n)$. On considère les alternatives*

$$\bar{\mathcal{F}}_1[n, \alpha, \beta] = \left\{ f(\cdot) = (1 - \varepsilon)\phi(\cdot - \mu_1) + \varepsilon\phi(\cdot - \mu_2); (\varepsilon, \mu_1, \mu_2) \in \bigcup_{k \in \mathcal{K}_n} \bar{\mathcal{S}}(\alpha_n, \rho(k, n), k) \right\}$$

où, pour tout $k \in \mathcal{K}_n$, $\bar{\mathcal{S}}(\alpha_n, \rho(k, n), k)$ est défini par (2.5) avec

$$\rho(k, n) = \frac{k}{n} + \frac{1 + \sqrt{1 + 2k\beta}}{n\beta}.$$

Alors Ψ_α^{ord} est un test de niveau α et

$$\sup_{f \in \bar{\mathcal{F}}_1[n, \alpha, \beta]} \mathbb{P}_f(\Psi_\alpha^{ord} = 0) \leq \beta.$$

Dans le Théorème 2.2.1, la condition $n \geq 8 \log(4/\alpha_n)$ assure qu'il existe $k \in \mathcal{K}_n$ tel que $k > 2 \log(4/\alpha_n)$. On a défini un ensemble $\bar{\mathcal{F}}_1[n, \alpha, \beta]$ pour lequel le test Ψ_α^{ord} est de niveau α et de puissance plus grande que $1 - \beta$. Mais cet ensemble d'alternatives s'appuie sur des ensembles $\bar{\mathcal{S}}(\alpha, \rho, k)$ (voir Equation (2.5)) dont la définition est peu intuitive pour le moment. Néanmoins, c'est le résultat fondamental sur lequel on s'appuie ensuite pour évaluer la puissance de notre procédure de test dans plusieurs situations, asymptotiques et non-asymptotiques.

2.2.2 Etude dans le régime dense

On se focalise ici sur le *régime dense* au sens suivant : on suppose que la différence entre μ_1 et μ_2 est bornée. L'ensemble des alternatives est donc, pour $M > 0$,

$$\mathcal{F}_1[M] = \{f(\cdot) = (1 - \varepsilon)\phi(\cdot - \mu_1) + \varepsilon\phi(\cdot - \mu_2), \varepsilon \in]0, 1[, 0 < \mu_2 - \mu_1 \leq M\}.$$

Dans ce régime, la "distance" à l'hypothèse nulle est portée par la quantité

$$d(\varepsilon, \mu_1, \mu_2) = \varepsilon(1 - \varepsilon)(\mu_2 - \mu_1)^2.$$

2.2.2.1 Borne inférieure non-asymptotique dans le cas gaussien

On se place ici dans le cas de mélanges gaussiens. Le théorème suivant donne une borne inférieure non-asymptotique pour le niveau de séparation minimax $\underline{\rho}(\mathcal{F}_{1,G}[M], \alpha, \beta)$ (voir (2.3)).

Théorème 2.2.2. *Soit $\alpha \in]0, 1[$ et $\beta \in]0, 1 - \alpha[$. Soit*

$$\rho^* = \frac{1}{C(M)} \left(\sqrt{\frac{-2 \log[c(\alpha, \beta)]}{n}} \sqrt{1 + \frac{\log[c(\alpha, \beta)]}{2n}} \right),$$

avec $c(\alpha, \beta) = 1 - \frac{(1-\alpha-\beta)^2}{2}$ et $C(M) = \sqrt{\frac{1}{2} + \frac{2M^2}{3} e^{M^2/4}}$. Alors pour tout $\rho < \rho^*$,

$$\inf_{\substack{\Psi_\alpha \\ f \in \mathcal{F}_{1,G}[M] \\ d(\varepsilon, \mu_1, \mu_2) \geq \rho}} \sup \mathbb{P}_f(\Psi_\alpha = 0) > \beta,$$

où l'infimum est pris sur tous les tests Ψ_α de niveau α . Ceci implique que

$$\underline{\rho}(\mathcal{F}_{1,G}[M], \alpha, \beta) \geq \rho^*.$$

Le Théorème 2.2.2 implique que quelque soit le test Ψ_α de niveau α , si $\rho < \rho^*$, il existe une densité $f \in \mathcal{F}_{1,G}[M]$ pour laquelle $\mathbb{P}_f(\Psi_\alpha = 0) > \beta$. En particulier on ne peut pas détecter si $\mu_2 - \mu_1$ est trop petit par rapport à $\varepsilon(1 - \varepsilon)$.

2.2.2.2 Comportement du test Ψ_α^{ord} dans le cas gaussien

Nous allons maintenant étudier le comportement de notre procédure de test Ψ_α^{ord} pour des mélanges gaussiens dans le régime dense et voir si la condition sur $(\varepsilon, \mu_1, \mu_2)$ est optimale (à une constante près).

Théorème 2.2.3. *Soit X_1, \dots, X_n un n -échantillon de densité f . Soit $\alpha \in]0, 1[$ et Ψ_α^{ord} la procédure de test de niveau α définie par (2.4). Soit $\beta \in]0, 1 - \alpha[$ et $M > 0$. On suppose que n satisfait $n \geq 3$ et $8.25 \times \log(4 \log_2(n/2)/\alpha)/n \leq \int_M^\infty \phi_G(x) dx$.*

Alors, il existe une constante positive $C(\alpha, \beta, M)$, dépendant seulement de α, β et M , telle que si

$$\rho \geq C(\alpha, \beta, M) \sqrt{\frac{\log \log(n)}{n}}, \quad (2.6)$$

alors,

$$\sup_{\substack{f \in \mathcal{F}_{1,G}[M] \\ d(\varepsilon, \mu_1, \mu_2) \geq \rho}} \mathbb{P}_f(\Psi_\alpha^{\text{ord}} = 0) \leq \beta.$$

La valeur de ρ dans (2.6) diffère de la borne inférieure ρ^* par un terme d'ordre $\sqrt{\log \log n}$. Ce terme logarithmique est classique en théorie des tests (voir par ex. Spokoiny, 1996; Fromont et Laurent, 2006). Il est dû à la procédure de test multiple (adaptative) : la valeur optimale pour $k \in \mathcal{K}_n$ dans le test Ψ_α^{ord} est choisie à partir des données.

On adopte maintenant un point de vue asymptotique comme dans Donoho et Jin (2004) pour le cas des mélanges gaussiens. Le régime dense est paramétré par

$$\varepsilon \underset{n \rightarrow +\infty}{\sim} n^{-\delta} \text{ et } \mu_2 - \mu_1 \underset{n \rightarrow +\infty}{\sim} n^{-r} \text{ avec } 0 < \delta \leq \frac{1}{2} \text{ et } 0 < r < \frac{1}{2}. \quad (2.7)$$

Corollaire 2.2.4. *La borne de détection dans le régime dense (2.7) est*

$$r^*(\delta) = \frac{1}{4} - \frac{\delta}{2} :$$

la détection est possible si $r < r^*(\delta)$ et impossible si $r > r^*(\delta)$.

Rappelons que dans le cas de la contamination en régime dense, la détection est possible si $r < \frac{1}{2} - \delta$ (Ingster, 1999; Donoho et Jin, 2004). On obtient donc ici une borne différente dans notre problème plus général, du au fait que la moyenne sous \mathcal{H}_0 est inconnue, ce qui rend le problème de test plus difficile.

2.2.2.3 Procédure de test fondée sur la variance

Dans cette section, on ne suppose plus uniquement le cas gaussien. On introduit une procédure de test basée sur la variance des X_i . On remarque que sous \mathcal{H}_0 , $\text{Var}(X_i) = \sigma^2$, où $\sigma^2 = \int_{\mathbb{R}} x^2 \phi(x) dx$, alors que sous \mathcal{H}_1 , $\text{Var}(X_i) = \sigma^2 + \varepsilon(1 - \varepsilon)(\mu_2 - \mu_1)^2$. On considère donc la procédure de test Ψ_α^{var} définie par

$$\Psi_\alpha^{\text{var}} = \mathbb{1}_{\{S_n^2 > v_{\alpha,n}\}}, \text{ avec } S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \text{ et } \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad (2.8)$$

et $v_{\alpha,n}$ est le $(1 - \alpha)$ -quantile de la variable S_n^2 sous \mathcal{H}_0 .

On peut alors vérifier que ce test atteint la borne inférieure établie dans le Théorème 2.2.2.

Proposition 2.2.5. *Soit $\alpha \in]0, 1[$ et $\beta \in]0, 1 - \alpha[$. On suppose que ϕ a un moment d'ordre 4 fini : $\int_{\mathbb{R}} x^4 \phi(x) dx \leq B$. Il existe une constante positive $C(\alpha, \beta, M, B)$ telle que si*

$$\rho \geq C(\alpha, \beta, M, B) / \sqrt{n},$$

alors

$$\sup_{\substack{f \in \mathcal{F}_1[M] \\ d(\varepsilon, \mu_1, \mu_2) \geq \rho}} \mathbb{P}_f(\Psi_\alpha^{\text{var}} = 0) \leq \beta.$$

La procédure de test Ψ^{var} atteint donc la borne inférieure donnée dans le Théorème 2.2.2 (à une constante près). Ce résultat est meilleur que celui énoncé dans le Théorème 2.2.3 puisque la condition pour contrôler l'erreur de seconde espèce par β est $\varepsilon(1 - \varepsilon)(\mu_2 - \mu_1)^2 > C/\sqrt{n}$ au lieu de $C\sqrt{\log \log(n)}/\sqrt{n}$. Néanmoins, on verra par la suite que le test basé sur la variance n'est pas adapté dans le cas du régime sparse (voir Section 2.2.3.3). Or d'un point de vue pratique on souhaite fournir une procédure de test qui s'adapte à toutes les situations possibles. Pour cette procédure de test, on obtient les mêmes résultats que dans le Corollaire 2.2.4 d'un point de vue asymptotique dans le cas gaussien.

2.2.3 Etude dans le régime sparse

Nous allons maintenant examiner la situation où $\mu_2 - \mu_1$ est autorisé à tendre vers l'infini lorsque n augmente. Dans ce cas, il est intéressant de considérer des alternatives où $\varepsilon \ll \frac{1}{\sqrt{n}}$ quand $n \rightarrow +\infty$. Ce régime est appelé le *régime sparse*. On va étudier le comportement de notre procédure de test Ψ_α^{ord} dans ce régime sparse, pour des mélanges gaussiens et des mélanges de lois de Laplace.

2.2.3.1 Cas des mélanges gaussiens

On se place ici dans le cas gaussien $\phi(\cdot) = \phi_G(\cdot)$. Dans ce contexte, le régime *sparse* introduit dans Donoho et Jin (2004) est caractérisé par

$$\varepsilon \underset{n \rightarrow +\infty}{\sim} n^{-\delta} \text{ et } \mu_2 - \mu_1 \underset{n \rightarrow +\infty}{\sim} \sqrt{2r \log(n)} \text{ avec } \frac{1}{2} < \delta < 1 \text{ et } 0 < r < 1. \quad (2.9)$$

Théorème 2.2.6. *Soit X_1, \dots, X_n un n -échantillon de densité f . Soit $\alpha \in]0, 1[$ et Ψ_α^{ord} la procédure de test de niveau α définie par (2.4). On suppose que le comportement de $(\varepsilon, \mu_1, \mu_2)$ suit (2.9) et que $r > r^*(\delta)$ avec*

$$r^*(\delta) = \begin{cases} \delta - \frac{1}{2} & \text{si } \frac{1}{2} < \delta < \frac{3}{4} \\ (1 - \sqrt{1 - \delta})^2 & \text{si } \frac{3}{4} \leq \delta < 1 \end{cases}.$$

Alors, pour $f(\cdot) = (1 - \varepsilon)\phi_G(\cdot - \mu_1) + \varepsilon\phi_G(\cdot - \mu_2)$, on a pour n assez grand,

$$\mathbb{P}_f(\Psi_\alpha^{ord} = 0) \leq \beta.$$

Dans le régime *sparse*, on retrouve exactement les frontières de séparation connues pour le problème de contamination dans le cas gaussien. Par conséquent, le fait que la moyenne sous \mathcal{H}_0 soit inconnue n'affecte pas la difficulté du problème de test dans le cas *sparse*. Ceci prouve l'optimalité de notre procédure dans le régime *sparse* puisque les bornes inférieures établies dans le problème de contamination (voir Ingster, 1999; Cai *et al.*, 2011) fournissent également des bornes inférieures pour notre problème de test.

2.2.3.2 Cas des mélanges de lois de Laplace

On considère ici le problème de test (2.1) dans le cas particulier où ϕ est la densité de Laplace

$$\phi_L(x) = \frac{1}{2}e^{-|x|}, \quad \forall x \in \mathbb{R}.$$

Dans ce cas, Donoho et Jin (2004) ont montré que les cas d'intérêt dans le régime *sparse* correspondent à la paramétrisation suivante

$$\varepsilon \underset{n \rightarrow +\infty}{\sim} n^{-\delta} \text{ et } \mu_2 - \mu_1 \underset{n \rightarrow +\infty}{\sim} r \log(n) \text{ avec } \frac{1}{2} < \delta < 1 \text{ et } 0 < r < 1. \quad (2.10)$$

Les performances de notre procédure de test Ψ_α^{ord} sont données par le théorème suivant.

Théorème 2.2.7. *Soit X_1, \dots, X_n un n -échantillon de densité f . Soit $\alpha \in]0, 1[$ et soit Ψ_α^{ord} la procédure de test de niveau α définie par (2.4). On suppose que le comportement de $(\varepsilon, \mu_1, \mu_2)$ est régi par (2.10) et que $r > r^*(\delta)$ avec*

$$r^*(\delta) = 2\delta - 1.$$

Alors, pour $f(\cdot) = (1 - \varepsilon)\phi_L(\cdot - \mu_1) + \varepsilon\phi_L(\cdot - \mu_2)$, on a pour n assez grand,

$$\mathbb{P}_f(\Psi_\alpha^{ord} = 0) \leq \beta.$$

On remarque que la limite de détection $r^*(\delta)$ est la même que celle présentée par Donoho et Jin (2004). En s'appuyant sur leurs résultats dans le problème de contamination, on obtient ici aussi que notre procédure est optimale.

2.2.3.3 Test sur la variance dans le cas sparse

Si l'on revient sur la procédure de test Ψ_α^{var} définie par (2.8), on peut facilement voir qu'elle n'est pas adaptée au régime sparse. En effet, on peut remarquer que

$$\text{Var}_\phi(X_i) = \int_{\mathbb{R}} x^2 \phi(x) dx,$$

tandis que, pour tout $f = (1 - \varepsilon)\phi(\cdot - \mu_1) + \varepsilon\phi(\cdot - \mu_2)$

$$\text{Var}_f(X_i) = \int_{\mathbb{R}} x^2 \phi(x) dx + \varepsilon(1 - \varepsilon)(\mu_2 - \mu_1)^2.$$

Pour les mélanges gaussien et Laplace, dans les schémas asymptotiques respectifs (2.9) et (2.10), nous obtenons que

$$\text{Var}_f(X_i) - \text{Var}_\phi(X_i) = \varepsilon(1 - \varepsilon)(\mu_2 - \mu_1)^2 \ll \frac{1}{\sqrt{n}}, \text{ quand } n \rightarrow +\infty.$$

Puisque la variance est estimée à un taux paramétrique $1/\sqrt{n}$, le test Ψ_α^{var} introduit dans (2.8) va échouer : il ne pourra pas séparer \mathcal{H}_0 de \mathcal{H}_1 avec une puissance appropriée.

2.3 Détection de mélanges gaussiens multidimensionnels à deux composantes

Cette section résume les travaux publiés dans [LMM18]. On considère le problème de contamination (voir Equation (2.2)) où $\phi(\cdot) = \phi_G(\cdot)$ est ici la densité de la loi gaussienne standard d -dimensionnelle $\mathcal{N}_d(0_d, I_d)$. Ainsi les mélanges dans l'ensemble d'alternatives $\mathcal{F}_{1,G}$ sont paramétrés par le couple $(\varepsilon, \mu) \in]0, 1[\times \mathbb{R}^d$ ($\mu = \mu_2$ pour alléger les notations dans la suite).

Pour ce problème de contamination, nous avons proposé plusieurs procédures de test et étudié leurs performances de détection pour le cas de mélanges régis par la contrainte

$$d(\varepsilon, \mu) = \varepsilon \|\mu\|_{\square} \geq \rho,$$

où $\rho > 0$ est un rayon donné et $\|\cdot\|_{\square}$ désigne alternativement la norme ℓ_2 ($\|\mu\|_2 = (\sum_{j=1}^d \mu_j^2)^{1/2}$) et la norme ℓ_∞ ($\|\mu\|_\infty = \max_{1 \leq j \leq d} |\mu_j|$).

Dans ce cadre multidimensionnel, la définition des régimes *dense* et *sparse* est plus complexe puisque la dimension d du problème a une réelle influence sur le problème de détection, et elle dépend aussi de la norme utilisée pour l'étude. Nous avons obtenu une description précise du rayon de séparation optimal ρ pour le cas *dense* pour les deux normes et établi des résultats partiels pour le régime *sparse*.

2.3.1 Frontière de détection pour la norme ℓ_2

Dans cette section, on parle de régime *dense* lorsque $\varepsilon \geq d^{1/4}/\sqrt{n}$ et de régime *sparse* sinon.

2.3.1.1 Borne inférieure pour la norme ℓ_2 dans le régime dense

Une borne inférieure pour le niveau de séparation minimax $\underline{\rho}(\mathcal{F}_{1,G}, \alpha, \beta)$ dans le cadre du régime *dense* pour la norme ℓ_2 est établie dans le théorème suivant.

Théorème 2.3.1. *Supposons que $\varepsilon \geq d^{1/4}/\sqrt{n}$. Soit $\alpha, \beta \in]0, 1[$ tels que $\alpha + \beta < 0.29$. Soit*

$$\rho^* := 0.4 \frac{d^{1/4}}{\sqrt{n}}.$$

Alors, si $\rho < \rho^*$,

$$\inf_{\substack{\Psi_\alpha \\ f \in \mathcal{F}_{1,G} \\ \varepsilon \|\mu\|_2 \geq \rho}} \sup \mathbb{P}_f(\Psi_\alpha = 0) > \beta, \quad (2.11)$$

où l'infimum est pris sur tous les tests Ψ_α de niveau α . En particulier, cela implique que

$$\underline{\rho}(\mathcal{F}_{1,G}, \alpha, \beta) \geq \rho^*.$$

Le Théorème 2.3.1 quantifie précisément le rôle joué par la dimension d du problème et la condition $\varepsilon \|\mu\|_2 \geq Cd^{1/4}/\sqrt{n}$ régit l'ordre de grandeur de $\|\mu\|_2$ par rapport à un $\varepsilon \geq d^{1/4}/\sqrt{n}$ donné pour pouvoir détecter une contamination potentielle dans l'échantillon. Nous verrons par la suite que cette borne inférieure est optimale, à une constante près.

2.3.1.2 Procédure de test basée sur la moyenne empirique

La première procédure de test proposée est simplement basée sur les fluctuations de la moyenne empirique. Intuitivement, $\mathbb{E}_f[X] = \varepsilon\mu$ pour tout $f \in \mathcal{F}_{1,G}$ et $\mathbb{E}_0[X] = 0$ sous \mathcal{H}_0 . Aussi, si la moyenne empirique de l'échantillon a une norme élevée, il est probable que l'échantillon est issu d'une loi de $\mathcal{F}_{1,G}$. Ainsi on considère la procédure de test

$$\Psi_{1,\alpha} = \mathbb{1}_{\{\|\sqrt{n}\bar{X}_n\|_2^2 > t_{d,\alpha}\}} \text{ avec } \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (2.12)$$

et $t_{d,\alpha}$ le $(1 - \alpha)$ quantile d'une loi du khi-deux à d degrés de liberté. Le théorème suivant donne les performances de ce test.

Théorème 2.3.2. *Soit $\alpha, \beta \in]0, 1[$ fixés. On suppose que $n\varepsilon \geq \frac{8}{\beta}$. Alors la procédure de test $\Psi_{1,\alpha}$ définie par (2.12) est de niveau α , et il existe une constante positive $C(\alpha, \beta)$ telle que pour tout $f = f_{(\varepsilon,\mu)} \in \mathcal{F}_{1,G}$ avec*

$$\varepsilon^2 \|\mu\|_2^2 \geq C(\alpha, \beta) \frac{\sqrt{d}}{n},$$

$$\mathbb{P}_f(\Psi_{1,\alpha} = 0) \leq \beta.$$

Le résultat indique que le test $\Psi_{1,\alpha}$ est puissant dès que $f \in \mathcal{F}_{1,G}$ avec $\varepsilon \|\mu\|_2 \gtrsim d^{1/4}/\sqrt{n}$. D'après la borne inférieure obtenue dans le Théorème 2.3.1, il apparaît que dans le régime *dense*, la frontière de détection minimax est d'ordre $d^{1/4}/\sqrt{n}$ à une constante près.

2.3.1.3 Procédure de test basée sur le max des normes ℓ_2

La procédure $\Psi_{1,\alpha}$ définie par (2.12) est optimale dans le régime dense mais n'est pas performante dans le régime sparse ($\varepsilon < d^{1/4}/\sqrt{n}$). Nous proposons donc une autre procédure de test, basée sur le maximum des normes ℓ_2 des X_i :

$$\Psi_{2,\alpha} = \max_{1 \leq i \leq n} \mathbb{1}_{\{\|X_i\|_2^2 > t_{d,\alpha/n}\}}, \quad (2.13)$$

où $t_{d,\alpha/n}$ désigne le quantile $(1 - \alpha/n)$ de la distribution du khi-deux à d degrés de liberté. Les performances de cette procédure sont données dans le théorème suivant.

Théorème 2.3.3. *Soit $\alpha, \beta \in]0, 1[$ fixés. Nous supposons que $n\varepsilon \geq \frac{8}{\beta}$. Alors, la procédure de test $\Psi_{2,\alpha}$ définie par (2.13) est de niveau α . De plus, il existe une constante positive $C(\alpha, \beta)$ telle que pour tout $f = f_{(\varepsilon, \mu)} \in \mathcal{F}_{1,G}$ pour laquelle*

$$\varepsilon^2 \|\mu\|_2^2 \geq C(\alpha, \beta) \varepsilon^2 \left[\sqrt{d \ln(n)} + \ln(n) \right],$$

$$\mathbb{P}_f(\Psi_{2,\alpha} = 0) \leq \beta.$$

Au vu des résultats des Théorèmes 2.3.2 et 2.3.3, on peut alors agréger les deux tests précédents de niveau $\alpha/2$ chacun pour avoir une nouvelle procédure de test de niveau α , notée $\Psi_{1,\alpha/2} \vee \Psi_{2,\alpha/2}$, qui exploite le bon comportement de chaque test selon les régimes. L'erreur de seconde espèce de $\Psi_{1,\alpha/2} \vee \Psi_{2,\alpha/2}$ est contrôlée par

$$\mathbb{P}_f(\Psi_{1,\alpha/2} \vee \Psi_{2,\alpha/2} = 0) \leq \inf \{ \mathbb{P}_f(\Psi_{1,\alpha/2} = 0), \mathbb{P}_f(\Psi_{2,\alpha/2} = 0) \}.$$

On obtient ainsi le résultat suivant.

Corollaire 2.3.4. *Soit $\alpha, \beta \in]0, 1[$ fixés. Soient $\Psi_{1,\alpha}$ et $\Psi_{2,\alpha}$ les deux tests définis respectivement par (2.12) et (2.13). Il existe une constante positive $C(\alpha, \beta)$ telle que, si $n\varepsilon \geq \frac{8}{\beta}$, pour tout $f = f_{(\varepsilon, \mu)} \in \mathcal{F}_{1,G}$ qui vérifie*

$$\varepsilon^2 \|\mu\|_2^2 \geq C(\alpha, \beta) \left[\left(\frac{\sqrt{d}}{n} \right) \wedge \left\{ \varepsilon^2 \left(\sqrt{d \ln(n)} + \ln(n) \right) \right\} \right], \quad (2.14)$$

on a

$$\mathbb{P}_f(\Psi_{1,\frac{\alpha}{2}} \vee \Psi_{2,\frac{\alpha}{2}} = 0) \leq \beta.$$

Si on regarde de plus près le terme de droite dans l'inégalité (2.14),

$$\rho_{n,d,\varepsilon}^2 := \left(\frac{\sqrt{d}}{n} \right) \wedge \left\{ \varepsilon^2 \left(\sqrt{d \ln(n)} + \ln(n) \right) \right\},$$

on obtient les conditions de séparation suivantes

- Si $d \leq \ln(n)$, alors

$$\rho_{n,d,\varepsilon}^2 \leq 2\varepsilon^2 \ln(n) \mathbb{1}_{\varepsilon \leq \frac{d^{1/4}}{\sqrt{n \ln(n)}}} + \frac{\sqrt{d}}{n} \mathbb{1}_{\varepsilon \geq \frac{d^{1/4}}{\sqrt{n \ln(n)}}}$$

- Si $d \geq \ln(n)$, alors

$$\rho_{n,d,\varepsilon}^2 \leq 2\varepsilon^2 \sqrt{d \ln(n)} \mathbf{1}_{\varepsilon \leq 1/(\sqrt{n}(\ln(n))^{1/4})} + \frac{\sqrt{d}}{n} \mathbf{1}_{\varepsilon \geq 1/(\sqrt{n}(\ln(n))^{1/4})}$$

On voit donc que le niveau de séparation du test agrégé est inférieur à celui du test $\Psi_{1,\alpha}$ dans le cas où $d \leq \ln(n)$ et $\varepsilon \leq d^{1/4}/\sqrt{n \ln(n)}$, et dans le cas où $d \geq \ln(n)$ et $\varepsilon \leq 1/(\sqrt{n}(\ln(n))^{1/4})$. Malheureusement, nous ne savons pas si ces résultats sont optimaux car nous n'avons pas réussi à obtenir des bornes inférieures pour la norme ℓ_2 dans le régime sparse. On peut noter que les deux tests proposés peuvent être facilement utilisés en pratique.

2.3.1.4 Procédure de test basée sur les sous-échantillons

On peut améliorer les résultats du Corollaire 2.3.4 dans le régime sparse en agrégeant le test $\Psi_{1,\alpha}$ avec le test $\Psi_{3,\alpha}$ suivant

$$\Psi_{3,\alpha} = \sup_{U \in \mathcal{U}} \mathbf{1}_{\{T_U > t_{n,d,|U|,\alpha}\}} \quad (2.15)$$

où \mathcal{U} est l'ensemble des sous-ensembles non vide de $\{1, \dots, n\}$, $|U|$ est le cardinal d'un sous-ensemble U ,

$$T_U = \frac{1}{|U|} \left\| \sum_{i \in U} X_i \right\|_2^2,$$

$t_{n,d,k,\alpha} = d + 2\sqrt{d} x_{n,k,\alpha} + 2 x_{n,k,\alpha}$ et $x_{n,k,\alpha} = k \ln(en/k) + \ln(n/\alpha)$. On s'appuie ici sur des résultats de déviation des lois du khi-deux et du khi-deux décentré proposés dans Massart (2000) et Birgé (2001). On obtient alors les performances suivantes pour le test agrégé $\Psi_{1,\frac{\alpha}{2}} \vee \Psi_{3,\frac{\alpha}{2}}$.

Théorème 2.3.5. *Soit $\alpha, \beta \in]0, 1[$ fixés. Soit $\Psi_{1,\alpha}$ et $\Psi_{3,\alpha}$ les deux tests définis respectivement par (2.12) et (2.15). Il existe une constante positive $C(\alpha, \beta)$ telle que, pour tout $f = f_{(\varepsilon, \mu)} \in \mathcal{F}_{1,G}$ qui vérifie $n\varepsilon \geq \frac{8}{\beta}$ et*

$$\varepsilon^2 \|\mu\|_2^2 \geq C(\alpha, \beta) \left[\left(\frac{\sqrt{d}}{n} \right) \wedge \left\{ \varepsilon^2 \ln \left(\frac{1}{\varepsilon} \right) + \varepsilon^{3/2} \sqrt{\frac{d}{n} \ln \left(\frac{1}{\varepsilon} \right)} \right\} \right], \quad (2.16)$$

on a

$$\mathbb{P}_f(\Psi_{1,\frac{\alpha}{2}} \vee \Psi_{3,\frac{\alpha}{2}} = 0) \leq \beta.$$

L'ensemble optimal $U \in \mathcal{U}$ qui permet de dériver la condition de séparation a un cardinal d'ordre εn . Les conditions de séparation sont résumées dans la Figure 2.1. Cependant, on peut remarquer que le calcul de la procédure de test $\Psi_{3,\alpha}$ est probablement NP-difficile, contrairement au test $\Psi_{1,\alpha/2} \vee \Psi_{2,\alpha/2}$. Malheureusement, nous n'avons pas réussi à construire des bornes inférieures appropriées. Par conséquent, la détermination de la condition de séparation minimax lorsque $\varepsilon < d^{1/4}/\sqrt{n}$ reste un problème ouvert.

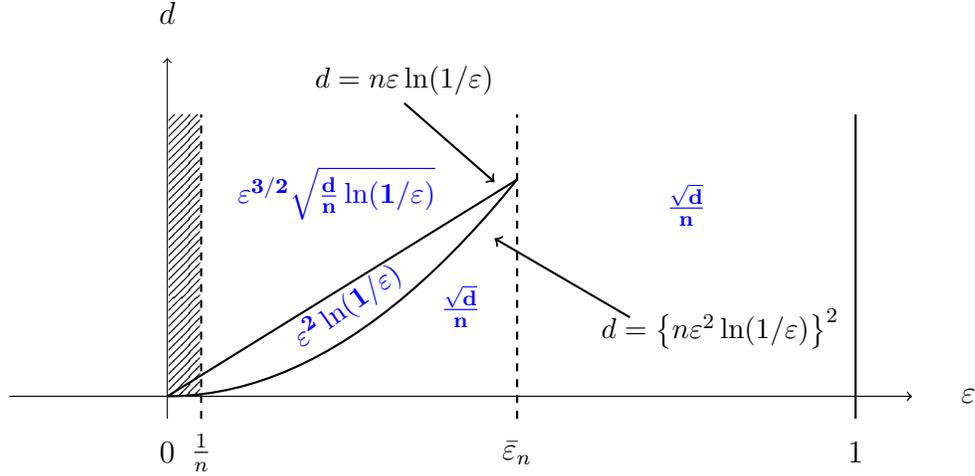


FIGURE 2.1 : Résumé de la condition de séparation sur $\varepsilon^2 \|\mu\|_2^2$ pour le test $\Psi_{1,\alpha/2} \vee \Psi_{3,\alpha/2}$, où $\bar{\varepsilon}_n = \inf \{ \varepsilon \in]0, 1[; n\varepsilon^3 \ln(1/\varepsilon) \geq 1 \}$

2.3.2 Frontière de détection pour la norme ℓ_∞

2.3.2.1 Borne inférieure pour la norme ℓ_∞

Dans cette section, on s'intéresse maintenant au niveau de séparation (α, β) -minimax pour $\mathcal{F}_{1,G}$ par rapport à la norme ℓ_∞ , noté $\rho_\infty(\mathcal{F}_{1,G}, \alpha, \beta)$.

Théorème 2.3.6. *Soit $\alpha, \beta \in]0, 1[$ tels que $\alpha + \beta < 1$. Soit $\eta(\alpha, \beta) = 2(1 - \alpha - \beta)$. On définit*

$$\rho^* = \varepsilon \sqrt{\ln \left[1 + \frac{1}{n\varepsilon^2} \ln(1 + d\eta(\alpha, \beta)^2) \right]}.$$

Alors, si $\rho \leq \rho^*$,

$$\inf_{\Psi_\alpha} \sup_{\substack{f \in \mathcal{F}_{1,G} \\ \varepsilon \|\mu\|_\infty \geq \rho}} \mathbb{P}_f(\Psi_\alpha = 0) > \beta, \quad (2.17)$$

où l'infimum est pris sur tous les tests Ψ_α de niveau α . En particulier, ce résultat implique que

$$\rho_\infty(\mathcal{F}_{1,G}, \alpha, \beta) \geq \rho^*.$$

De cette borne inférieure, on peut distinguer deux régimes :

- Cas 1 : Si $n\varepsilon^2 \gg \ln(d)$, ou si $c_1 \ln(d) \leq n\varepsilon^2 \leq c_2 \ln(d)$ pour $0 < c_1 < c_2$, alors

$$\rho^* \sim \sqrt{\ln(d)/n}.$$

- Cas 2 : Si $n\varepsilon^2 \ll \ln(d)$, alors $\rho^* \sim \varepsilon \sqrt{\ln(\ln(d)/n\varepsilon^2)}$.

Par analogie avec la discussion menée dans la section précédente, le premier cas peut être considéré comme le *régime dense* : ce cas permet notamment de détecter une contamination bornée. Le contrôle dans le Cas 1 fournit une condition de séparation similaire par rapport au

cas ℓ_2 , sauf que la dépendance par rapport à d change : la quantité $d^{1/4}$ est remplacée ici par $\sqrt{\ln(d)}$. En revanche, dans le cas où $n\varepsilon^2 \ll \ln(d)$, qui peut être considéré comme un régime *sparse* dans cette analyse ℓ_∞ , la condition de séparation est de la forme $\varepsilon\sqrt{\ln(\ln(d)/n\varepsilon^2)}$. L'optimalité de ces bornes est discutée dans la section suivante.

2.3.2.2 Procédures de test pour la norme ℓ_∞

Deux procédures de test pour la norme ℓ_∞ sont ici proposées. Elles consistent à appliquer une procédure de test sur chaque direction canonique afin de détecter une éventuelle contamination. Pour chaque $i \in \{1, \dots, n\}$, on note $X_i = (X_{ij})_{1 \leq j \leq d}$. Pour un $j \in \{1, \dots, d\}$ donné, on peut remarquer que $(X_{ij})_{i=1, \dots, n}$ est un échantillon unidimensionnel, distribué selon

1. la distribution gaussienne standard $\mathcal{N}(0, 1)$ sous \mathcal{H}_0 ,
2. le mélange unidimensionnel $(1 - \varepsilon)\phi_1(\cdot) + \varepsilon\phi_1(\cdot - \mu_j)$ sous \mathcal{H}_1 .

Première procédure de test La première procédure s'appuie sur le comportement moyen par direction :

$$\Psi_{4,\alpha} = \max_{j=1, \dots, d} \mathbb{1} \left\{ n\bar{X}_j^2 > t_{1, \frac{\alpha}{d}} \right\} \quad (2.18)$$

où $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$ et $t_{1, \frac{\alpha}{d}}$ est le $(1 - \frac{\alpha}{d})$ -quantile d'une distribution du khi-deux à un degré de liberté.

Théorème 2.3.7. *Soit $\alpha, \beta \in]0, 1[$ fixés. Nous supposons que $n\varepsilon \geq \frac{8}{\beta}$. Alors, la procédure de test $\Psi_{4,\alpha}$ introduite dans (2.18) est de niveau α . De plus, il existe une constante positive $C(\alpha, \beta)$ telle que pour tout $f = f_{(\varepsilon, \mu)} \in \mathcal{F}_{1,G}$ pour laquelle*

$$\varepsilon \|\mu\|_\infty \geq C(\alpha, \beta) \sqrt{\frac{\ln(d)}{n}},$$

$$\mathbb{P}_f(\Psi_{4,\alpha} = 0) \leq \beta.$$

Seconde procédure de test La seconde procédure est basée sur le comportement maximal par direction :

$$\Psi_{5,\alpha} = \max_{j=1, \dots, d} \mathbb{1} \left\{ \max_{i=1, \dots, n} X_{ij}^2 > t_{1, \frac{\alpha}{dn}} \right\} \quad (2.19)$$

où $t_{1, \frac{\alpha}{dn}}$ est le $(1 - \frac{\alpha}{dn})$ -quantile du khi-deux à un degré de liberté.

Théorème 2.3.8. *Soit $\alpha, \beta \in]0, 1[$ fixés. Nous supposons que $n\varepsilon \geq \frac{8}{\beta}$. Alors, la procédure de test $\Psi_{5,\alpha}$ introduite dans (2.19) est de niveau α . De plus, il existe une constante positive $C(\alpha, \beta)$ telle que pour tout $f = f_{(\varepsilon, \mu)} \in \mathcal{F}_{1,G}$ pour laquelle*

$$\|\mu\|_\infty \geq C(\alpha, \beta) \sqrt{\ln(dn)},$$

$$\mathbb{P}_f(\Psi_{5,\alpha} = 0) \leq \beta.$$

Comme expliqué dans la Section 2.3.1, on peut alors profiter de l'avantage de chaque test en les agrégeant. On obtient donc le test $\Psi_{4,\alpha/2} \vee \Psi_{5,\alpha/2}$ avec pour borne supérieure du niveau de séparation

$$\varepsilon \|\mu\|_\infty \geq C(\alpha, \beta) \left[\sqrt{\frac{\ln(d)}{n}} \wedge \varepsilon \sqrt{\ln(dn)} \right]. \quad (2.20)$$

Ainsi, si $n\varepsilon^2 \geq \ln(d)/\ln(dn)$ le terme de droite dans (2.20) est d'ordre $\sqrt{\ln(d)/n}$; sinon il est d'ordre $\varepsilon \sqrt{\ln(dn)}$.

Comparons maintenant ces bornes supérieures avec les bornes inférieures obtenues dans le Théorème 2.3.6. Dans le régime dense (noté Cas 1) où soit $n\varepsilon^2 \gg \ln(d)$, soit $c_1 \ln(d) \leq n\varepsilon^2 \leq c_2 \ln(d)$, les bornes supérieure et inférieure coïncident et nos résultats sont optimaux. Dans le régime sparse (noté Cas 2), un écart subsiste entre les bornes inférieure et supérieure.

2.4 Conclusions et perspectives

Dans le cas unidimensionnel, nous avons pu étudier le problème général (2.1) de détection d'une distribution de mélange. Grâce au Théorème 2.2.1, on peut envisager d'étendre l'étude de notre test basé sur les différences de statistiques d'ordre à d'autres lois dès lors que l'on peut avoir un contrôle assez fin de la fonction de répartition. Dans le cas gaussien, on met en évidence la différence de comportement selon le régime dense ou sparse et par rapport à un problème de contamination. Le passage au cas multidimensionnel n'est pas immédiat et assez technique. Cela explique que nous sommes revenus au cas plus simple de la contamination pour le cas gaussien. Les résultats présentés dans les Sections 2.3.1 et 2.3.2 constituent une première tentative de description des valeurs optimales de détection. Dans le régime *dense* ($\varepsilon \gtrsim d^{1/4}/\sqrt{n}$ et $\varepsilon \gtrsim \sqrt{\ln(d)/n}$ pour les normes l_2 et l_∞ respectivement) les résultats établis fournissent une caractérisation précise du rayon de séparation ρ . En revanche, dans le régime *sparse*, nous n'avons pas réussi à fournir une caractérisation précise de la dépendance par rapport à la dimension d pour ce problème de détection. Plusieurs perspectives de recherche pourraient être menées, notamment pour l'extension de ces résultats à d'autres distributions, pour des mélanges hétéroscédastiques, pour compléter les résultats en contamination pour le régime *sparse*.

Estimation des paramètres d'un mélange de contamination à deux composantes

Résumé Dans ce chapitre, nous présentons les résultats publiés dans [GKMM20]. Nous étudions un modèle de contamination paramétrique : Nous considérons un échantillon i.i.d de densité $f^* = (1 - \varepsilon^*)\phi + \varepsilon^*\phi(\cdot - \mu^*)$, où la fonction ϕ est supposée connue. Nous établissons des vitesses de convergence optimales pour l'estimation des paramètres de mélange $(\varepsilon^*, \mu^*) \in]0, 1[\times \mathbb{R}^d$. En particulier, nous prouvons que la vitesse paramétrique usuelle $1/\sqrt{n}$ ne peut pas être atteinte quand au moins un de ces paramètres est amené à tendre vers 0 avec n .

[GKMM20] Gadat S., Kahn J., Marteau C. and Maugis-Rabusseau C. (2020). Parameter recovery in two-component contamination mixtures : The L^2 strategy. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(2), 1391–1418.

Sommaire

3.1. Contexte et stratégie d'estimation	24
3.1.1. Cadre d'étude	24
3.1.2. Stratégie d'estimation pour la norme \mathbb{L}^2	25
3.1.3. Hypothèses sur ϕ	26
3.2. Vitesses d'estimation des paramètres (ε^*, μ^*)	26
3.3. Bornes inférieures	28
3.4. Lien entre Wasserstein et $\ \cdot\ _2$	29

3.1 Contexte et stratégie d'estimation

3.1.1 Cadre d'étude

On considère un échantillon $\underline{X} := (X_1, \dots, X_n)$ i.i.d. où chaque X_i admet la densité de probabilité f^* par rapport à la mesure de Lebesgue sur \mathbb{R}^d . Pour un modèle de mélange fini, la densité f^* est supposée avoir la forme suivante :

$$f^*(\cdot) = \sum_{k=1}^K \varepsilon_k \phi_k(\cdot) \quad (3.1)$$

où $(\varepsilon_1, \dots, \varepsilon_K) \in]0, 1[^K$ est le vecteur des proportions et ϕ_k est la densité de la k ème composante du mélange. Pour des raisons pratiques, on considère souvent des modèles paramétriques. Dans ce cas, les densités ϕ_k sont supposées connues, à certains paramètres finis près, et le problème de l'estimation des paramètres est souvent abordé à l'aide d'un algorithme de type EM (Dempster *et al.*, 1977). En pratique, cette étape d'estimation des paramètres est un point clé. Par exemple, on peut l'utiliser pour en déduire une classification non supervisée des données par la règle du Maximum A Posteriori (MAP) (voir par ex. McLachlan et Peel, 2000; Bouveyron *et al.*, 2019).

D'un point de vue théorique, l'intérêt pour la question de l'estimation des paramètres d'un mélange a reçu moins d'attention que celui de l'estimation de densité. Parmi les contributions, on peut citer le travail de Bunea *et al.* (2010). Se donnant un dictionnaire de densités, ils proposent une procédure d'estimation des proportions de mélange basée sur la minimisation d'un critère empirique \mathbb{L}^2 avec une contrainte de sparsité. Dans le cas particulier d'un dictionnaire $\{\phi_1, \dots, \phi_K; \phi_k = \phi(\cdot - \mu_k^*)\}$, leurs travaux fournissent une estimation des paramètres d'intérêt lorsque les paramètres de localisation μ_k^* sont assez séparés. Chen (1995) a étudié l'estimation de la distribution de mélange sous une forte condition d'identifiabilité. Comme observé dans les travaux de Nguyen (2013), Ho et Nguyen (2016) et Heinrich et Kahn (2018), la vitesse optimale dépend de la connaissance du nombre de composantes K . Ho et Nguyen (2016) montrent que les vitesses d'estimation des paramètres sont plus lentes pour certains mélanges faiblement identifiables. D'autres extensions sont disponibles dans Ho *et al.* (2016). Les problèmes d'identifiabilité (et d'estimation) sont discutés dans Hunter *et al.* (2007) sous l'hypothèse que $\phi_k = \phi(\cdot - \mu_k^*)$ pour une certaine séquence $(\mu_k^*)_{k=1, \dots, K}$ et une densité symétrique ϕ . Concernant l'estimation des paramètres via l'algorithme EM, les résultats théoriques de convergence sont limités. Un premier résultat est donné dans Wu *et al.* (1983) lorsque la densité est unimodale et que certaines conditions de régularité sont vérifiées. Plus récemment Balakrishnan *et al.* (2017) donnent un résultat général pour une convergence de l'algorithme EM sous des conditions d'initialisation, de Lipschitz et de concavité. Dans le cadre de l'estimation des paramètres d'un mélange à deux composantes où $f^* = (1 - \varepsilon^*)\phi(\cdot - \mu_1^*) + \varepsilon^*\phi(\cdot - \mu_2^*)$, Bordes *et al.* (2006) proposent une procédure de récurrence basée sur une formule d'inversion pour estimer $\theta^* = (\varepsilon^*, \mu_1^*, \mu_2^*)$ et la fonction ϕ (supposée symétrique inconnue). En particulier, le paramètre ε^* est estimé à la vitesse paramétrique classique $1/\sqrt{n}$, tandis que la vitesse $n^{-1/4}$ est obtenue pour les paramètres de localisation (μ_1^*, μ_2^*) . Un problème similaire est traité dans Butucea et Vandekerckhove (2014) où la vitesse $1/\sqrt{n}$ est atteinte pour l'estimation de θ^* . La procédure d'estimation est basée sur le calcul d'une transformée de Fourier empirique. Plus récemment, Patra et Sen (2016) ont considéré la situation où la distribution d'une des composantes du mélange est connue. Dans un tel cas de contamination, ils fournissent à la fois un estimateur du paramètre de mélange et de la distribution de la seconde composante.

Dans ce chapitre, on va également considérer un problème de contamination. On observe un échantillon i.i.d. $\underline{X} := (X_1, \dots, X_n)$ de densité $f^* \in \mathcal{F}$ par rapport à la mesure de Lebesgue sur \mathbb{R}^d où

$$\mathcal{F} = \left\{ f_\theta = f_{\varepsilon, \mu} = (1 - \varepsilon)\phi + \varepsilon\phi_\mu, \theta = (\varepsilon, \mu) \in]0, 1[\times \mathbb{R}^d \right\} \quad (3.2)$$

avec ϕ une densité connue et $\phi_\mu : x \mapsto \phi(x - \mu)$. En particulier, on note $f^* = f_{\theta^*} = f_{\varepsilon^*, \mu^*}$. Dans ce cadre de contamination, Cai *et al.* (2007) proposent une procédure basée sur la fonction de répartition empirique. Dans le régime dit *sparse* où $\varepsilon^* \ll 1/\sqrt{n}$ et $\mu^* \sim \sqrt{2r \log(n)}$ pour un certain $r \in]0, 1[$ comme $n \rightarrow +\infty$, les auteurs déduisent des taux de convergence pour l'estimation de ε^* .

L'objectif ici est d'établir des vitesses de convergence optimales pour l'estimation du vecteur des paramètres θ^* du mélange à partir de l'échantillon \underline{X} , en permettant aux paramètres d'évoluer avec n . Nous proposons une procédure inspirée de Bunea *et al.* (2010) et dérivons un estimateur $(\hat{\varepsilon}_n, \hat{\mu}_n)$ pour le couple (ε^*, μ^*) . Cet estimateur est basé sur la minimisation d'un contraste \mathbb{L}^2 au lieu d'un estimateur habituel de maximum de vraisemblance des paramètres du mélange. On va s'appuyer sur le fait que tout modèle de mélange de contamination à deux composantes est identifiable et on va chercher à combler l'écart entre l'estimation de densité et l'estimation des paramètres. Dans la suite, on suppose que le paramètre de translation inconnu μ^* appartient à un intervalle borné $[-M, M]^d$ où $M > 0$ est connu.

3.1.2 Stratégie d'estimation pour la norme \mathbb{L}^2

Notre estimateur est construit selon une estimation de densité \mathbb{L}^2 optimale, contrainte aux modèles de contamination. Pour cela, on commence par définir une grille sur les valeurs possibles de ε et μ :

$$\mathcal{M}_{\Lambda, \mathfrak{M}} := \{(\varepsilon, \mu) : \varepsilon \in \Lambda = \{\varepsilon_1, \dots, \varepsilon_p\} \text{ et } \mu \in \mathfrak{M} = \{\mu_1, \dots, \mu_q\}\},$$

où Λ et \mathfrak{M} seront précisés dans la suite en fonction de n . Visant à minimiser $\|f^* - f_{\varepsilon, \mu}\|_2^2$ sur la grille $\mathcal{M}_{\Lambda, \mathfrak{M}}$, notre estimateur est défini par

$$\hat{\theta} = (\hat{\varepsilon}_n, \hat{\mu}_n) = \underset{(\varepsilon, \mu) \in \mathcal{M}_{\Lambda, \mathfrak{M}}}{\operatorname{argmin}} -\frac{2}{n} \sum_{i=1}^n f_{\varepsilon, \mu}(X_i) + \|f_{\varepsilon, \mu}\|_2^2. \quad (3.3)$$

Dans un premier temps, on établit les performances de l'estimateur de densité $\hat{f}_n := f_{\hat{\varepsilon}_n, \hat{\mu}_n}$.

Théorème 3.1.1. *Soit $(\varepsilon^*, \mu^*) \in]0, 1[\times \mathbb{R}^d$. Soit $(\hat{\varepsilon}_n, \hat{\mu}_n)$ l'estimateur défini par (3.3). Alors, il existe une constante positive \mathcal{C} telle que pour tout $0 < \alpha < 1$:*

$$\mathbb{E} \left[\|\hat{f}_n - f^*\|_2^2 \right] \leq \left(\frac{1 + \alpha}{1 - \alpha} \right) \inf_{(\varepsilon, \mu) \in \mathcal{M}_{\Lambda, \mathfrak{M}}} \|f_{\varepsilon, \mu} - f^*\|_2^2 + \frac{\mathcal{C}}{2\alpha} \frac{\log^2(|\mathcal{M}_{\Lambda, \mathfrak{M}}|)}{n}, \quad (3.4)$$

où $|\mathcal{M}_{\Lambda, \mathfrak{M}}|$ est le cardinal de la grille $\mathcal{M}_{\Lambda, \mathfrak{M}}$.

Le Théorème 3.1.1 est une exigence minimale pour étendre notre analyse sur l'estimation des paramètres des modèles de mélange eux-mêmes. En particulier, la question suivante se pose maintenant : est-ce que le fait que \hat{f}_n est un "bon" estimateur au sens \mathbb{L}^2 de f^* implique que $\hat{\theta}_n$ associé fournit un estimateur satisfaisant de θ^* ? Pour répondre à cette question, nous allons faire quelques hypothèses sur la densité ϕ .

3.1.3 Hypothèses sur ϕ

Nous introduisons maintenant des hypothèses suffisantes sur ϕ pour établir des résultats d'estimation optimale de θ^* à partir de l'inégalité oracle (3.4). Dans la suite, on note $\mathcal{C}^k(\mathbb{R}^d)$ l'ensemble des fonctions continues qui admettent k dérivées continues et $\|\cdot\|$ la norme euclidienne usuelle de \mathbb{R}^d .

Hypothèse (\mathbf{H}_S) La densité ϕ appartient à $\mathcal{C}^3(\mathbb{R}^d) \cap \mathbb{L}^2(\mathbb{R}^d)$.

Hypothèse (\mathbf{H}_{Lip}) La densité ϕ satisfait

$$\exists g \in \mathbb{L}^2(\mathbb{R}^d) \quad \forall x \in \mathbb{R}^d \quad \forall \mu \in [-M, M]^d \quad |\phi(x) - \phi_\mu(x)| \leq \|\mu\|g(x), \quad (3.5)$$

et g satisfait

$$\mathcal{J} := \int_{\mathbb{R}^d} g^2(x)\phi^{-1}(x)dx < +\infty.$$

Hypothèse (\mathbf{H}_D) La densité ϕ satisfait

$$\mathcal{I}_\phi := \sup_{1 \leq j \leq d} \int \{d_{j,j}\phi(x)\}^2 \phi^{-1}(x)dx < +\infty, \quad (3.6)$$

où $d_{j,j}\phi$ est la dérivée seconde de ϕ par rapport à la variable j .

Ces hypothèses sont suffisantes, elles peuvent être assouplies (voir [GKMM20]). On peut noter que (\mathbf{H}_S) implique que la densité ϕ est nécessairement bornée sur \mathbb{R}^d . L'hypothèse (\mathbf{H}_{Lip}) est primordiale pour obtenir des résultats d'estimation sur les paramètres du mélange eux-mêmes. En particulier, cela permettra d'avoir une relation entre la norme \mathbb{L}^2 de $\phi - \phi_\mu$ et $\|\mu\|$. Par conséquent, sous l'hypothèse (\mathbf{H}_{Lip}), une bonne estimation de la densité f^* pour la norme \mathbb{L}^2 est supposée produire une bonne estimation des paramètres du mélange. L'hypothèse (\mathbf{H}_D) est seulement nécessaire pour établir nos résultats de borne inférieure.

On a montré dans [GKMM20] que ces trois hypothèses sont satisfaites pour toutes les densités log-concaves

$$\phi(\cdot) = e^{-u(\cdot)} \quad \text{avec } u \text{ convexe telle que } \|\nabla u\| + \|D^2 u\| = o_\infty(u),$$

où $o_\infty(u)$ est une quantité négligeable devant $u(x)$ quand $\|x\| \rightarrow +\infty$. Par cette remarque ou en adaptant un peu les hypothèses, on peut ainsi par exemple considérer pour ϕ la densité gaussienne standard, la densité de Laplace, la densité de Cauchy, la densité gaussienne asymétrique, ...

3.2 Vitesses d'estimation des paramètres (ε^*, μ^*)

Nous allons maintenant nous appuyer sur l'inégalité oracle du Théorème 3.1.1 pour établir des vitesses d'estimation des paramètres du mélange (ε^*, μ^*) . D'après l'hypothèse $\mu^* \in [-M, M]^d$ pour un $M > 0$ donné, on définit la grille $\mathcal{M}_n = \mathcal{M}_{\Lambda, \mathfrak{M}}$ par

$$\mathcal{M}_n = \left\{ (\varepsilon, \mu) : \varepsilon = \frac{i}{\sqrt{n}}, \mu = (\mu^{(1)}, \dots, \mu^{(d)}) \quad \text{avec} \quad \mu^{(j)} = \pm \frac{k_j}{\sqrt{n}} \right. \\ \left. \text{où } i \in \{1, \dots, \sqrt{n}\}, j \in \{1, \dots, d\}, k_j \in \{1, \dots, M\sqrt{n}\} \right\}, \quad (3.7)$$

de sorte que le terme d'approximation $\inf_{(\varepsilon, \mu) \in \mathcal{M}_n} \|f_{\varepsilon, \mu} - f^*\|_2^2$ dans l'équation (3.4) peut être rendu inférieur à n^{-1} , tout en gardant $\log(|\mathcal{M}_n|)$ raisonnable d'ordre $d \log(n)$. Le résultat suivant donne explicitement la vitesse non asymptotique de l'estimation de μ^* .

Théorème 3.2.1. *Soit $(\hat{\varepsilon}_n, \hat{\mu}_n)$ l'estimateur défini par (3.3) avec \mathcal{M}_n donné dans (3.7). Si ϕ satisfait les hypothèses (\mathbf{H}_S) et $(\mathbf{H}_{\text{Lip}})$, une constante $C_1 > 0$ existe telle que :*

$$\forall n \in \mathbb{N} \quad \sup_{(\varepsilon^*, \mu^*) \in]0, 1[\times [-M, M]^d} \mathbb{E}_{\varepsilon^*, \mu^*} [(\varepsilon^* \|\mu^*\|)^2 \|\hat{\mu}_n - \mu^*\|^2] \leq \frac{C_1 \log^2 n}{n}.$$

Dans le cas unidimensionnel ($d = 1$), une conséquence immédiate du Théorème 3.2.1 est que pour un couple $(\varepsilon^*, \mu^*) \in]0, 1[\times \mathbb{R}^*$:

$$\mathbb{E}_{\varepsilon^*, \mu^*} \left[\left(\frac{\hat{\mu}_n}{\mu^*} - 1 \right)^2 \right] \leq \frac{C_1 \log^2 n}{n \{\varepsilon^*\}^2 \{\mu^*\}^4}.$$

En particulier, puisque μ^* est autorisé à tendre vers 0 avec n , l'estimateur $\hat{\mu}_n$ sera cohérent dès que $\sqrt{n} \varepsilon^* \{\mu^*\}^2 \rightarrow +\infty$ comme $n \rightarrow +\infty$. Dans un contexte de détection (voir Chapitre 2), une loi de mélange à deux composantes peut être détectée de celle d'une seule composante dès que $\sqrt{n} \varepsilon^* |\mu^*| > C$ pour une constante positive C . Naturellement, la détection est "plus facile" que l'estimation. Puisque μ^* est supposé majoré, on suppose implicitement que $\varepsilon^* \gg 1/\sqrt{n}$ quand $n \rightarrow +\infty$.

Avant de vérifier l'optimalité de ce résultat (voir Section 3.3), nous étudions l'estimation de la proportion de contamination ε^* . D'après la discussion précédente, nous supposons que $\varepsilon^* \|\mu^*\|^2$ est significativement plus grand que $\log^2(n)/\sqrt{n}$. Cela garantit que le niveau de contamination μ^* est estimé de manière cohérente. Pour cela, nous introduisons l'ensemble $\Theta_n(M, (\ell_n)_n, \bar{\varepsilon})$ indexé par une suite $(\ell_n)_n$:

$$\Theta_n(M, (\ell_n)_n, \bar{\varepsilon}) := \left\{ \theta = (\varepsilon, \mu) : \frac{\ell_n}{\|\mu\|^2 \sqrt{n}} \leq \varepsilon \leq \bar{\varepsilon}, \|\mu\|_\infty \leq M \right\},$$

pour un $\bar{\varepsilon} \in]0, 1[$.

Théorème 3.2.2. *Si ϕ satisfait les hypothèses (\mathbf{H}_S) et $(\mathbf{H}_{\text{Lip}})$, et la suite $(\ell_n)_n$ est telle que $\lim_{n \rightarrow +\infty} \frac{\ell_n}{\log n} = +\infty$, alors il existe une constante $C_2 > 0$ telle que :*

$$\sup_{(\varepsilon^*, \mu^*) \in \Theta_n(M, (\ell_n)_n, \bar{\varepsilon})} \mathbb{E}_{\varepsilon^*, \mu^*} [\|\mu^*\|^4 (\hat{\varepsilon}_n - \varepsilon^*)^2] \leq \frac{C_2 \log^2(n)}{n}.$$

Encore une fois, on peut immédiatement déduire de cette borne que

$$\mathbb{E}_{\varepsilon^*, \mu^*} \left[\left(\frac{\hat{\varepsilon}_n}{\varepsilon^*} - 1 \right)^2 \right] \leq \frac{C_2 \log^2(n)}{n \{\varepsilon^*\}^2 \|\mu^*\|^4},$$

ce qui n'a de sens que lorsque $\sqrt{n} \varepsilon^* \|\mu^*\|^2 \rightarrow +\infty$ quand $n \rightarrow +\infty$. Soulignons que dans le cas particulier où ε^* et μ^* sont fixes (par rapport à n), ces quantités peuvent être estimées à la vitesse paramétrique classique $1/\sqrt{n}$ (à un terme logarithmique près).

Remarque 3.1. Les bornes supérieures obtenues dans les Théorèmes 3.2.1 et 3.2.2 impliquent un terme $\log^2(n)$. Ce terme logarithmique provient de l'inégalité oracle du Théorème 2.1 et est lié à la complexité de l'ensemble $\mathcal{M}_{\Lambda, \mathfrak{M}}$ sur lequel notre contraste est minimisé. Comme nous le verrons dans la Section 3.3, un tel terme est absent des bornes inférieures. Cet écart logarithmique entre les bornes inférieure et supérieure est un résultat classique lorsqu'il s'agit d'estimateurs de minimisation de contraste.

Remarque 3.2. Un point essentiel pour établir le Théorème 3.2.1 à partir du Théorème 3.1.1 est d'obtenir une minoration de $\|\hat{f}_n - f^*\|_2^2$ en fonction des paramètres (ε^*, μ^*) et $(\hat{\varepsilon}_n, \hat{\mu}_n)$. L'inégalité suivante, pour une constante $c > 0$,

$$\|\hat{f}_n - f^*\|_2^2 \geq c (\varepsilon^* - \hat{\varepsilon}_n)^2 \|\hat{\mu}_n\|^2 \|\mu^*\|^2 + c \{\varepsilon^*\}^2 \|\mu^*\|^2 \|\hat{\mu}_n - \mu^*\|^2 \quad (3.8)$$

a pu être établie à l'aide d'un raffinement technique de l'inégalité de Cauchy-Schwarz (prouvée dans la prépublication associée à [GKMM20]). Pour la preuve du Théorème 3.2.2, il faut assurer que $\hat{\mu}_n$ est proche de μ^* avec grande probabilité pour isoler l'estimation de ε^* .

3.3 Bornes inférieures

Dans cette section, on présente les bornes inférieures sur l'estimation de (ε^*, μ^*) établies dans [GKMM20], où l'on a divisé notre étude en deux cas :

Contamination forte : $\|\mu^*\|$ est minoré par une constante indépendante de n , ainsi la contamination n'est pas négligeable lorsque $n \rightarrow +\infty$.

Contamination faible : $\|\mu^*\|$ n'est pas minoré et peut donc tendre vers 0 lorsque $n \rightarrow +\infty$.

Dans le cas de la contamination forte, on considère le sous-ensemble de paramètres suivant : soit m et c deux constantes positives et

$$\Theta_n(m, c) := \left\{ \theta = (\varepsilon, \mu) : \frac{c}{\|\mu\|^2 \sqrt{n}} \leq \varepsilon, m \leq \|\mu\| \right\}.$$

Notons que la contamination n'est pas négligeable lorsque $n \rightarrow +\infty$ mais cela permet tout de même un faible effet de contamination puisque ε^* peut être de l'ordre de $n^{-1/2}$. On a alors établi les bornes inférieures suivantes.

Théorème 3.3.1. *Contamination forte*

Soit deux constantes strictement positives m et c telles que $0 < \frac{c}{m^2 \sqrt{n}} < 1$ afin que $\Theta_n(m, c)$ soit non vide. Une densité ϕ qui satisfait (\mathbf{H}_S) et $(\mathbf{H}_{\text{Lip}})$ existe telle que :

(i) il existe une constante $C_1 > 0$ telle que

$$\inf_{(\hat{\varepsilon}, \hat{\mu})} \sup_{(\varepsilon, \mu) \in \Theta_n(m, c)} \mathbb{E}[\varepsilon^2 \|\hat{\mu} - \mu\|^2] \geq \frac{C_1}{n}, \quad (3.9)$$

(ii) il existe une constante $C_2 > 0$ telle que

$$\inf_{(\hat{\varepsilon}, \hat{\mu})} \sup_{(\varepsilon, \mu) \in \Theta_n(m, c)} \mathbb{E}[(\hat{\varepsilon} - \varepsilon)^2] \geq \frac{C_2}{n}, \quad (3.10)$$

où l'infimum est pris sur tous les estimateurs $\hat{\theta} = (\hat{\varepsilon}, \hat{\mu})$ dans (3.9) et (3.10). Les constantes C_1 et C_2 dépendent de c , m et \mathcal{J} (défini dans $(\mathbf{H}_{\text{Lip}})$).

Dans ce cas, on obtient les bornes inférieures qui correspondent (à un terme logarithmique près) aux bornes supérieures obtenues dans les Théorèmes 3.2.1 et 3.2.2.

Pour la contamination faible, on s'intéresse au sous-ensemble de paramètres suivant : Soit $c > 0$, et

$$\Theta_n(c) := \left\{ \theta = (\varepsilon, \mu) : \frac{c}{\|\mu\|^2 \sqrt{n}} \leq \varepsilon \right\}.$$

On a alors le résultat suivant.

Théorème 3.3.2. *Contamination faible*

Il existe un entier $N > 0$ et une densité ϕ qui satisfait $(\mathbf{H}_{\mathcal{S}})$ et $(\mathbf{H}_{\mathbf{D}})$ tels que, pour tout $n > N$:

(i) une constante $C_3 > 0$ existe telle que :

$$\inf_{(\hat{\varepsilon}, \hat{\mu})} \sup_{(\varepsilon, \mu) \in \Theta_n(c)} \mathbb{E}[\|\mu\|^4 (\varepsilon - \hat{\varepsilon})^2] \geq \frac{C_3}{n}, \quad (3.11)$$

(ii) une constante $C_4 > 0$ existe telle que :

$$\inf_{(\hat{\varepsilon}, \hat{\mu})} \sup_{(\varepsilon, \mu) \in \Theta_n(c)} \mathbb{E}[\varepsilon^2 \|\mu\|^2 \|\mu - \hat{\mu}\|^2] \geq \frac{C_4}{n}, \quad (3.12)$$

où l'infimum est pris sur tous les estimateurs $\hat{\theta} = (\hat{\varepsilon}, \hat{\mu})$ dans (3.11) et (3.12). Les constantes C_3 et C_4 dépendent de c et \mathcal{I}_ϕ (défini dans $(\mathbf{H}_{\mathbf{D}})$).

On peut remarquer que l'hypothèse $(\mathbf{H}_{\mathbf{D}})$ ne sert que pour établir les bornes inférieures dans le cas de la contamination faible. La preuve des Théorèmes 3.3.1 et 3.3.2 s'appuie sur une généralisation de la méthode de Le Cam pour une perte asymétrique (voir Lemme B.1 dans [GKMM20]). Ce résultat est ensuite appliqué à $\rho(\theta_1, \theta_2) = \varepsilon_1 \|\mu_1 - \mu_2\|$ pour prouver le Théorème 3.3.1 et $\rho(\theta_1, \theta_2) = |\varepsilon_1 - \varepsilon_2| \|\mu_1\|$ pour le Théorème 3.3.2. Dans ces résultats, on peut constater la relation différente entre ε , μ et n dans chacun des cas. On peut également remarquer qu'estimer μ lorsque ε devient négligeable par rapport à $n^{-1/2}$ semble être impossible comme indiqué dans (ii) du Théorème 3.3.2.

3.4 Lien entre Wasserstein et $\|\cdot\|_2$

Pour conclure ce chapitre, on peut faire un lien intéressant entre la perte \mathbb{L}^2 et la perte de Wasserstein pour notre modèle de mélange contaminé. La distance de Wasserstein (L^p) entre deux mesures de probabilités m_1 et m_2 sur Ω est définie par

$$W_p(m_1, m_2) := \left[\inf_{\pi \in \Pi(m_1, m_2)} \int d^p(x, y) d\pi(x, y) \right]^{\frac{1}{p}}, \quad (3.13)$$

$\Pi(m_1, m_2)$ étant l'ensemble des mesures de probabilité sur $\Omega \times \Omega$ telles que leurs marginales sont m_1 et m_2 . On note également $G_{\varepsilon, \mu} = (1 - \varepsilon)\delta_0 + \varepsilon\delta_\mu$ la distribution mélangeante associée à la densité $f_{\varepsilon, \mu}$, où δ_μ est le Dirac en μ .

Théorème 3.4.1. *Pour toute densité ϕ qui satisfait (\mathbf{H}_S) et $(\mathbf{H}_{\text{Lip}})$, une constante $c_\phi > 0$ existe telle que*

$$\forall(\varepsilon, \varepsilon') \in]0, 1]^2 \quad \forall(\mu, \mu') \in [-M, M]^d \quad \|f_{\varepsilon, \mu} - f_{\varepsilon', \mu'}\|_2 \geq c_\phi W_2^2(G_{\varepsilon, \mu}, G_{\varepsilon', \mu'}).$$

Ainsi, $\hat{f}_n := f_{\hat{\varepsilon}_n, \hat{\mu}_n}$ défini par (3.3) satisfait

$$\mathbb{E}_{\varepsilon^*, \mu^*} [W_2^4(G_{\hat{\varepsilon}_n, \hat{\mu}_n}, G_{\varepsilon^*, \mu^*})] \lesssim \mathbb{E} [\|\hat{f}_n - f_{\varepsilon^*, \mu^*}\|_2^2] \lesssim \frac{(\log n)^2}{n}.$$

La stratégie \mathbb{L}^2 implique donc un contrôle de la distance de Wasserstein entre la distribution de mélange estimée $G_{\hat{\varepsilon}_n, \hat{\mu}_n}$ et la vraie distribution G_{ε^*, μ^*} . La preuve du Théorème 3.4.1 s'appuie sur l'inégalité (3.8) et la majoration suivante

$$W_2^4(G_{\varepsilon, \mu}, G_{\varepsilon', \mu'}) \leq 8 [(\varepsilon' - \varepsilon)^2 \|\mu'\|^2 \|\mu\|^2 + (\varepsilon \|\mu\| + \varepsilon' \|\mu'\|)^2 \|\mu - \mu'\|^2]. \quad (3.14)$$

On s'appuie donc sur des contrôles des distances entre mélanges en passant par leurs paramètres. Par contre, une borne inférieure sur la vitesse de convergence minimax en terme de distance de Wasserstein ne peut pas être directement déduite de nos résultats présentés dans les Théorèmes 3.3.1 ou 3.3.2 en raison du manque de symétrie en (ε, μ) par rapport à $(\hat{\varepsilon}, \hat{\mu})$.

Estimation adaptative de densités par mélanges gaussiens

Résumé Dans [MM11a], on s’est intéressé à l’estimation de la densité inconnue d’un échantillon par des mélanges gaussiens. Nous avons construit un estimateur par maximum de vraisemblance dont le nombre de composantes du mélange est sélectionné via un critère pénalisé non-asymptotique. Dans ce chapitre, nous présentons les travaux publiés dans [MM13] où l’on considère une collection de densités univariées dont le logarithme est β -höldérien et qui satisfont entre autres des conditions de moment. On montre que l’estimateur proposé dans [MM11a] est adaptatif au sens minimax pour la régularité β de ces densités pour le risque Hellinger.

[MM13] Maugis-Rabusseau C. and Michel B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM : Probability and Statistics*, 17, 698–724.

[BMM12] Baudry J.-P., Maugis C. and Michel B. (2012). Slope heuristics : overview and implementation. *Statistics and Computing*, 22(2), 455–470.

[MM11b] Maugis C. and Michel B. (2011). Data-driven penalty calibration : a case study for Gaussian mixture model selection. *ESAIM : Probability and Statistics*, 15, 320–339.

[MM11a] Maugis C. and Michel B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM : Probability and Statistics*, 15, 41–68.

Sommaire

4.1. Contexte	32
4.2. Les classes de densités $\mathcal{H}(\beta, \mathcal{P})$	34
4.3. Résultat d’approximation	35
4.4. Borne inférieure	36
4.5. Estimation adaptative de densités	37
4.6. Conclusion	38

4.1 Contexte

Dans ce chapitre, on considère un n -échantillon $\underline{X} = (X_1, \dots, X_n)$ à valeurs dans \mathbb{R} , de densité inconnue f appartenant à l'ensemble \mathcal{S} de toutes les densités par rapport à la mesure de Lebesgue sur \mathbb{R} . On s'intéresse à l'estimation adaptative de f par des mélanges gaussiens, motivé par le lien entre estimation de densité par mélange et récupération d'une classification non supervisée des données. On considère donc des mélanges gaussiens unidimensionnels caractérisés par leur nombre de composantes K et leurs paramètres de moyenne bornés. Ces densités de mélanges sont organisées en une collection de modèles $(\mathcal{S}_K)_{K \in \mathcal{M}_n}$, sous-ensembles de \mathcal{S} , définis par

$$\mathcal{S}_K = \left\{ x \in \mathbb{R} \mapsto \sum_{k=1}^K \varepsilon_k \phi_\sigma(x - \mu_k); \mu_k \in [-\bar{\mu}(K), \bar{\mu}(K)], \sigma = \sigma(K); \varepsilon_k \in [0, 1], \sum_{k=1}^K \varepsilon_k = 1 \right\} \quad (4.1)$$

où ϕ est le noyau gaussien pour tout $x \in \mathbb{R}$ et $\phi_\sigma(\cdot) = \sigma^{-1} \phi(\frac{\cdot}{\sigma})$ pour tout $\sigma > 0$. Contrairement à [MM11a], les mélanges gaussiens d'un modèle \mathcal{S}_K ont ici une variance commune connue $\sigma^2(K)$. Pour un modèle \mathcal{S}_K , le nombre de paramètres libres communs à toutes les densités de ce modèle vaut $D(K) := 2K - 1$ et est appelé *la dimension* du modèle. D'un point de vue non-asymptotique (voir par ex. Massart, 2007), $\bar{\mu}(K)$, $\sigma^2(K)$ ainsi que le nombre maximum de composantes du mélange dans la collection peuvent dépendre de n . La collection de modèles est indexée par l'ensemble \mathcal{M}_n qui contrôle le nombre de composantes du mélange. Par exemple, \mathcal{M}_n peut être l'ensemble des entiers entre 2 et $K_{\max}(n)$, où K_{\max} est une fonction croissante de n .

Dans chaque modèle \mathcal{S}_K , un estimateur du maximum de vraisemblance (MLE) \hat{f}_K est obtenu en minimisant le contraste empirique

$$\gamma_n(t) = -\frac{1}{n} \sum_{i=1}^n \ln \{t(X_i)\}, \quad \forall t \in \mathcal{S}$$

La fonction de perte associée à ce contraste est la divergence de Kullback-Leibler : pour deux densités g et h de \mathcal{S} , la divergence de Kullback-Leibler est définie par

$$\text{KL}(g, h) = \int \ln \left[\frac{g(x)}{h(x)} \right] g(x) dx$$

si $g dx$ est absolument continue par rapport à $h dx$ et $+\infty$ sinon. La distance d'Hellinger entre deux densités g et h de \mathcal{S} est définie par $d_H(g, h) = \frac{1}{\sqrt{2}} \|\sqrt{g} - \sqrt{h}\|_2$.

Idéalement, on voudrait estimer la vraie densité f par \hat{f}_{K^*} (appelée *oracle*) où K^* minimise sur \mathcal{M}_n le risque de Kullback-Leibler :

$$K^* \in \underset{K \in \mathcal{M}_n}{\operatorname{argmin}} \mathbb{E}_f[\text{KL}(f, \hat{f}_K)].$$

Néanmoins, K^* et la densité associée \hat{f}_{K^*} sont inconnues car elles dépendent de la vraie densité f . On sélectionne donc \hat{K} qui minimise sur \mathcal{M}_n le critère pénalisé

$$\text{crit}(K) = \gamma_n(\hat{f}_K) + \text{pen}(K)$$

où la fonction de pénalité $\text{pen} : K \in \mathcal{M}_n \mapsto \text{pen}(K) \in \mathbb{R}^+$ doit être choisie telle que le risque de Kullback-Leibler $\mathbb{E}_f[\text{KL}(f, \hat{f}_{\hat{K}})]$ de $\hat{f}_{\hat{K}}$ est proche du risque oracle $\mathbb{E}_f[\text{KL}(f, \hat{f}_{K^*})]$. Dans

[MM11a], on s'est intéressé à la construction d'une telle fonction de pénalité. On rappelle ici le résultat établi, adapté à notre collection de mélanges gaussiens unidimensionnels définis par (4.1).

Théorème 4.1.1. *Il existe quatre constantes κ , C , c_1 et c_2 telles que, si*

$$\text{pen}(K) \geq \kappa \frac{D(K)}{n} \left\{ 1 + 2\mathcal{A}^2 + \ln \left(\frac{1}{1 \wedge \frac{D(K)}{n} \mathcal{A}^2} \right) \right\}$$

où

$$\mathcal{A} = c_2 + \sqrt{\ln \left(\frac{c_1 \bar{\mu}(K)}{\sigma(K)} \right)}, \quad (4.2)$$

alors le modèle \hat{K} qui minimise $\text{crit}(K) = \gamma_n(\hat{f}_K) + \text{pen}(K)$ sur \mathcal{M}_n existe et

$$\mathbb{E} \left[d_H^2(f, \hat{f}_{\hat{K}}) \right] \leq C \left[\inf_{K \in \mathcal{M}_n} \{ \text{KL}(f, \mathcal{S}_K) + \text{pen}(K) \} + \frac{1}{n} \right]. \quad (4.3)$$

Ce résultat est issu d'un résultat plus général établi dans [MM11a] dans le cadre de la classification non supervisée par mélanges gaussiens de données multivariées avec une problématique de sélection des variables. La méthode a été implémentée et testée en pratique dans [MM11b]. Elle consiste à déterminer l'estimateur \hat{f}_K en utilisant un algorithme EM pour chaque modèle \mathcal{S}_K . Ensuite, le critère pénalisé permet de sélectionner $\hat{f}_{\hat{K}}$. Puisque ce critère pénalisé dépend d'une constante inconnue κ , cette dernière est calibrée en utilisant une méthode d'heuristique de pente (voir [BMM12] pour plus de détails).

L'estimation adaptative minimax a fait l'objet d'une étude intensive dans le domaine des statistiques non paramétriques (voir par exemple Tsybakov, 2009; Massart, 2007). Un critère d'optimalité naturel est le risque minimax, introduit par Wolfowitz (1950). Ici le risque Hellinger minimax sur une classe de densités \mathcal{H}_β (que l'on précisera par la suite) est défini par

$$\mathcal{R}_n(\mathcal{H}_\beta) = \inf_{\tilde{f}_n} \mathcal{R}(\tilde{f}_n, \mathcal{H}_\beta) \text{ avec } \mathcal{R}(\tilde{f}_n, \mathcal{H}_\beta) := \sup_{f \in \mathcal{H}_\beta} \mathbb{E}_f[d_H^2(f, \tilde{f}_n)]$$

où l'infimum est pris sur tous les estimateurs \tilde{f}_n de f . Un estimateur \tilde{f}_n est dit minimax sur \mathcal{H}_β si son risque maximal sur \mathcal{H}_β , $\mathcal{R}(\tilde{f}_n, \mathcal{H}_\beta)$, atteint le risque minimax sur cette classe de densités. Considérons maintenant une collection $(\mathcal{H}_\beta)_{\beta \in \mathcal{B}}$ de classes de densités indexées par un ensemble \mathcal{B} de paramètres de régularité β . Un estimateur est dit *adaptatif minimax* s'il atteint le risque minimax sur \mathcal{H}_β pour tous les β de \mathcal{B} , sans utiliser la connaissance de β . Notre objectif est ici d'établir que l'estimateur $\hat{f}_{\hat{K}}$ proposé dans [MM11a], qui est un mélange gaussien, est adaptatif minimax sur une collection particulière de classes de densité de Hölder $(\mathcal{H}_\beta)_{\beta \in \mathcal{B}}$ définies par la suite.

Le lien entre la sélection de modèle et l'estimation adaptative est fait par le biais de la théorie de l'approximation. En effet, une estimation adaptative est possible uniquement pour les classes fonctionnelles \mathcal{H}_β qui peuvent être efficacement approximées par notre collection de mélanges gaussiens. La convolution est largement utilisée dans la théorie de l'approximation. Il est bien connu que la convolution d'une densité f avec des versions échelonnées ϕ_σ du noyau gaussien ϕ converge vers f (voir par exemple Cheney et Light, 2009, chapitre 20). La méthode dite de quasi-interpolation consiste à remplacer les fonctions $\phi_\sigma \star f$ par des combinaisons linéaires infinies de noyaux gaussiens mis à l'échelle et translatés (voir par ex. Cheney et

Light, 2009, chapitre 36). Dans Hangelbroek et Ron (2010), un algorithme d'approximation non linéaire reposant sur des combinaisons finies de noyaux gaussiens échelonnés et translatés est défini pour donner des résultats d'approximation en norme L_p sur certaines classes de densité. Néanmoins, tous ces résultats ne peuvent pas être appliqués directement à l'étude des capacités d'approximation des mélanges gaussiens. En effet, les coefficients de ces combinaisons linéaires ne sont pas nécessairement positifs et leur somme n'est pas contrainte d'être égale à un. De plus, les résultats d'approximation fournis par toutes ces méthodes ne sont pas donnés pour la divergence de Kullback-Leibler comme l'exige notre contexte.

La capacité d'approximation des mélanges gaussiens a également été étudiée dans des travaux bayésiens non paramétriques. Le lemme 3.1 de Ghosal *et al.* (2001) donne un résultat de discrétisation pour les mélanges gaussiens avec un contrôle en normes L_1 et L_∞ . Ghosal et van der Vaart (2007) tirent avantage de cette méthode pour approximer par des mélanges gaussiens finis des fonctions différentiables avec des conditions de régularité supplémentaires. Kruijer *et al.* (2010) ont prouvé un résultat d'approximation par des mélanges gaussiens finis pour des densités dont le logarithme est localement Hölder, au sens de la divergence de Kullback-Leibler. Dans [MM13], nous avons repris leur résultat d'approximation afin d'assurer des bornes d'approximation avec des constantes uniformes vis à vis d'une classe de densités \mathcal{H}_β que nous définissons ci-dessous. Ce résultat permet alors un contrôle du terme de biais dans l'inégalité oracle (4.3) sur ces classes de densités \mathcal{H}_β .

Dans la suite, nous allons définir les classes de densités \mathcal{H}_β considérées dans la Section 4.2. Un résultat d'approximation adapté de Kruijer *et al.* (2010) est énoncé dans la Section 4.3. Ensuite, une borne inférieure pour le risque minimax est donnée dans la Section 4.4 et la propriété adaptative de notre estimateur sur \mathcal{H}_β est abordée dans la Section 4.5.

4.2 Les classes de densités $\mathcal{H}(\beta, \mathcal{P})$

On considère des densités dont le logarithme est localement β -Hölder et qui satisfont des conditions supplémentaires sur leur queue de distribution, de moments et de monotonie. Plus précisément, pour $\beta > 0$, $r = \lfloor \beta \rfloor$ est le plus grand entier inférieur à β et $b \in \mathbb{N}$ tel que $\beta \in]2b, 2b + 2]$. Soit également \mathcal{P} l'ensemble des paramètres $\{\gamma, l^+, L, \varepsilon, C, \alpha, \xi, M\}$ où L est une fonction polynomiale sur \mathbb{R} et les autres paramètres sont des constantes positives. Nous définissons ensuite la classe $\mathcal{H}(\beta, \mathcal{P})$ de toutes les densités f satisfaisant les conditions suivantes :

1. **Régularité** $\ln(f)$ est supposé être localement β -Hölder : pour tout x et y tels que $|y - x| \leq \gamma$,

$$\left| (\ln f)^{(r)}(x) - (\ln f)^{(r)}(y) \right| \leq r! L(x) |y - x|^{\beta - r}. \quad (4.4)$$

De plus pour tout $j \in \{0, \dots, r\}$,

$$|(\ln f)^{(j)}(0)| \leq l^+. \quad (4.5)$$

2. **Moments** Les dérivées $(\ln f)^{(j)}$ pour $j = 1, \dots, r$ et la fonction polynomiale L vérifient

$$\int_{\mathbb{R}} \left| (\ln f)^{(j)}(x) \right|^{\frac{2\beta + \varepsilon}{j}} f(x) dx \leq C, \quad \int_{\mathbb{R}} |L(x)|^{2 + \frac{\varepsilon}{\beta}} f(x) dx \leq C. \quad (4.6)$$

3. **Queue de distribution** Pour tout $x \in \mathbb{R}$,

$$f(x) \leq M\phi(x). \quad (4.7)$$

4. **Monotonie** f est strictement positive, f is croissante sur $] -\infty, -\alpha[$ et décroissante sur $]\alpha, +\infty[$, et $f(x) \geq \xi$ pour tout $x \in [-\alpha, \alpha]$.

Remarque 4.1. Pour faciliter la lecture nous prenons ici une hypothèse "forte" de monotonie sur un intervalle symétrique. On peut relaxer ces hypothèses pour un intervalle non symétrique avec seulement une minoration de $f \star \phi_\sigma$ par cf pour $c > 0$.

Remarque 4.2. Ces classes de densités sont plus restrictives que celles considérées dans Kruijer *et al.* (2010). En effet, les limites supérieures dans (4.6) doivent être uniformes sur la classe de densités $\mathcal{H}(\beta, \mathcal{P})$ et nous avons également besoin de la condition supplémentaire (4.5). Ces restrictions nous permettent de contrôler la divergence de Kullback-Leibler entre une densité de $\mathcal{H}(\beta, \mathcal{P})$ et un mélange gaussien fini, uniformément sur $\mathcal{H}(\beta, \mathcal{P})$. Notez que la condition (4.7) est ici supposée sur \mathbb{R} mais elle pourrait être supposée seulement à l'extérieur d'un intervalle comme dans Kruijer *et al.* (2010).

4.3 Résultat d'approximation

Pour toute fonction f , $\Phi_\sigma f$ désigne la convolution $f \star \phi_\sigma$ et $\Gamma_\sigma f$ est le terme d'erreur $\Phi_\sigma f - f$. Comme expliqué dans Kruijer *et al.* (2010), pour une densité lisse f avec $\beta \leq 2$ et sous des hypothèses de régularité raisonnables, il est possible de définir un mélange gaussien fini \wp_σ tel que $\text{KL}(f, \wp_\sigma) = O(\sigma^{2\beta})$. L'approche habituelle consiste à discrétiser le mélange continu $\Phi_\sigma f$. Mais comme $\|f - \Phi_\sigma f\|_\infty$ reste d'ordre σ^2 lorsque $\beta > 2$, cette approche semble inefficace pour des densités plus lisses. Une stratégie alternative est proposée dans Kruijer *et al.* (2010), basée sur les convolutions successives suivantes de f : $f_0 = f$ et pour tout $j \geq 0$, $f_{j+1} = f - \Gamma_\sigma f_j$. Dans leur article, la densité est approximée par une version discrétisée du mélange continu $\Phi_\sigma f_b$ où $b \in \mathbb{N}$ est tel que $\beta \in]2b, 2b + 2]$.

Dans notre cadre, le Lemme 4 de Kruijer *et al.* (2010) ne peut pas être utilisé directement puisque la borne supérieure sur la divergence de Kullback-Leibler entre f et le mélange gaussien fini n'est pas uniforme sur $\mathcal{H}(\beta, \mathcal{P})$. Ainsi, un travail supplémentaire est nécessaire pour obtenir une version uniforme de ce résultat d'approximation. Une autre raison de revoir les résultats d'approximation donnés dans Kruijer *et al.* (2010) est que ceux-ci sont énoncés pour $\sigma \leq \bar{\sigma}$ où $\bar{\sigma}$ dépend de la densité approximée f . Ainsi, nous devons également vérifier qu'il est possible de choisir le même $\bar{\sigma}$ pour toutes les densités de $\mathcal{H}(\beta, \mathcal{P})$. La preuve du Théorème 4.3.1 consiste à suivre soigneusement la méthode proposée dans Kruijer *et al.* (2010) afin d'obtenir cette version uniforme.

Théorème 4.3.1. *Il existe une constante positive $\bar{\sigma}(\beta) < 1$ telle que pour tout $f \in \mathcal{H}(\beta, \mathcal{P})$ et pour tout $\sigma < \bar{\sigma}(\beta)$, il existe un mélange gaussien de densité \wp_σ avec moins de $G_\beta \sigma^{-1} |\ln \sigma|^{\frac{3}{2}}$ points de support, avec la même variance σ pour chaque composante et avec des moyennes appartenant à $[-\mu_\sigma, \mu_\sigma]$ où*

$$\mu_\sigma \leq \tilde{G}_\beta |\ln \sigma|^{\frac{1}{2}}$$

tel que

$$\text{KL}(f, \wp_\sigma) = \int_{\mathbb{R}} f(x) \ln \left(\frac{f(x)}{\wp_\sigma(x)} \right) dx \leq c_\beta \sigma^{2\beta} \quad (4.8)$$

où c_β est uniforme sur $\mathcal{H}(\beta, \mathcal{P})$ et continue en β . La constante $\bar{\sigma}(\beta)$ dépend seulement de $\mathcal{H}(\beta, \mathcal{P})$ et est une fonction continue de β . De plus, G_β et \tilde{G}_β sont deux constantes positives qui dépendent seulement de $\mathcal{H}(\beta, \mathcal{P})$, et sont deux fonctions croissantes de β .

4.4 Borne inférieure

Afin de prouver que notre estimateur $\hat{f}_{\hat{K}}$ est adaptatif au paramètre de lissage β , une borne inférieure du risque minimax $\mathcal{R}_n(\mathcal{H}(\beta, \mathcal{P}))$ est nécessaire. Pour tout $0 < \underline{\beta} < \bar{\beta}$, on va donc déterminer un ensemble de paramètres "suffisamment grand" $\mathcal{P}(\underline{\beta}, \bar{\beta})$ tel que pour tout $\beta \in [\underline{\beta}, \bar{\beta}]$, $\mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))$ est bien définie et établir une borne inférieure pour les classes de densités $\mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))$. Notez que dans le Théorème 4.3.1, les constantes c_β , $\bar{\sigma}(\beta)$, G_β et \tilde{G}_β ne peuvent pas être bornées uniformément pour tous les $\beta \in \mathbb{R}^+$. Néanmoins, on peut prouver que $\hat{f}_{\hat{K}}$ est minimax adaptatif sur un intervalle de régularité $[\underline{\beta}, \bar{\beta}]$.

Tout d'abord, l'ensemble de paramètres $\mathcal{P}(\underline{\beta}, \bar{\beta})$ doit être défini rigoureusement. Sa définition est assez technique puisqu'elle dépend de la façon dont la borne inférieure est prouvée. La preuve est basée sur la construction de certaines fonctions oscillantes, cette méthode est présentée par exemple dans Massart (2007) (voir section 7.5). Prenons une fonction infiniment différentiable $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ à support compact inclus dans $]\frac{1}{4}, \frac{3}{4}[$ telle que

$$\int_{\mathbb{R}} \varphi(x) dx = 0 \text{ et } \int_{\mathbb{R}} \varphi(x)^2 dx = 1.$$

On pose $A = \max_{0 \leq k \leq r+1} \|\varphi^{(k)}\|_\infty > 1$ et D est un nombre entier pair positif. Pour un entier positif $j \in \{1, \dots, D\}$, on considère la fonction

$$\begin{aligned} \varphi_j : \mathbb{R} &\rightarrow \mathbb{R} \\ x &\mapsto \frac{\xi D^{-\beta}}{A} \varphi\left(\frac{D}{\alpha}(x + \frac{\alpha}{2}) - (j-1)\right). \end{aligned}$$

Soit $\mathcal{T}(\alpha, \xi)$ l'ensemble des fonctions $\omega : \mathbb{R} \rightarrow \mathbb{R}^+$ tel que ω est croissante sur $]-\infty, -\frac{\alpha}{2}[$, décroissante sur $]\frac{\alpha}{2}, +\infty[$, $\omega(x) = 2\xi$ pour tout $x \in [-\frac{3\alpha}{4}, \frac{3\alpha}{4}]$, et $\omega(-\alpha) = \omega(\alpha) = \xi$. Soit $\tilde{\mathcal{P}} = \left\{ \frac{\alpha}{4}, \ln(2\xi), \tilde{L}, \tilde{\varepsilon}, \tilde{C}, \alpha, \xi, \tilde{M} \right\}$ un ensemble de paramètres tel que $\mathcal{T}(\alpha, \xi) \cap \mathcal{H}(\beta, \tilde{\mathcal{P}})$ est non vide. Basé sur une fonctions $\omega \in \mathcal{T}(\alpha, \xi) \cap \mathcal{H}(\beta, \tilde{\mathcal{P}})$ et les fonctions φ_j , on considère l'espace fonctionnel $\mathcal{J}(\beta, D) = \{f_\theta; \theta \in \{0, 1\}^D\}$ où pour tout $\theta \in \{0, 1\}^D$ et pour tout $x \in \mathbb{R}$,

$$f_\theta(x) = \omega(x) + \sum_{j=1}^D (2\theta_j - 1)\varphi_j(x). \quad (4.9)$$

Proposition 4.4.1. *Il existe un ensemble de paramètres $\mathcal{P}(\underline{\beta}, \bar{\beta})$ tel que pour tout $D \in \mathbb{N}^*$ et pour tout $\beta \in [\underline{\beta}, \bar{\beta}]$,*

$$\mathcal{J}(\beta, D) \subset \mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta})).$$

Remarque 4.3. Un point clé pour prouver la borne inférieure énoncée dans le théorème suivant est que l'ensemble de paramètres $\mathcal{P}(\underline{\beta}, \bar{\beta})$ ne dépend pas de D .

Théorème 4.4.2. *Soit le n -échantillon $\underline{X} = (X_1, \dots, X_n)$ de densité f par rapport à la mesure de Lebesgue sur \mathbb{R} . Pour tout $\beta \in [\underline{\beta}, \bar{\beta}]$ et tout ensemble de paramètres $\mathcal{P}(\underline{\beta}, \bar{\beta})$ donné*

par la Proposition 4.4.1, il existe une constante positive κ_β telle que

$$\mathcal{R}_n(\mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))) := \inf_{\tilde{f}} \sup_f \mathbb{E}[d_H^2(f, \tilde{f})] \geq \kappa_\beta n^{-\frac{2\beta}{2\beta+1}}$$

où le supremum (resp. l'infimum) est pris sur toutes les densités f dans $\mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))$ (resp. sur tous les estimateurs \tilde{f}).

L'idée de la preuve est qu'après avoir établi la Proposition 4.4.1, la distance de Hellinger et la divergence de Kullback-Leibler entre deux fonctions de $\mathcal{J}(\beta, D)$ sont contrôlées afin de combiner un corollaire du Lemme de Birgé et le Lemme de Varshamov-Gilbert (voir Corollaire 2.19 et Lemme 4.7 dans Massart, 2007).

4.5 Estimation adaptative de densités

Dans une approche non asymptotique de sélection de modèles, la collection de modèles peut croître avec la taille de l'échantillon n , ce qui conduit à une procédure adaptative. Comme nous l'avons déjà expliqué, les propriétés adaptatives de $\hat{f}_{\hat{K}}$ sont étudiées sur un intervalle de régularité $[\underline{\beta}, \bar{\beta}]$. Préalablement, on fixe $0 < \underline{\beta} < \bar{\beta}$ et on choisit également $a_{\bar{\beta}} > 1$ assez grand pour que

$$\frac{G_{\bar{\beta}}}{a_{\bar{\beta}}} \left(\frac{\ln a_{\bar{\beta}}}{\ln 2} + 3 \right)^{3/2} \leq 1, \quad (4.10)$$

où $G_{\bar{\beta}}$ est défini dans le Théorème 4.3.1. Les paramètres des modèles de mélange gaussiens $(\mathcal{S}_K)_{K \in \mathcal{M}_n}$ sont maintenant spécifiés dans le but d'appliquer les résultats d'approximation établis dans le Théorème 4.3.1 : on reprend la définition (4.1) avec

$$\sigma(K) := a_{\bar{\beta}} K^{-1} (\ln K)^{3/2} \text{ et } \bar{\mu}(K) := \tilde{G}_{\bar{\beta}} |\ln \sigma(K)|^{1/2}.$$

Il est naturel de supposer que le nombre de composantes du mélange K est inférieur à n et nous supposons également que les mélanges ont au moins deux composantes : $\mathcal{M}_n = \{2, \dots, n\}$.

Théorème 4.5.1. *Supposons que $n \geq 3$ et soit $\hat{f}_{\hat{K}}$ l'estimateur du maximum de vraisemblance où \hat{K} minimise le critère pénalisé défini dans le Théorème 4.1.1. Alors il existe une constante $c_{\underline{\beta}, \bar{\beta}}$ telle que pour tout $\beta \in [\underline{\beta}, \bar{\beta}]$ et pour tout $f \in \mathcal{H}(\beta, \mathcal{P}(\underline{\beta}, \bar{\beta}))$,*

$$\mathbb{E} \left[d_H^2(f, \hat{f}_{\hat{K}}) \right] \leq c_{\underline{\beta}, \bar{\beta}} (\ln n)^{\frac{5\beta}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}}.$$

Le Théorème 4.5.1 montre que l'estimateur pénalisé $\hat{f}_{\hat{K}}$ est minimax adaptatif à la régularité β des classes de densités définies dans la Section 4.2, à une puissance de $\ln(n)$ près. Ce terme logarithmique est dû à la forme de la pénalité donnée dans le Théorème 4.1.1. Il n'est pas détecté en pratique comme montré dans [MM11b] et nous pensons qu'il pourrait être supprimé de la forme de la pénalité. Notez que l'estimateur bayésien non paramétrique défini dans Kruijer *et al.* (2010) a un niveau de convergence similaire avec une plus grande puissance du terme logarithmique.

Remarque 4.4. Contrairement à [MM11a], les mélanges gaussiens considérés dans ce chapitre ont une variance commune et connue $\sigma(K)$. Un résultat analogue au Théorème 4.5.1 pour

les mélanges gaussiens à variances inconnues différentes peut être facilement énoncé. Plus précisément, considérons

$$\tilde{\mathcal{S}}_K = \left\{ x \in \mathbb{R} \mapsto \sum_{k=1}^K \varepsilon_k \phi_{\sigma_k}(x - \mu_k); \mu_k \in [-\bar{\mu}(K), \bar{\mu}(K)], \sigma_k \in [\sigma(K), \bar{\sigma}], \varepsilon_k \in [0, 1], \sum_{k=1}^K \varepsilon_k = 1 \right\}.$$

Puisque \mathcal{S}_K est inclus dans $\tilde{\mathcal{S}}_K$, les résultats d'approximation pour \mathcal{S}_K sont également valables pour $\tilde{\mathcal{S}}_K$. En partant de l'inégalité oracle donnée dans [MM11a] pour $\tilde{\mathcal{S}}_K$ (la constante \mathcal{A} étant modifiée), on en déduit alors que l'estimateur pénalisé $\hat{f}_{\tilde{K}}$, défini sur les modèles $\tilde{\mathcal{S}}_K$, est minimax adaptatif à la régularité β des classes de densités définies dans la Section 4.2. Notons que ce résultat est en fait plus faible que le Théorème 4.5.1 puisque les modèles $\tilde{\mathcal{S}}_K$ sont plus riches que les modèles \mathcal{S}_K ; on devrait obtenir une adaptation sur une classe beaucoup plus grande que Hölderienne uniforme.

4.6 Conclusion

Dans ce travail nous avons établi que l'estimateur pénalisé $\hat{f}_{\tilde{K}}$, défini dans [MM11a] pour les modèles de mélange gaussiens ayant une variance identique et connue pour chaque composante, est adaptatif à la régularité des classes de densité \mathcal{H}_β (densités univariées dont le logarithme est localement β -Hölder). La preuve a nécessité d'adapter le résultat d'approximation de Kruijer *et al.* (2010) pour contrôler le terme de biais entre les modèles de mélange gaussiens et les espaces de densité \mathcal{H}_β . Une borne inférieure pour le risque minimax sur les classes de densité \mathcal{H}_β a également été énoncée pour finalement prouver que notre estimateur atteint la vitesse minimax.

Comme indiqué en Remarque 4.4, un résultat similaire (mais plus faible) peut être donné pour l'adaptation d'estimateurs pénalisés définis sur des modèles de mélange gaussiens avec des variances hétéroscédastiques inconnues. Mais la collection de modèles étant beaucoup plus riche, on devrait obtenir l'adaptation sur une classe beaucoup plus large que la Hölder uniforme. Cela nécessite d'établir un nouveau résultat d'approximation sur une classe de fonctions à déterminer.

Dans [MM11a], l'estimateur basé sur des mélanges gaussiens, remplissant une inégalité d'oracle comme (4.3), concerne des données multivariées. Il serait intéressant d'étendre notre résultat au cas multidimensionnel mais là encore, le point clé est de déterminer une classe de densités multidimensionnelles pour laquelle on peut établir un résultat d'approximation comme dans le Théorème 4.3.1.

Estimation de mélanges par super-résolution

Résumé Dans [DGMM21], on étudie l'estimation d'une mesure de mélange discrète G^* . On introduit un programme convexe basé sur de la régularisation ℓ_1 sur l'espace des mesures pour estimer G^* sans grille à partir d'un échantillon de loi de mélange, cette méthode est appelée *Beurling-LASSO*. Nous établissons une limite inférieure sur la largeur de bande de notre terme d'ajustement des données qui ne dépend que du support de G^* et de la "séparation minimale" pour garantir un contrôle de l'erreur de localisation du support. Sous des conditions dite "de source non dégénérée", nous obtenons une propriété de stabilité du support non asymptotique. Cette dernière montre que pour un échantillon de taille n suffisamment grande, notre estimateur possède exactement autant de masses de Dirac pondérées que la cible G^* , convergeant en amplitude et en localisation vers les vraies. Les performances de cet estimateur sont étudiées en construisant un "certificat dual" adapté à notre situation. Certaines situations classiques, comme les mélanges de lois *super-smooth* (dont le cas gaussien) ou de lois *ordinary smooth* (dont la loi de Laplace), sont étudiées.

[DGMM21] De Castro Y., Gadat S., Marteau C. and Maugis-Rabusseau C. (2021). Supermix : sparse regularization for mixtures. *Annals of Statistics*, 49(3) : 1779-1809.

Sommaire

5.1. Contexte	40
5.2. Hypothèses et premiers résultats	41
5.3. Estimation via le Beurling-LASSO (BLASSO)	43
5.4. Performances statistiques de \hat{G}_n	46
5.5. Application à quelques mélanges usuels	50
5.6. Conclusion et perspectives	53

5.1 Contexte

A partir d'un échantillon $\underline{X} := (X_1, \dots, X_n) \in (\mathbb{R}^d)^n$, on souhaite estimer une loi de mélange G^* à K composantes :

$$G^* := \sum_{k=1}^K \varepsilon_k^* \delta_{\mu_k^*} \quad (5.1)$$

où l'ensemble des poids $(\varepsilon_k^*)_{1 \leq k \leq K}$ définit une loi de probabilité discrète : chaque $\delta_{\mu_k^*}$ est une masse de Dirac au point $\mu_k^* \in \mathbb{R}^d$ et

$$\sum_{k=1}^K \varepsilon_k^* = 1 \quad \text{et} \quad \forall k \in [K] := \{1, \dots, K\} : \varepsilon_k^* > 0.$$

On note $S^* := \{\mu_1^*, \dots, \mu_K^*\}$ le support de la loi G^* et la séparation minimale Δ^* de S^* est la distance minimale entre deux pics (*spikes*)

$$\Delta^* := \min_{k \neq \ell} \|\mu_k^* - \mu_\ell^*\|_2.$$

La loi G^* est indirectement observée via l'échantillon \underline{X} qui admet une densité f^* par rapport à la mesure de Lebesgue sur \mathbb{R}^d :

$$f^*(x) = \sum_{k=1}^K \varepsilon_k^* \phi(x - \mu_k^*), \quad \forall x \in \mathbb{R}^d \quad (5.2)$$

où ϕ est une densité connue.

Lorsque le nombre de composantes K est connu, l'estimateur du maximum de vraisemblance (MLE) est un candidat naturel pour f^* , souvent approximé via l'algorithme EM. Cependant, le MLE souffre de plusieurs inconvénients (voir par ex. Le Cam, 1990) dont la non-unicité de la solution et l'obtention de garanties théoriques pour l'algorithme EM reste une question difficile (voir Wu *et al.*, 1983; Balakrishnan *et al.*, 2017; Dwivedi *et al.*, 2020). Plusieurs méthodes alternatives ont été proposées dans ce contexte. Certaines contributions utilisent largement le point de vue MLE pour dériver des propriétés cohérentes dans les modèles semi-paramétriques généraux, y compris le cas gaussien (voir par ex. Van der Vaart, 1996), tandis que d'autres ont développé certaines fonctions de contraste dans un cadre semi-paramétrique : avec des hypothèses de symétrie et de nombre de composantes dans Bordes *et al.* (2006) et Butucea et Vandekerkhove (2014), ou avec un nombre fixe de composantes et un contraste L^2 dans [GKMM20]. Le cas des mélanges gaussiens a attiré une attention particulière : par exemple, une stratégie de sélection de modèle est développée dans [MM11a] et une analyse spécifique de l'algorithme EM avec deux composantes gaussiennes est fournie dans Xu *et al.* (2016). Certains des articles mentionnés précédemment fournissent de meilleurs résultats (par exemple avec des vitesses de convergence paramétriques pour l'estimation des poids ε_k^* , voir Nguyen, 2013; Heinrich et Kahn, 2018), mais sont obtenus dans des contextes plus contraints : nombre fixe connu de composantes (souvent $K = 2$), cas univarié...

Dans [DGMM21], on propose de construire un estimateur \hat{G}_n de G^* comme solution d'un problème convexe (régularisation ℓ_1 sur l'espace des mesures) et ne nécessitant pas la connaissance de K . Cet estimateur est inspiré de méthodes de "*off-the-grid super-resolution*" (Bhaskar *et al.*, 2013; Candès et Fernandez-Granda, 2014), où l'on cherche à déterminer une mesure discrète à partir de mesures linéaires. Le problème dit de "*déconvolution sparse*" s'inscrit dans ce

cadre puisqu'il s'agit d'estimer une mesure cible à partir de l'observation d'un certain produit de convolution entre la mesure cible et un noyau connu, comme f^* dans (5.2). Notez que dans notre problème, nous n'observons pas f^* mais un échantillon tiré de celle-ci, et les stratégies standard telles que (1.15) dans Candès et Fernandez-Granda (2014) ne peuvent pas être utilisées ici. Cependant, l'idée fondamentale de la construction d'un "*certificat dual*" de discrétisation dans Candès et Fernandez-Granda (2014) est l'élément clé qui a été repris dans De Castro et Gamboa (2012), Duval et Peyré (2015) et De Castro *et al.* (2017) entre autres. Des travaux récents ont abordé les modèles de mélange en supposant que la loi d'échantillonnage est connue (voir par ex. Poon *et al.*, 2019). Dans notre travail, on cherche à combler le fossé entre ces méthodes récentes de "*off-the-grid super-resolution*" et une procédure d'estimation à partir d'un n -échantillon.

Dans la suite, on va tout d'abord redonner des ingrédients concernant les méthodes de régularisation ℓ_1 et donner un résultat déterministe préliminaire dans un contexte "idéal" (estimation exacte de G^* à partir de f^*) afin de mieux comprendre les difficultés du problème et les leviers. On présentera ensuite notre estimateur \hat{G}_n basé sur une déconvolution avec une stratégie Beurling-LASSO (BLASSO) (voir par ex. De Castro et Gamboa, 2012). Un point clé est la construction d'un *certificat dual* adapté à notre problème. On étudiera les performances statistiques de \hat{G}_n à retrouver la vraie mesure de mélange G^* à partir d'un n échantillon. On finira par interpréter ces résultats pour des mélanges de lois *super-smooth* (dont les lois gaussiennes) et les lois *ordinary smooth* (comme la loi de Laplace).

5.2 Hypothèses et premiers résultats

5.2.1 Hypothèses et notations

On va ici introduire quelques notations nécessaires pour la suite et préciser les hypothèses sur la densité connue ϕ .

On note tout d'abord $(\mathcal{M}(\mathbb{R}^d, \mathbb{R}), \|\cdot\|_1)$ l'ensemble des mesures de \mathbb{R}^d muni de la norme en variation totale $\|\cdot\|_1$ définie par

$$\|G\|_1 := \int_{\mathbb{R}^d} d|G| \quad \forall G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}),$$

où $|G| = G^+ + G^-$ et $G = G^+ - G^-$ est la décomposition de Jordan associée à la mesure $G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$.

On note \mathcal{F} la transformée de Fourier définie par

$$\forall x \in \mathbb{R}^d, \forall f \in L^1(\mathbb{R}^d), \quad \mathcal{F}[f](x) := \int_{\mathbb{R}^d} e^{-ix^\top \omega} f(\omega) d\omega$$

et étendue à $\mathcal{M}(\mathbb{R}^d, \mathbb{R})$ par

$$\forall x \in \mathbb{R}^d, \forall G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}), \quad \mathcal{F}[G](x) := \int_{\mathbb{R}^d} e^{-ix^\top \omega} dG(\omega).$$

L'ensemble des fonctions continues à valeurs réelles qui tendent vers 0 à l'infini sur \mathbb{R}^d est noté $\mathcal{C}_0(\mathbb{R}^d, \mathbb{R})$ et on rappelle que $\mathcal{F}(L^1(\mathbb{R}^d))$ est un sous-ensemble dense de $\mathcal{C}_0(\mathbb{R}^d, \mathbb{R})$.

On introduit également l'opérateur de convolution Φ

$$G \mapsto \Phi G := \phi \star G = \int_{\mathbb{R}^d} \phi(\cdot - x) dG(x), \quad G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}). \quad (5.3)$$

Concernant la densité $\phi \in L^1(\mathbb{R}^d)$, on fait l'hypothèse suivante :

$$\text{La fonction } \phi \text{ est symétrique continue bornée définie positive.} \quad (\mathcal{H}_0)$$

L'hypothèse (\mathcal{H}_0) est peu limitante, elle est vérifiée en outre pour les lois gaussiennes, de Laplace ou de Cauchy.

Dans la suite, on considère $h : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ la fonction définie par $h(x, y) = \phi(x - y)$ pour tout $x, y \in \mathbb{R}^d$. L'hypothèse (\mathcal{H}_0) implique que $h(\cdot, \cdot)$ est un noyau défini positif continu symétrique. D'après le théorème de Bochner (voir par ex. le Théorème 11.32 de Rudin, 1991), ϕ est la transformée de Fourier inverse d'une mesure spectrale Σ admettant une densité positive σ par rapport à la mesure de Lebesgue sur \mathbb{R}^d telle que $\sigma \in L^1(\mathbb{R}^d)$. Par la suite l'ensemble des points où la transformée de Fourier d'une fonction ne s'annule pas jouera un rôle important. Ce support est noté $\text{Supp}(\sigma)$:

$$\text{Supp}(\sigma) = \left\{ \omega \in \mathbb{R}^d : \sigma(\omega) \neq 0 \right\}.$$

5.2.2 Estimation exacte de G^* à partir de f^*

Nous allons ici considérer un problème "idéal" où l'on cherche à estimer la mesure G^* non pas à partir de l'échantillon \underline{X} mais à partir de la loi f^* elle-même. Bien sûr, cette situation ne se produit pas en pratique puisque nous n'observons pas f^* mais ce problème jouet fournit déjà les premiers ingrédients pour une meilleure compréhension des difficultés qui se présentent dans notre contexte. Comme $f^* := \Phi G^*$, on cherche à retrouver G^* à partir de sa convolution par ϕ . Nous sommes donc confrontés à un problème inverse (déconvolution) que l'on cherche ici à résoudre par une approche de régularisation ℓ_1 :

$$\min_{G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}) : \Phi G = f^*} \|G\|_1. \quad (5.4)$$

En particulier, on recherche les conditions pour lesquelles l'ensemble des solutions de (5.4) est le singleton $\{G^*\}$ ("*Perfect Recovery*"). Nous introduisons l'ensemble des points admissibles du programme (5.4) :

$$\mathcal{M}(f^*) := \{G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}) : \Phi G = f^*\}.$$

Une première situation est celle où le support de la densité spectrale $\sigma = \mathcal{F}(\phi)$ est \mathbb{R}^d :

$$\phi = \mathcal{F}^{-1}[\sigma], \sigma(\omega) = \sigma(-\omega) \text{ a.e. avec } \text{Supp}(\sigma) = \mathbb{R}^d : \forall \omega \in \mathbb{R}^d \quad \sigma(\omega) > 0. \quad (\mathcal{H}_\infty)$$

L'ensemble des densités ϕ qui satisfont les hypothèses (\mathcal{H}_0) et (\mathcal{H}_∞) comprend la densité gaussienne, de Laplace, $B_{2\ell+1}$ -spline, ... (voir par exemple Sriperumbudur *et al.*, 2011, page 2397). Sous ces hypothèses (\mathcal{H}_0) et (\mathcal{H}_∞) , on a la propriété suivante de "*Perfect Recovery*".

Théorème 5.2.1 ("*Perfect Recovery*" sous (\mathcal{H}_0) et (\mathcal{H}_∞)). *Supposons que le noyau de convolution ϕ satisfait les hypothèses (\mathcal{H}_0) et (\mathcal{H}_∞) , alors pour tout G^* , $\mathcal{M}(f^*) = \{G^*\}$.*

Par la théorie de la super-résolution (popularisée par Candès et Fernandez-Granda, 2014, en outre), l'hypothèse (\mathcal{H}_∞) peut être relaxée et la reconstruction peut être obtenue au prix d'une hypothèse sur la transformée de Fourier de ϕ . Nous supposons ici que la densité spectrale σ a un support qui contient l'hypercube $[-\eta, \eta]^d$ pour un certain seuil de fréquence $\eta > 0$:

$$\phi = \mathcal{F}^{-1}[\sigma], \sigma(\omega) = \sigma(-\omega) \text{ presque partout avec } [-\eta, \eta]^d \subset \text{Supp}(\sigma). \quad (\mathcal{H}_\eta)$$

Les densités ϕ qui satisfont (\mathcal{H}_η) et pour lesquelles $\text{Supp}(\sigma) = [-\eta, \eta]^d$ agissent comme des filtres passe-bas. L'opérateur de convolution Φ décrit dans (5.3) annule toutes les fréquences supérieures à η . Bien entendu, plus la valeur de η est grande, plus le problème de déconvolution inverse est facile à résoudre.

Sous (\mathcal{H}_0) et (\mathcal{H}_η) , la mesure cible $G^* \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$ n'est pas le seul point admissible dans $\mathcal{M}(f^*)$. Nous devons nous assurer de l'existence d'une fonction spécifique, appelée *certificat dual*, qui entraînera que G^* est toujours la seule solution du problème (5.4).

Théorème 5.2.2 (Certificat Dual pour (5.4)). *Supposons que la densité ϕ satisfait (\mathcal{H}_0) et (\mathcal{H}_η) pour un $\eta > 0$. Supposons que G^* et $S^* = \{\mu_1^*, \dots, \mu_K^*\}$ sont donnés par l'Equation (5.1) et qu'une fonction \mathcal{P}_η existe, satisfaisant les conditions d'interpolation*

$$\forall t \in S^* : \mathcal{P}_\eta(t) = 1 \quad \text{et} \quad \forall t \notin S^* : |\mathcal{P}_\eta(t)| < 1,$$

et les conditions de régularité

- $\mathcal{P}_\eta \in \mathcal{C}_0(\mathbb{R}^d, \mathbb{R}) \cap L^1(\mathbb{R}^d)$,
- $\text{Supp}(\mathcal{F}[\mathcal{P}_\eta]) \subset [-\eta, \eta]^d$.

Alors le problème (5.4) a pour unique solution G^* .

Un des challenges majeurs dans [DGMM21] est la construction d'un tel *certificat dual* \mathcal{P}_η dans notre problème statistique.

5.3 Estimation via le Beurling-LASSO (BLASSO)

Dans cette section, nous revenons à notre problème initial où la densité f^* n'est pas disponible, on observe un échantillon $\underline{X} = (X_1, \dots, X_n)$ i.i.d. distribué selon la densité f^* . Dans ce contexte, seule la mesure empirique

$$\hat{f}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}, \tag{5.5}$$

est disponible, et notre objectif est d'estimer G^* à partir de \hat{f}_n . Nous considérons une procédure *BLASSO* (voir par ex. Azaïs *et al.*, 2015). La mesure discrète inconnue G^* est estimée par \hat{G}_n , défini par

$$\hat{G}_n := \underset{G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})}{\text{argmin}} \{C(\Phi G, \hat{f}_n) + \kappa \|G\|_1\}, \tag{5.6}$$

où κ est le paramètre de régularisation (dont la valeur sera précisée plus tard), et $C(\Phi G, \hat{f}_n)$ est un terme de *fidélité aux données* (dépendant de l'échantillon \underline{X}) qui mesure la "distance" entre la cible G^* et toute mesure $G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$.

5.3.1 Approche par noyau

La difficulté dans la conception de $C(\Phi G, \hat{f}_n)$ réside dans le fait que la loi empirique \hat{f}_n n'appartient pas à $\mathcal{C}_0(\mathbb{R}^d, \mathbb{R})$. Pour comparer ΦG avec \hat{f}_n , nous avons besoin d'intégrer ces quantités dans un même espace. Nous considérons ici une approche par noyau pour profiter des propriétés d'une structure d'un espace de Hilbert à noyau reproduisant (RKHS) (voir par ex. Sriperumbudur *et al.*, 2011). Nous rappelons brièvement la définition d'un tel espace.

Définition 5.3.1. Soit $(\mathbb{L}, \|\cdot\|_{\mathbb{L}})$ un espace de Hilbert contenant une fonction de \mathbb{R}^d à \mathbb{R} . On dit que l'espace \mathbb{L} est un RKHS si $L_x : f \mapsto f(x)$ sont continues pour tout $x \in \mathbb{R}^d$ de $(\mathbb{L}, \|\cdot\|_{\mathbb{L}})$ à $(\mathbb{R}, |\cdot|)$.

Le théorème de représentation de Riesz conduit à l'existence d'une fonction ℓ qui satisfait :

$$\langle f, \ell(x, \cdot) \rangle_{\mathbb{L}} = f(x) \quad \forall f \in \mathbb{L}, \quad \forall x \in \mathbb{R}^d. \quad (5.7)$$

La fonction ℓ est appelée le *noyau reproduisant* associé à \mathbb{L} . Ci-dessous, nous considérons un noyau λ tel que $\ell(x, y) = \lambda(x - y)$ pour tout $x, y \in \mathbb{R}^d$ où λ satisfait (\mathcal{H}_0) . De nouveau, le théorème de Bochner donne l'existence d'une mesure positive $\Lambda \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$ telle que λ est sa transformée de Fourier inverse $\lambda = \mathcal{F}^{-1}(\Lambda)$. De plus, puisque λ est continu, Λ est alors une mesure bornée.

Enfin un opérateur de convolution L , similaire à celui défini dans l'Equation (5.3), peut être associé au RKHS :

$$\forall \nu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}), L\nu = \lambda \star \nu \in \mathbb{L}.$$

Finalement, comme pour tout $G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$, $L\hat{f}_n$ et $L \circ \Phi G$ appartiennent à \mathbb{L} , on utilise comme critère

$$C_{\lambda}(\Phi G, \hat{f}_n) := \|L\hat{f}_n - L \circ \Phi G\|_{\mathbb{L}}^2, \quad \forall G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}). \quad (5.8)$$

Exemple 5.3.2 (Noyau sinus-cardinal sinc).

Se donnant une coupure des fréquences $1/\tau > 0$, on peut considérer le noyau

$$\lambda_{\tau}(x) := \frac{1}{\tau^d} \lambda_{\text{sinc}}\left(\frac{x}{\tau}\right) \quad \text{où} \quad \lambda_{\text{sinc}}(x) := \prod_{j=1}^d \frac{\sin(\pi x_j)}{\pi x_j} \quad \forall x \in \mathbb{R}^d.$$

Alors, la mesure spectrale est donnée par

$$d\Lambda_{\tau}(\omega) = d\Lambda_{\text{sinc}}(\omega\pi\tau) := \frac{1}{2^d} \prod_{j=1}^d \mathbb{1}_{[-1/\tau, 1/\tau]}(\omega_j) d\omega, \quad \forall \omega \in \mathbb{R}^d.$$

Dans ce cas particulier, nous déduisons que la convolution L est un filtre passe-bas avec une fréquence de coupure de $1/\tau$ et le RKHS (noté \mathbb{L}_{τ}) est donné par :

$$\mathbb{L}_{\tau} = \left\{ f \text{ s.t. } \|f\|_{\mathbb{L}_{\tau}}^2 = \frac{1}{2^d} \int_{B_{\infty}(1/\tau)} |\mathcal{F}[f]|^2 < +\infty \text{ et } \text{Supp}(\mathcal{F}[f]) \subseteq B_{\infty}(1/\tau) \right\}, \quad (5.9)$$

où $B_{\infty}(1/\tau)$ désigne la boule centrée ℓ_{∞} de rayon $1/\tau$. Dans ce contexte, le critère devient

$$C_{\lambda_{\tau}}(\Phi G, \hat{f}_n) = \frac{1}{2^d} \int_{\mathbb{R}^d} \left| \lambda_{\tau} \star (\Phi G - \hat{f}_n)(x) \right|^2 dx.$$

Cette perte se concentre sur l'erreur L^2 de $\Phi G - \hat{f}_n$ pour les fréquences dans le domaine de Fourier $[-1/\tau, 1/\tau]^d$. Dans un certain sens, l'estimateur à noyau $\lambda_{\tau} \star \hat{f}_n$ a une largeur de bande τ qui empêchera le sur-ajustement.

5.3.2 Estimation par programme convexe

Notre estimateur \hat{G}_n est donc défini comme une solution du problème d'optimisation suivant

$$\inf_{G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})} \left\{ \frac{1}{2} \|L\hat{f}_n - L \circ \Phi G\|_{\mathbb{L}}^2 + \kappa \|G\|_1 \right\}, \quad (\mathbf{P}_\kappa)$$

où $\|\cdot\|_{\mathbb{L}}$ est la norme associée au RKHS généré par λ (voir Section 5.3.1) et $\kappa > 0$ est un paramètre dont la valeur sera réglée plus tard.

Dans Candès et Fernandez-Granda (2014), la super-résolution permet de retrouver une mesure discrète sur le tore à partir de certains coefficients de Fourier alors que nous voulons ici retrouver une mesure discrète sur \mathbb{R}^d à partir d'une transformée de Fourier sur \mathbb{R}^d . Par conséquent, de nouvelles garanties théoriques sont nécessaires afin de définir correctement l'estimateur \hat{G}_n . Dans cette optique, nous considérons les variables primales $G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$ et $z \in \mathbb{L}$ et introduisons la variable duale $c \in \mathbb{L}$ ainsi que le lagrangien suivant :

$$\mathcal{L}(G, z, c) := \frac{1}{2} \|L\hat{f}_n - z\|_{\mathbb{L}}^2 + \kappa \|G\|_1 - \langle c, L \circ \Phi G - z \rangle_{\mathbb{L}}. \quad (5.10)$$

Il est immédiat de vérifier que si $z \neq L \circ \Phi G$, alors le supremum de $\mathcal{L}(G, z, c)$ sur c est $+\infty$. Par conséquent, l'expression primale coïncide avec le supremum dans les variables duales, à savoir

$$\inf_{G, z} \sup_c \mathcal{L}(G, z, c) = \inf_{G \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})} \left\{ \frac{1}{2} \|L\hat{f}_n - L \circ \Phi G\|_{\mathbb{L}}^2 + \kappa \|G\|_1 \right\} \iff (\mathbf{P}_\kappa).$$

Le programme dual de (\mathbf{P}_κ) est donné par

$$\sup_{c \in \mathbb{L}} \inf_{(G, z) \in \mathcal{M}(\mathbb{R}^d, \mathbb{R}) \times \mathbb{L}} \mathcal{L}(G, z, c). \quad (\mathbf{P}_\kappa^*)$$

Théorème 5.3.3. *Les affirmations suivantes sont vraies.*

i) *Le problème primal (\mathbf{P}_κ) a au moins une solution et*

$$\hat{z}_n := L \circ \Phi \hat{G}_n \quad \text{et} \quad \hat{m}_n := \|\hat{G}_n\|_1 \quad \text{sont définis de façon unique,}$$

donc ils ne dépendent pas du choix de la solution \hat{G}_n .

ii) *Le problème dual (\mathbf{P}_κ^*) de (\mathbf{P}_κ) satisfait*

$$\frac{\|L\hat{f}_n\|_{\mathbb{L}}^2}{2} - \inf \left\{ \frac{1}{2} \|L\hat{f}_n - c\|_{\mathbb{L}}^2 : c \in \mathbb{L} \text{ s.t. } \|\Phi c\|_{\infty} \leq \kappa \right\} \iff (\mathbf{P}_\kappa^*),$$

et il n'y a pas de gap de dualité. De plus, il a une solution unique

$$\hat{c}_n = L\hat{f}_n - \hat{z}_n.$$

iii) *Une solution \hat{G}_n de (\mathbf{P}_κ) satisfait*

$$\text{Supp}(\hat{G}_n) \subseteq \left\{ x \in \mathbb{R}^d : |\hat{\eta}_n|(x) = 1 \right\} \quad \text{et} \quad \int_{\mathbb{R}^d} \hat{\eta}_n d\hat{G}_n = \hat{m}_n,$$

où

$$\hat{\eta}_n := \frac{\Phi \hat{c}_n}{\kappa} = \frac{1}{\kappa} \Phi(L\hat{f}_n - z_n). \quad (5.11)$$

iv) Si $d = 1$ et si au moins une des mesures spectrales Λ ou σ a un support borné, alors $\{x \in \mathbb{R} : |\hat{\eta}_n|(x) = 1\}$ est discret avec aucun point d'accumulation, toute solution primale \hat{G}_n a un support discret (au plus dénombrable) $\hat{S} \subset \mathbb{R}$ avec aucun point d'accumulation

$$\hat{G}_n = \sum_{\mu \in \hat{S}} \hat{\varepsilon}_\mu \delta_\mu. \quad (5.12)$$

Il est généralement admis numériquement (voir par ex. Candès et Fernandez-Granda, 2014, page 939), que les extrema du polynôme dual $\hat{\eta}_n = \Phi \hat{c}_n$ sont situés dans un ensemble discret, de sorte que toute solution de (\mathbf{P}_κ) a un support discret en utilisant *iii*). Cependant, ce problème reste une question ouverte. En pratique, tous les solveurs de (\mathbf{P}_κ) conduisent à des solutions discrètes : les méthodes gloutonnes sont discrètes par construction, et les méthodes de régularisation ℓ_1 conduisent empiriquement à des solutions discrètes (voir par ex. Candès et Fernandez-Granda, 2014).

Exemple 5.3.4. Le filtre passe-bas défini dans l'Exemple 5.3.2 satisfait aux exigences du *iv*) du Théorème 5.3.3 : on en déduit que lorsque $d = 1$, toutes les solutions \hat{G}_n sont de la forme (5.12).

5.4 Performances statistiques de \hat{G}_n

Cette section fournit quelques résultats théoriques pour \hat{G}_n , construit comme la solution de (\mathbf{P}_κ) . L'analyse statistique est effectuée en utilisant un *certificat dual* \mathcal{P}_m comme dans la super-résolution. Mais le défi est de construire un certificat dual \mathcal{P}_m sur \mathbb{R}^d (et non sur le tore de dimension d comme dans Candès et Fernandez-Granda, 2014) et dont le support des fréquences (à savoir $4m$ dans la suite) s'adapte correctement à la taille de l'échantillon n .

5.4.1 Construction du certificat dual

Pour tout $m \in \mathbb{N}^*$, on considère la fonction $p_m^{\alpha, \beta}$ paramétrée par un vecteur α et une matrice β de coefficients, définie par :

$$p_m^{\alpha, \beta}(t) = \sum_{k=1}^K \{\alpha_k \psi_m(t - t_k) + \langle \beta_k, \nabla \psi_m(t - t_k) \rangle\}, \quad \forall t \in \mathbb{R}^d, \quad (5.13)$$

où $\alpha = (\alpha_1, \dots, \alpha_K)^T$, $\beta = (\beta_k^i)_{1 \leq k \leq K, 1 \leq i \leq d}$ avec

$$\psi_m(\cdot) = \psi^4(m \cdot) \text{ et } \forall u = (u^1, \dots, u^d) \in \mathbb{R}^d \quad \psi(u) = \prod_{j=1}^d \text{sinc}(u^j) \text{ et } \text{sinc}(x) = \frac{\sin(x)}{x}. \quad (5.14)$$

A partir de cette fonction $p_m^{\alpha, \beta}$, nous construisons le *certificat dual* et énonçons ses propriétés dans le théorème suivant.

Théorème 5.4.1 (Certificat dual). *Se donnant un ensemble de K points $S^* = \{\mu_1^*, \dots, \mu_K^*\}$ dans \mathbb{R}^d avec $\Delta^* := \min_{k \neq \ell} \|\mu_k^* - \mu_\ell^*\|_2$ and $\Delta^*_+ = \min(\Delta^*, 1)$. Alors les propriétés suivantes sont vérifiées*

- *i)* Il existe une fonction \mathcal{P}_m définie par $\mathcal{P}_m(t) = [p_m^{\alpha,\beta}(t)]^2$ avec $m \gtrsim \sqrt{K}d^{3/2}/\Delta_+^*$ telle que

$$\forall k \in [K], \mathcal{P}_m(\mu_k^*) = 1, 0 \leq \mathcal{P}_m \leq 1, \text{ et } \mathcal{P}_m(t) = 1 \iff t \in S^*.$$

- *ii)* Un couple de constantes universelles (v, γ) indépendantes de n, m et d existe tel que pour $\epsilon = \frac{v}{md}$:

– "Near region" : Si on définit

$$\mathbf{N}(\epsilon) := \bigcup_{k=1}^K \mathbf{N}_k(\epsilon) \text{ où } \mathbf{N}_k(\epsilon) := \{t : \|t - \mu_k^*\|_2 \leq \epsilon\},$$

une constante $\mathcal{C} > 0$ existe telle que

$$\forall t \in \mathbf{N}_k(\epsilon) : \quad 0 \leq \mathcal{P}_m(t) \leq 1 - \mathcal{C}m^2\|t - \mu_k^*\|_2^2.$$

– "Far region" :

$$\forall t \in \mathbf{F}(\epsilon) := \mathbb{R}^d \setminus \mathbf{N}(\epsilon) : \quad 0 \leq \mathcal{P}_m(t) \leq 1 - \gamma \frac{v^2}{d^3}.$$

- *iii)* Le support de la transformée de Fourier de \mathcal{P}_m croît linéairement avec m :

$$\text{Supp}(\mathcal{F}[\mathcal{P}_m]) \subset [-4m, 4m]^d \quad \text{et} \quad \|\mathcal{P}_m\|_2 \lesssim K^2 m^{-d/2}.$$

- *iv)* Si (\mathcal{H}_η) est vérifiée avec $\eta = 4m$, alors il existe $c_{0,m} \in \mathbb{L}$ tel que $\mathcal{P}_m = \Phi_{c_{0,m}}$.

Cette construction est inspirée de celle donnée dans Candès et Fernandez-Granda (2014), adaptée à notre cadre spécifique. Nous soulignons que la taille du spectre de \mathcal{P}_m augmente linéairement avec m , tandis que l'effet du nombre de points K , de la dimension d , et de l'espacement Δ^* entre les paramètres de localisation $\{\mu_1^*, \dots, \mu_K^*\}$ se traduit dans la contrainte initiale $m \gtrsim \sqrt{K}d^{3/2}/\Delta_+^*$.

De la même façon, pour chaque pic μ_k^* , on peut construire sous les mêmes conditions un certificat dual \mathcal{Q}_m^k qui interpole 1 au point μ_k^* , dont le comportement est différent dans la région proche de μ_k^* par rapport aux régions proches des autres pics et tel que $\exists c_{k,m} \in \mathbb{L}$ tel que $\mathcal{Q}_m^k = \Phi_{c_{k,m}}$ (voir Corollaire 8 dans [DGMM21]).

5.4.2 Divergence de Bregman $D_{\mathcal{P}_m}(\hat{G}_n, G^*)$

La perte statistique entre \hat{G}_n et G^* est exprimée par la divergence de Bregman associée au certificat dual \mathcal{P}_m (obtenu dans le Théorème 5.4.1) :

$$D_{\mathcal{P}_m}(\hat{G}_n, G^*) := \|\hat{G}_n\|_1 - \|G^*\|_1 - \int_{\mathbb{R}^d} \mathcal{P}_m d(\hat{G}_n - G^*) \geq 0. \quad (5.15)$$

Le résultat suivant donne un contrôle entre \hat{G}_n et G^* avec la divergence de Bregman.

Proposition 5.4.2. *Soit $\mathcal{P}_m = \Phi_{c_{0,m}}$ le certificat dual obtenu dans le Théorème 5.4.1. Soit $(\rho_n)_{n \in \mathbb{N}^*}$ une suite telle que $\mathbb{E}[\|L\hat{f}_n - L \circ \Phi G^*\|_{\mathbb{L}}^2] \leq \rho_n^2$ pour tout $n \in \mathbb{N}^*$. Si κ est choisi tel que*

$$\kappa = \frac{\rho_n}{\|c_{0,m}\|_{\mathbb{L}}} \quad (5.16)$$

et si \hat{G}_n est défini par (\mathbf{P}_κ) , alors :

i) pour tout entier n :

$$\mathbb{E} \left[D_{\mathcal{P}_m}(\hat{G}_n, G^*) \right] \leq \frac{3\sqrt{2}}{2} \rho_n \|c_{0,m}\|_{\mathbb{L}},$$

ii) $c_{0,m} \in \mathbb{L}$ satisfait

$$\|c_{0,m}\|_{\mathbb{L}} \leq \sqrt{\frac{\|\mathcal{P}_m\|_2^2}{\inf_{\|t\|_\infty \leq 4m} \{\sigma^2(t)\mathcal{F}[\lambda](t)\}}} \lesssim \frac{K^2 m^{-d/2}}{\underbrace{\sqrt{\inf_{\|t\|_\infty \leq 4m} \{\sigma^2(t)\mathcal{F}[\lambda](t)\}}}_{:=C_m(\phi,\lambda)}}. \quad (5.17)$$

Les résultats précédents indiquent que la divergence de Bregman entre notre estimateur \hat{G}_n et la mesure cible G^* dépend, à constantes près, de trois quantités principales :

- La variance de la mesure empirique par l'opérateur L quantifiée par ρ_n
- La transformée de Fourier σ du noyau de convolution ϕ sur l'intervalle $[-4m; 4m]^d$. Ce terme mesure le caractère mal posé du problème inverse, qui est associé à la difficulté de retrouver G^* avec des observations indirectes
- La structure du RKHS via le noyau λ .

Remarque 5.1.

- Par des arguments similaires pour le certificat dual \mathcal{Q}_m^k , on peut prouver qu'il existe $c_{k,m} \in \mathbb{L}$ tel que $\mathcal{Q}_m^k = \Phi c_{k,m}$ et $\|c_{k,m}\|_{\mathbb{L}} \lesssim C_m(\phi, \lambda)$.
- La Proposition 5.4.2 permet d'obtenir des vitesses de convergence explicites dans chaque situation spécifique, en explicitant les quantités dans $C_m(\phi, \lambda)$. Des exemples seront discutés en Section 5.5. On constate que le paramètre m joue un rôle important dans ces vitesses. En effet, κ est lié à m par la relation $\kappa = \rho_n / \|c_{0,m}\|_{\mathbb{L}}$. De même, nous verrons dans la Section 5.5 que le noyau λ est également lié à m de manière explicite. Notons que la condition $m \gtrsim \sqrt{K} d^{3/2} \Delta_+^{-1}$ sera satisfaite si on laisse m aller à l'infini avec n alors que K , Δ et d sont maintenus fixes.
- La borne supérieure proposée dans la Proposition 5.4.2 utilise uniquement les points iii) et iv) du Théorème 5.4.1.

5.4.3 Détection des régions proche et éloignée

Le résultat suivant met en lumière les performances de l'estimateur BLASSO \hat{G}_n à retrouver la vraie mesure de mélange G^* . La bonne adéquation se traduit par les propriétés statistiques de la charge de \hat{G}_n autour des pics de G^* (les points de support $S^* = \{\mu_1^*, \dots, \mu_K^*\}$), qui définiront une famille de K régions proches, ainsi que le comportement de \hat{G}_n sur l'ensemble complémentaire, la région éloignée. Les ensembles $\mathbb{F}(\epsilon)$ et $\mathbb{N}(\epsilon)$ ont déjà été introduits dans le Théorème 5.4.1. Notre résultat exploite les bornes précédentes et les points i) et ii) du Théorème 5.4.1.

Théorème 5.4.3. Soit $m \gtrsim \sqrt{K}d^{3/2}/\Delta_+^*$ et soit \mathcal{P}_m le certificat dual donné dans le Théorème 5.4.1. Supposons que \hat{G}_n est l'estimateur BLASSO donné par (\mathbf{P}_κ) avec $\kappa = \kappa_n$ choisi comme en Proposition 5.4.2. Soit $\mathcal{C}_m(\phi, \lambda)$ la quantité introduite dans la Proposition 5.4.2, $\hat{G}_n = \hat{G}_n^+ - \hat{G}_n^-$ la décomposition de Jordan de \hat{G}_n . Un couple de constantes universelles (γ, ν) existe tel que, si

$$\epsilon = \frac{\nu}{md}, \quad (5.18)$$

i) Région éloignée et partie négative :

$$\mathbb{E} \left[\hat{G}_n^-(\mathbb{R}^d) \right] \lesssim \rho_n \mathcal{C}_m(\phi, \lambda) \text{ et } \mathbb{E} \left[\hat{G}_n^+(\mathbb{F}(\epsilon)) \right] \lesssim \frac{d^3}{\gamma \nu^2} \rho_n \mathcal{C}_m(\phi, \lambda).$$

ii) Détection des pics : une constante $\mathcal{C} > 0$ existe telle que

$$\forall A \subset \mathbb{R}^d, \quad \mathbb{E}[\hat{G}_n^+(A)] \gtrsim \frac{d^3}{\gamma \nu^2} \rho_n \mathcal{C}_m(\phi, \lambda) \implies \min_{k \in [K]} \inf_{t \in A} \|t - \mu_k^*\|_2 \leq \frac{\gamma \nu^2}{\mathcal{C} d^3 m^2}.$$

iii) Reconstruction des poids : pour tout $k \in [K]$:

$$\mathbb{E} \left[|\epsilon_k^* - \hat{G}_n(\mathbb{N}_k(\epsilon))| \right] \lesssim \rho_n \mathcal{C}_m(\phi, \lambda).$$

Remarque 5.2. On peut montrer dans des situations spécifiques (voir *iv*) du Théorème 5.3.3) que la solution de (\mathbf{P}_κ) est en effet une mesure discrète qui peut s'écrire comme suit

$$\hat{G}_n = \sum_{\mu \in \hat{S}} \hat{\epsilon}_\mu \delta_\mu.$$

Dans un tel cas, la pertinence des emplacements \hat{S} des pics reconstruits $\hat{\epsilon}_\mu$ peut être déduite des résultats du Théorème 5.4.3.

5.4.4 Stabilité du support pour n assez grand

On introduit $\mathcal{P}_0 := \Phi c_0$ le “minimal norm certificate” (voir Duval et Peyré, 2015), défini par

$$c_0 = \operatorname{argmin} \left\{ \|c\|_{\mathbb{L}}^2 : c \in \mathbb{L} \text{ s.t. } \|\Phi c\|_\infty \leq 1 \text{ et } (\Phi c)(\mu_k^*) = 1, k \in [K] \right\},$$

quand il existe.

Le support S^* de G^* satisfait la condition *Non-Degenerate Bandwidth* (NDB) s'il existe $0 < q < 1$, $r > 0$ et $\rho > 0$ tels que

$$\mathcal{P}_0 \text{ existe, } \forall t \in \mathbb{F}(r), |\mathcal{P}_0(t)| < 1 - q, \quad \forall t \in \mathbb{N}(r), \nabla^2 \mathcal{P}_0(t) \prec -\rho \operatorname{Id}_d. \quad (\text{NDB})$$

Nous avons alors le résultat de stabilité du support pour des grandes valeurs de n .

Théorème 5.4.4. Soit le triplet λ, ϕ, G^* tel que (NDB) est satisfaite. Soit $r_\kappa \in]0, \frac{1}{2}[$ et $\kappa_n = \sqrt{\lambda(0)} n^{-r_\kappa}$. Soit \hat{G}_n l'estimateur BLASSO de (\mathbf{P}_κ) avec le paramètre $\kappa = \kappa_n$. Alors pour n assez grand, et avec probabilité supérieure à $1 - Ce^{-n^{\frac{1}{2}-r_\kappa}}$ pour une constante $C > 0$, il vient que \hat{G}_n a K pics avec exactement un pic $\hat{\mu}_k$ dans chaque région $\mathbb{N}_k(r)$. Ces pics convergent vers les vrais pics, et il en est de même pour les amplitudes $\hat{\epsilon}_k$, lorsque n tend vers l'infini.

Nous soulignons que C est indépendant de la dimension d , du RKHS \mathbb{L} et de l'emplacement des pics. Dans le Théorème 5.4.4, on note que le noyau λ est fixe mais en pratique, la largeur de bande de λ dépend souvent de la taille de l'échantillon n . Le Théorème 5.4.4 suggère l'heuristique que le noyau $\lambda = \lambda_n$ peut dépendre de n et il pourrait être tel que $\kappa_n = \sqrt{\lambda_n(0)} n^{-r_\kappa}$ s'annule quand n tend vers l'infini.

5.5 Application à quelques mélanges usuels

5.5.1 Noyau sinus cardinal et coupure des fréquences

Dans cette section, nous décrivons les conséquences du Théorème 5.4.3 pour certains modèles de mélange basés sur des densités usuelles ϕ . On considère le noyau sinus cardinal sinc avec une coupure de fréquence $1/\tau$, introduit dans l'Exemple 5.3.2. Comme

$$\|t\|_\infty \geq \tau^{-1} \implies \mathcal{F}[\lambda_\tau](t) = 0,$$

on choisit $\tau^{-1} = 4m$ pour avoir $\mathcal{C}_m(\phi, \lambda) < +\infty$ (voir Equation (5.17)). Dans ce cas, $\mathcal{F}[\lambda_\tau]$ est une fonction constante sur son support et le terme $\mathcal{C}_m(\phi, \lambda_\tau)$ vaut

$$\mathcal{C}_m(\phi, \lambda_\tau) = \frac{K^2 m^{-d/2} 2^{d/2}}{\inf_{\|t\|_\infty \leq 4m} \sigma(t)}.$$

Pour utiliser le Théorème 5.4.3, nous avons également besoin d'une expression explicite de $(\rho_n)_{n \in \mathbb{N}^*}$, qui dépend du noyau λ_τ . Dans ce contexte, quelques calculs donnent

$$\mathbb{E} \left[\|L\hat{f}_n - L \circ \Phi G^* \|_{\mathbb{L}}^2 \right] \leq \frac{1}{n\tau^d}$$

ce qui conduit au choix

$$\forall n \in \mathbb{N}^* \quad \rho_n = \frac{1}{\sqrt{n\tau^d}} = \frac{2^d m^{d/2}}{\sqrt{n}}.$$

Par conséquent, le taux obtenu dans le Théorème 5.4.3 satisfait

$$\rho_n \mathcal{C}_m(\phi, \lambda_\tau) \leq \frac{K^2 2^{3d/2}}{\sqrt{n} \times \inf_{\|t\|_\infty \leq 4m} \sigma(t)}. \quad (5.19)$$

Nous devons comprendre l'inégalité précédente comme une limite supérieure qui traduit un compromis entre la finesse de la fenêtre où sont localisés les pics (donnée par $\epsilon = \mathcal{O}(1/(md))$ dans (5.18)) et la capacité statistique associée à récupérer une telle précision ciblée (donnée par la borne $\rho_n \mathcal{C}_m(\phi, \lambda_\tau)$ sur la divergence de Bregman) : la taille de la fenêtre ϵ est améliorée pour de grandes valeurs de m mais la variabilité statistique est alors dégradée en fonction du taux de décroissance de la transformée de Fourier σ de ϕ . Enfin, nous soulignons que l'effet de la dimension n'est pas seulement impliqué dans le terme $2^{3d/2}$ de l'Equation (5.19) mais est également caché dans la contrainte

$$m \gtrsim \sqrt{K} d^{3/2} / \Delta_+^*,$$

utilisée pour construire notre certificat dual dans le Théorème 5.4.1.

Nous allons maintenant nous intéresser à deux catégories de mélange :

- le cas de densités *super-smooth* où la transformée de Fourier décroît exponentiellement dans les hautes fréquences, dont l'exemple le plus connu est la densité gaussienne
- le cas de densités *ordinary smooth* où la transformée de Fourier décroît polynomialement, qui comprend les distributions de Laplace multivariée, de Gamma...

5.5.2 Mélanges de densités "super-smooth" dont le cas gaussien

La densité de mélange ϕ est supposée avoir une décroissance exponentielle de sa transformée de Fourier :

$$\exists j \in \mathbb{N}^*; \mathcal{F}[\phi](t) = \sigma(t) = e^{-\alpha \|t\|_j^\beta} \quad \forall t \in \mathbb{R}^d, \alpha > 0, \beta > 0, \quad (\mathcal{H}_{\alpha, \beta}^{\text{super-smooth}})$$

où pour tout $j \in \mathbb{N}^*$, $\|\cdot\|_j$ désigne la norme ℓ_j . L'hypothèse $(\mathcal{H}_{\alpha, \beta}^{\text{super-smooth}})$ inclut en particulier la distribution gaussienne. Sous cette hypothèse $(\mathcal{H}_{\alpha, \beta}^{\text{super-smooth}})$, on peut vérifier que

$$\inf_{\|t\|_\infty \leq 4m} \sigma(t) = e^{-\alpha(4d^{1/j}m)^\beta}$$

et donc

$$\rho_n \mathcal{C}_m(\phi, \lambda_\tau) \lesssim K^2 2^{3d/2} \times \frac{e^{\alpha(4d^{1/j}m)^\beta}}{\sqrt{n}}.$$

On peut alors appliquer le Théorème 5.4.3 avec cette majoration. Ainsi \hat{G}_n conduit à une estimation cohérente dès que m est choisi tel que

$$m = \left(\frac{\delta \log n}{\alpha} \right)^{1/\beta} \frac{1}{4d^{1/j}} \quad \text{avec} \quad \delta \in \left] 0, \frac{1}{2} \right[.$$

Dans ce cas,

$$\max \left(\mathbb{E} \left[\hat{G}_n^-(\mathbb{R}^d) \right], \mathbb{E} \left[\hat{G}_n^+(\mathbb{F}(\epsilon)) \right], \mathbb{E} \left[|\varepsilon_k^* - \hat{G}_n(\mathbb{N}_k(\epsilon))| \right] \right) \lesssim n^{-\frac{1}{2} + \delta},$$

et tout ensemble A tel que $\mathbb{E}[\hat{G}_n^+(A)] \gtrsim n^{-\frac{1}{2} + \delta}$ est au moins à une distance logarithmique ($\mathcal{O}(m^{-2})$) d'un vrai pic. On retrouve cette limitation logarithmique dans la situation "super-smooth" (voir par ex. Fan, 1991).

Dans le cas particulier de la densité gaussienne ($\beta = j = 2$ et $\alpha = 1/2$), nous obtenons que si $m \gtrsim \sqrt{K} d^{3/2} / \Delta_+^*$ et si $\epsilon = \mathcal{O}(\frac{1}{md})$, alors

$$\mathbb{E} \left[\hat{G}_n^-(\mathbb{R}^d) \right] \lesssim K^2 2^{3d/2} \times \frac{e^{8dm^2}}{\sqrt{n}} \quad \text{et} \quad \mathbb{E} \left[\hat{G}_n^+(\mathbb{F}(\epsilon)) \right] \lesssim K^2 d^3 2^{3d/2} \times \frac{e^{8dm^2}}{\sqrt{n}}. \quad (5.20)$$

De plus, un couple de constantes (c, \mathcal{C}) existe tel que

$$\forall A \subset \mathbb{R}^d, \quad \mathbb{E}[\hat{G}_n^+(A)] > cd^3 2^{3d/2} K^2 \times \frac{e^{8dm^2}}{\sqrt{n}} \implies \min_{k \in [K]} \inf_{t \in A} \|t - \mu_k^*\|_2^2 \leq \frac{1}{\mathcal{C} d^3 m^2}, \quad (5.21)$$

alors que la récupération des poids est assurée par l'inégalité suivante :

$$\forall k \in [K], \quad \mathbb{E} \left[|\varepsilon_k^* - \hat{G}_n(\mathbb{N}_k(\epsilon))| \right] \lesssim 2^{3d/2} K^2 \times \frac{e^{8dm^2}}{\sqrt{n}}.$$

Remarque 5.3.

- Lorsque d , K et Δ^* sont fixes, les performances de l'estimateur BLASSO \hat{G}_n sont d'ordre e^{8dm^2}/\sqrt{n} . En particulier, ce terme converge vers 0 si

$$e^{8dm^2} \ll \sqrt{n} \text{ i.e. } m = \mathcal{O}\left(\sqrt{\frac{\log(n)}{d}}\right) \text{ et } m \rightarrow +\infty \text{ quand } n \rightarrow +\infty. \quad (5.22)$$

En d'autres termes, la valeur maximale admissible pour m est $\sqrt{\frac{\log(n)}{16d}}$. En particulier, si l'on considère $m = \sqrt{\frac{\delta}{16} \frac{\log(n)}{d}}$ pour δ assez petit, on observe que

$$\mathbb{E}\left[\hat{G}_n^-(\mathbb{R}^d)\right] + \mathbb{E}\left[\hat{G}_n^+(\mathbb{F}(\epsilon_n))\right] \lesssim \sqrt{n}^{\delta-1}.$$

La contrepartie de cette taille admissible pour m est un taux lent pour ϵ_n :

$$\epsilon_n = \mathcal{O}\left(\frac{1}{md}\right) = \frac{\delta^{-1/2}}{\sqrt{d \log n}}.$$

- L'Equation (5.20) montre que la masse admise sur la région proche $\mathbb{N}(\epsilon)$ par \hat{G}_n s'approche de 1 quand n croît sous la Condition (5.22). On souligne que le paramètre m joue le rôle d'un indice de précision : si m est constant, la masse de la région proche converge vers 1 à une vitesse paramétrique. En revanche, si m est proche de la valeur limite admissible exprimée en (5.22), l'Equation (5.21) traduit que la région proche est proche du support S^* mais la convergence de la masse associée sera assez lente.
- Il est également intéressant de prêter attention à l'effet de K , Δ^* et d sur la précision statistique de notre méthode. Dans le cas gaussien, le taux est de l'ordre de $K^2 C^d e^{8dm^2} n^{-1/2}$ mais un effet important est caché dans la contrainte

$$m \gtrsim \sqrt{K} d^{3/2} / \Delta_+^*.$$

Le comportement de \hat{G}_n est sérieusement dégradé lorsque $((\Delta^*)^{-1} \vee K \vee d) \rightarrow +\infty$ puisque dans ce cas, en prenant la valeur minimale de m satisfaisant la contrainte précédente, on obtient un taux de l'ordre de $e^{d^4 K \Delta^{-2}} n^{-1/2}$. Ainsi d , K et $(\Delta^*)^{-1}$ ne peuvent pas augmenter plus vite qu'une puissance de $\log(n)$: $d^4 K \Delta^{-2} \ll \log(n)$.

5.5.3 Mélanges de densités "ordinary smooth"

Dans cette section, nous supposons que la densité ϕ est "ordinary smooth" c'est-à-dire que sa transformée de Fourier décroît polynomialement

$$\mathcal{F}[\phi] = \sigma \quad \text{et} \quad \|x\|_2^{-\beta} \lesssim \sigma(x) \lesssim \|x\|_2^{-\beta} \quad \text{quand} \quad \|x\|_2 \rightarrow +\infty. \quad (\mathcal{H}_\beta^{\text{smooth}})$$

On peut vérifier dans ce cas que

$$\inf_{\|t\|_\infty \leq 4m} \sigma(t) \leq \inf_{\|t\|_2 \leq 4m\sqrt{d}} \sigma(t) \lesssim [\sqrt{dm}]^{-\beta}$$

et donc

$$\rho_n \mathcal{C}_m(\phi, \lambda_\tau) \lesssim \frac{K^2 2^{3d/2} m^\beta d^{\beta/2}}{\sqrt{n}}.$$

On peut alors appliquer le Théorème 5.4.3 avec cette majoration.

Si K , d et Δ sont bornés, nous obtenons une bonne estimation avec \hat{G}_n lorsque m est choisi tel que

$$m = n^\delta \quad \text{avec} \quad \delta < \frac{1}{2\beta} \quad \text{quand} \quad n \rightarrow +\infty.$$

Dans ce cas, $\epsilon_n = \mathcal{O}(d^{-1}n^{-\delta})$.

Maintenant, si on laisse $K \vee d \vee (\Delta^*)^{-1}$ croître vers $+\infty$, en fixant $m \sim \sqrt{K} d^{3/2} / \Delta_+^*$ on obtient

$$\max \left(\mathbb{E} \left[\hat{G}_n^-(\mathbb{R}^d) \right], \mathbb{E} \left[\hat{G}_n^+(\mathbb{F}(\epsilon)) \right], \mathbb{E} \left[|\epsilon_k^* - \hat{G}_n(\mathbb{N}_k(\epsilon))| \right] \right) \lesssim \frac{2^{3d/2} K^{2+\beta/2} (\Delta_+^*)^{-\beta} d^{2\beta+3}}{\sqrt{n}}.$$

En particulier, l'ordre maximal pour la dimension est $\mathcal{O}(\log(n))$ quand $n \rightarrow +\infty$. De la même manière, la taille minimale des espacements permettant une bonne estimation ne doit pas être inférieure à $n^{-1/(2\beta)}$. En particulier, cela indique qu'une précision polynomiale est possible. Cela souligne le rôle important joué par la densité de mélange ϕ dans notre analyse.

Dans le cas de la densité de Laplace multivariée ($\beta = 2$), on a

$$\sigma(x) = \frac{2}{2 + \|x\|_2^2}.$$

L'espacement minimal pour des pics détectables est donc de l'ordre de $n^{-1/4}$ tandis que la contrainte sur la dimension n'est pas affectée par la valeur de β . Concernant le nombre de composantes K , sa valeur ne doit pas dépasser $n^{1/6}$ et la taille minimale de la fenêtre ϵ_n est de $n^{-1/4}$.

Si ϕ est le produit tensoriel de distributions univariées standard de Laplace :

$$\phi(x) = \frac{1}{2^d} e^{-\sum_{j=1}^d |x_j|} \quad \text{et} \quad \mathcal{F}[\phi](x) := \sigma(x) = \prod_{j=1}^d \frac{1}{1 + x_j^2} \quad \forall x \in \mathbb{R}^d.$$

Dans ce cas, $\beta = 2d$ et les commentaires précédents s'appliquent : la valeur maximale de m est $n^{1/4d}$ avec une taille optimale de la fenêtre de l'ordre de $n^{-1/(4d)}$ alors que d devrait être au moins d'ordre $\mathcal{O}(\log(n))$.

5.6 Conclusion et perspectives

L'article de Candès et Fernandez-Granda (2014), établissant un résultat de stabilité et la construction d'un *certificat dual*, fut le point de départ de notre travail. Cependant, ces auteurs travaillent sur le tore et leur construction fournit un certificat dual périodique non utilisable dans notre contexte de travail. Aussi, une contribution importante de notre travail fut une nouvelle construction de certificat dual, interpolant les phases/signes sur \mathbb{R}^d . Nous avons de plus étudié la capacité de notre procédure d'estimation à retrouver le mélange à partir de l'évaluation de \hat{G}_n selon un échantillon $(X_i)_{i \in [n]}$ avec $n \rightarrow +\infty$.

Notre procédure s'appuie sur deux paramètres clés : le paramètre $m \geq 1$ de bande passante du terme d'ajustement des données et le paramètre $\kappa > 0$ de réglage de la régularisation ℓ_1 .

Nos résultats montrent que le paramètre m ne dépend que de la difficulté intrinsèque à estimer le support S^* de la cible G^* (à travers la séparation minimale Δ^*).

Un des challenges maintenant est l'évaluation en pratique de l'estimateur BLASSO \hat{G}_n pour des mélanges. Les algorithmes actuels pour résoudre (\mathbf{P}_κ) avec une approche "off-the-grid" peuvent être globalement classés en deux catégories : les algorithmes gloutons et les méthodes de descente riemannienne. Une première idée est d'envisager un algorithme de Franck-Wolfe. Malheureusement, les mesures générées le long de cet algorithme ne sont pas sparses par rapport à \hat{G}_n , au sens où chaque Dirac de \hat{G}_n est approché par de multiples masses de Dirac avec une position proche. Aussi Denoyelle *et al.* (2019) suggèrent d'ajouter une étape supplémentaire dite de "sliding". Une seconde piste est d'adapter un algorithme CPGD (Conic Particle Gradient Descent) proposé par Chizat (2021) pour résoudre notre problème BLASSO pour des modèles de mélange. L'idée est de discrétiser une mesure positive en un système de particules et utiliser une approximation de champ moyen dans l'espace de Wasserstein conjointement associée à une descente de gradient riemannienne avec la métrique conique. La Figure 5.1 illustre une première *preuve de concept*. On peut voir que cette méthode découvre le bon nombre de masses cibles de Dirac et leurs emplacements car certaines particules se regroupent autour de trois pôles. Certaines particules disparaissent et ne détectent pas le support.

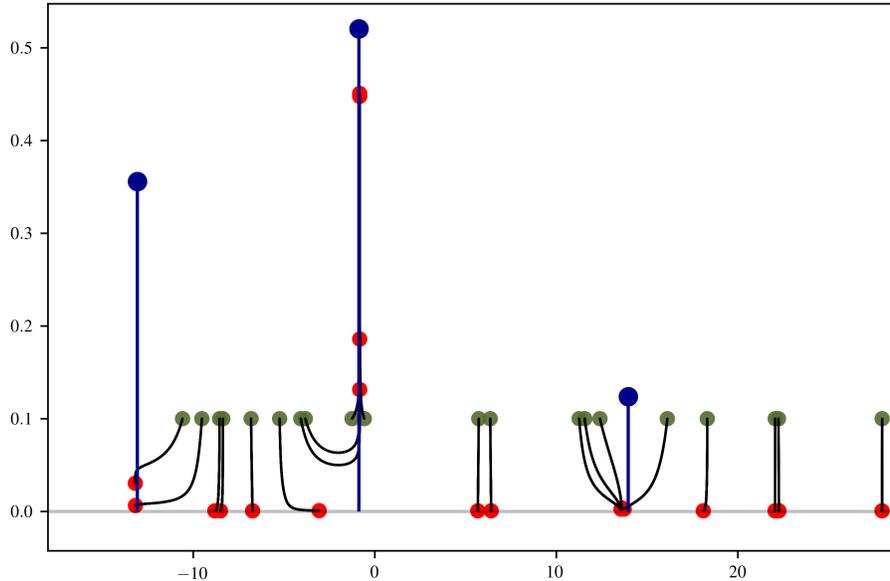


FIGURE 5.1 : La loi de mélange G^* est composée de trois masses de Dirac (en bleu) positionnées en $(-13.1, -0.9, 14.0)$ (choisi au hasard) et d'amplitudes $(0.36, 0.52, 0.12)$ (choisi au hasard). On tire $n = 200$ échantillons par rapport au mélange de densité $f^* = G^* \star \phi$ où ϕ est la densité gaussienne standard. L'algorithme CPGD pour BLASSO (avec les paramètres $\kappa = 0.01$ et $\tau = 0.1$) est initialisé avec 20 particules (en vert) localisées au hasard et on exécute des pas de gradient de 2.5. Les emplacements finaux (μ_k) et les poids (ε_k) sont indiqués en rouge (pour plus de lisibilité nous avons représenté $(\mu_k, 2 * \varepsilon_k)$).

Deuxième partie

Contributions pour l'analyse de
données transcriptomiques

Clustering de gènes co-exprimés à partir de données bulk RNA-seq

Résumé On s’intéresse ici à la classification non supervisée de gènes co-exprimés à partir de données de séquençage à haut-débit (bulk RNA-seq). Une première approche par mélanges de Poisson est proposée dans [RMMC15]. Dans [RM18] et [GMR19], le problème est reformulé en la classification de profils d’expression via des données compositionnelles. La classification est menée à l’aide de mélanges gaussiens sur des profils transformés dans [RM18]. Dans [GMR19], une adaptation des K -means avec la distance d’Aitchison est proposée pour prendre en compte les spécificités du problème de la co-expression.

[GMR19] Godichon-Baggioni A., Maugis-Rabusseau C. and Rau A. (2019). Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. *Journal of Applied Statistics*, 46, pp. 47–65.

[RM18] Rau A. and Maugis-Rabusseau C. (2018). Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics*, 19(3), pp. 425–436.

[RMMC15] Rau A., Maugis-Rabusseau C., Martin-Magniette M.-L. and Celeux G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. In : *Bioinformatics*, 31(9), pp. 1420–1427.

Sommaire

6.1. Contexte et objectif	58
6.1.1. Principe des données bulk RNA-seq	58
6.1.2. Classes de gènes co-exprimés	58
6.1.3. La classification par modèles de mélange	59
6.2. Classification à partir des comptages	61
6.2.1. Modèles de mélange de Poisson	61
6.2.2. Illustration	63
6.2.3. Discussion	64
6.3. Classification à partir des profils transformés	64
6.3.1. Mélanges gaussiens sur profils transformés	64
6.3.2. Classification par K -means de profils logCLR-transformés	68
6.4. Conclusions et perspectives	74

6.1 Contexte et objectif

6.1.1 Principe des données bulk RNA-seq

La technologie du séquençage à haut-débit (HTS), communément appelé séquençage de nouvelle génération (NGS), est devenue un outil essentiel pour les études génomiques et transcriptomiques. En quantifiant et en comparant les transcriptomes de différents types de tissus, stades de développement ou conditions expérimentales, les chercheurs ont ainsi acquis une meilleure compréhension de l'impact de changements de l'activité transcriptionnelle sur des types de cellules et impliquant des différences phénotypiques. En particulier, l'application de la technologie HTS pour le séquençage de l'ARN (RNA-seq) a révolutionné l'étude de l'expression des gènes en fournissant un niveau de détail plus poussé que les puces à ADN (*microarray*) et permettant un large éventail d'applications nouvelles. Cette technique permet une couverture élevée du génome, la détection de gènes faiblement exprimés et la quantification de l'expression des gènes sans connaissance préalable du génome. Les séquences cibles fragmentées (appelées librairies) sont directement séquencées sur une lame de verre, plutôt qu'hybridées sur une puce, et doivent ensuite être assemblées ou alignées sur un génome de référence. Après l'assemblage ou l'alignement, le niveau d'expression de chaque région d'intérêt doit être quantifié. Ces étapes de prétraitement, qui ne sont pas triviales et représentent des domaines de recherche majeurs dans le domaine de la bioinformatique, conduisent généralement à une mesure de l'expression génétique basée sur le comptage. Ainsi, contrairement aux données continues des puces à ADN, les données RNA-seq sont des comptages hétérogènes (jusqu'à 5 ordres de grandeur) pour des régions génomiques d'intérêt (généralement les gènes), et présentent une inflation de zéros et une surdispersion entre les réplicats biologiques. En outre, la profondeur de séquençage (c'est-à-dire la taille de la librairie) et la couverture varient d'une expérience à l'autre, et l'on sait que le nombre de lectures est corrélé à la longueur des gènes (Oshlack et Wakefield, 2009; Łabaj *et al.*, 2011). Ainsi les méthodes d'analyse des données de puces à ADN, principalement basées sur l'utilisation de lois gaussiennes après prétraitements, ne sont pas adaptées aux données RNA-seq. Cette évolution de technologie a donc conduit au développement de nouvelles méthodes d'analyse, notamment pour la normalisation (Mortazavi *et al.*, 2008; Robinson et Oshlack, 2010) et l'analyse de l'expression différentielle (Anders et Huber, 2010; Auer et Doerge, 2011; Dillies *et al.*, 2013) de données RNA-seq. Dans ce chapitre, nous nous concentrons sur la question de la classification non supervisée de gènes co-exprimés à partir de données RNA-seq.

6.1.2 Classes de gènes co-exprimés

L'identification de gènes qui partagent des profils d'expression similaires dans plusieurs conditions de traitement, dit gènes co-exprimés, peut aider à mettre en évidence des groupes de gènes qui sont impliqués dans les mêmes processus biologiques (Eisen *et al.*, 1998; Jiang *et al.*, 2004) et déterminer des fonctions biologiques de gènes orphelins (gènes pour lesquels aucune fonction biologique n'est connue). Dans la suite, on considère que l'on observe un tableau de données RNA-seq $\mathbf{y} = (y_{ij})$ où y_{ij} est le comptage représentant l'expression du gène i ($i = 1, \dots, n$) dans l'échantillon j ($j = 1, \dots, d$). Ces échantillons peuvent être des réplicats biologiques de différentes conditions expérimentales, on notera $c(j)$ la condition dont j est le réplicat et q le nombre de conditions du jeu de données. On suppose qu'une procédure de normalisation des données est effectuée, on note $\tilde{\mathbf{y}} = (\tilde{y}_{ij})$ les comptages normalisés. Ainsi à partir de données RNA-seq normalisées $\tilde{\mathbf{y}}$, on cherche à résoudre un problème de classification

non supervisée afin de mettre en évidence des groupes de gènes co-exprimés.

Pour ce problème de classification non supervisée, l'une des premières questions à aborder est de définir en quel sens des gènes "se ressemblent" à partir de données RNA-seq. Dans ces données, il faut prendre en compte de grandes différences de niveau d'expression des gènes. On ne souhaite pas regrouper des gènes qui ont un niveau d'expression similaire mais ceux qui ont une évolution d'expression similaire dans les différentes expériences. Pour illustrer la réflexion, on considère un sous-ensemble de gènes provenant de données RNA-seq de souris étudiées par Fietz *et al.* (2012) (voir Section 6.3.1.4 pour une description de ces données). Nous considérons dix gènes représentatifs de quatre groupes distincts : les gènes non différenciellement exprimés (NDE) (Groupe 1) ; et les gènes exprimés uniquement dans la première, seconde ou troisième condition expérimentale (Groupe 3, 4 et 2 resp.). On peut clairement constater que les grandes différences d'amplitude observables pour les comptes normalisés \tilde{y}_{ij} (Figure 6.1 A) sont grandement réduites par une transformation logarithmique $\log(\tilde{y}_{ij} + 1)$ (Figure 6.1 B), bien qu'une certaine dispersion reste présente entre les gènes très fortement et faiblement exprimés. Pour ne prendre en compte que l'évolution d'expression, on introduit les profils d'expression c'est-à-dire la proportion de lectures normalisées observées pour le gène i par rapport au total observé pour ce gène dans tous les échantillons :

$$\mathbf{p}_i = (p_{i1}, \dots, p_{id}) \text{ avec } p_{ij} = \frac{\tilde{y}_{ij} + 1}{\sum_{\ell} \tilde{y}_{i\ell} + 1}, \quad \forall j \in \{1, \dots, d\} \quad (6.1)$$

Les profils \mathbf{p}_i sont donc par définition des éléments du simplexe de \mathbb{R}^d :

$$\mathbf{p}_i \in \mathbb{S}^d := \left\{ x = (x_1, \dots, x_d) \in \mathbb{R}^d \mid \sum_{i=1}^d x_i = 1; x_i > 0, \forall i \right\}.$$

On voit sur notre exemple jouet (Figure 6.1 C) que les groupes sont plus clairement visibles avec les profils d'expression normalisés. Cet exemple illustre donc l'intérêt de considérer les profils d'expression normalisés, rendu indépendant du niveau d'expression absolue d'un gène pour ne prendre en compte que l'expression relative du gène dans les différentes expériences.

Dans la suite, nous allons résumer les contributions publiées dans [RMMC15], [RM18] et [GMR19]. Dans [RMMC15], une classification non supervisée basée sur une modélisation par mélanges de Poisson sur les comptages normalisés $\tilde{\mathbf{y}}$ est proposée. Dans [RM18] et [GMR19], les procédures de classification sont développées sur les profils d'expression normalisés \mathbf{p} : dans la première contribution, on souhaite exploiter la flexibilité des mélanges gaussiens ce qui demande de transformer les profils d'expression ; dans la seconde, on s'inspire de la méthode usuelle des K -means avec la distance d'Aitchison pour l'étude de données compositionnelles en l'adaptant à notre contexte biologique. Afin de présenter ces différentes contributions, quelques rappels sur la classification par modèles de mélange sont donnés dans la section suivante.

6.1.3 La classification par modèles de mélange

On fait ici quelques rappels nécessaires pour la suite concernant la classification non supervisée à partir de modèles de mélange. On suppose que $\mathbf{x} = (x_1, \dots, x_n)$ sont des observations de variables aléatoires (X_1, \dots, X_n) i.i.d de densité inconnue f , qui est estimée par un mélange fini

$$f(\cdot | \theta_K) = \sum_{k=1}^K \pi_k f_k(\cdot | \alpha_k) \quad (6.2)$$

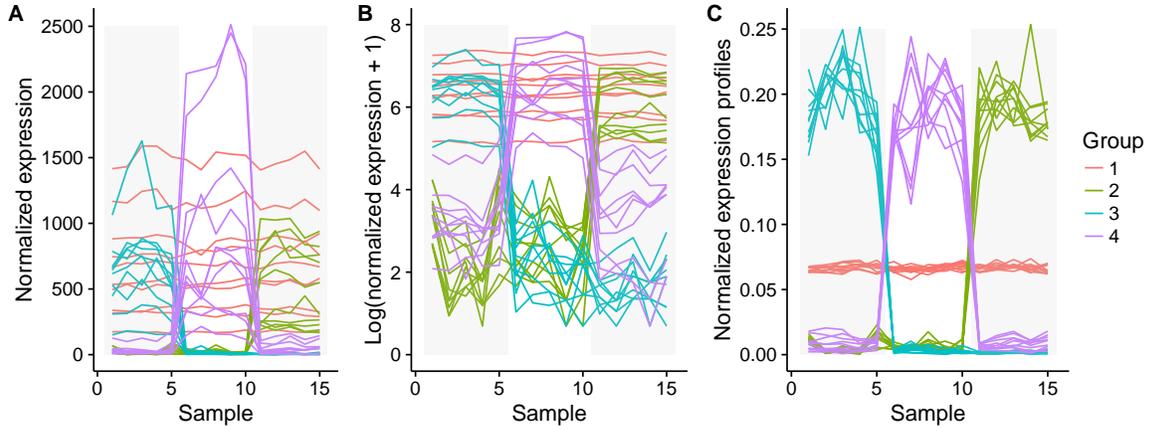


FIGURE 6.1 : Comptages normalisés \tilde{y}_{ij} (A), $\log(\tilde{y}_{ij} + 1)$ (B) et profils d'expression normalisés p_{ij} (C) pour un sous-ensemble des données RNA-seq étudiées dans Fietz et al. (2012). Le sous-ensemble de gènes considéré comprend des gènes non différentiellement exprimés dans toutes les conditions (Groupe 1), des gènes seulement exprimés dans la première condition (échantillons 1 à 5, Groupe 3), des gènes seulement exprimés dans la deuxième condition (échantillons 6 à 10, Groupe 4), et des gènes seulement exprimés dans la troisième condition (échantillons 11 à 15, Groupe 2). Les fonds gris transparents délimitent les réplicats des trois conditions expérimentales.

avec $\theta_K = (\pi_1, \dots, \pi_K, \alpha_1, \dots, \alpha_K)$, $(\pi_1, \dots, \pi_K) \in]0, 1[^K$, $\sum_{k=1}^K \pi_k = 1$. Cette modélisation part de l'idée que s'il y a effectivement une classification sous-jacente des données, les individus d'une même classe k se ressemblent donc partagent la même densité $f_k(\cdot|\alpha_k)$ et ainsi l'ensemble des données est distribué selon un mélange de ces densités. Le modèle de mélange de l'équation (6.2) peut être vu comme un modèle de données incomplètes, les données complètes étant

$$(\mathbf{x}, \mathbf{z}) = ((x_1, z_1), \dots, (x_n, z_n))$$

avec les données manquantes $\mathbf{z} = (z_1, \dots, z_n) = (z_{ik}; i = 1, \dots, n, k = 1, \dots, K)$, telles que $z_{ik} = 1$ si l'observation i provient du groupe k et 0 sinon. La variable latente \mathbf{z} définit donc une partition $\mathcal{P}_K = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ des données observées \mathbf{x} avec $\mathcal{C}_k = \{i; z_{ik} = 1\}$.

En rassemblant les mélanges par nombre de classes K

$$\mathcal{S}_K = \left\{ f(\cdot|\theta_K) = \sum_{k=1}^K \pi_k f_k(\cdot|\alpha_k), \theta_K \in \Theta \right\},$$

on est alors ramené à un problème de sélection de modèle. A K fixé, on détermine le "meilleur" mélange de \mathcal{S}_K en estimant le vecteur des paramètres θ_K par maximum de vraisemblance. Classiquement, un algorithme EM (Expectation-Maximization) introduit par Dempster *et al.* (1977) ou une variante de cet algorithme (voir par ex. McLachlan et Krishnan, 2007) est utilisé pour cette étape d'estimation. On note $\hat{\theta}_K$ l'estimateur obtenu. De l'estimation de f par $f(\cdot|\hat{\theta}_K)$, une classification des individus est déduite à l'aide de la règle du Maximum A Posteriori (MAP) : pour tout $i = 1, \dots, n$, pour tout $k = 1, \dots, K$,

$$\hat{z}_{ik} = \begin{cases} 1 & \text{si } \tau_{ik}(\hat{\theta}_K) > \tau_{il}(\hat{\theta}_K) \forall l \neq k \\ 0 & \text{sinon,} \end{cases} \quad (6.3)$$

où $\tau_{ik}(\hat{\theta}_K)$ est la probabilité conditionnelle que l'individu i appartienne à la k -ème composante de $f(\cdot|\hat{\theta}_K)$

$$\tau_{ik}(\hat{\theta}_K) = \frac{\hat{\pi}_k f_k(x_i|\hat{\alpha}_k)}{f(x_i|\hat{\theta}_K)}. \quad (6.4)$$

Pour résoudre ensuite la sélection de modèle (c'est-à-dire ici le choix du nombre de classes K), on peut considérer le critère BIC (Bayesian Information Criterion) de Schwarz (1978) :

$$\text{BIC}(K) = -\mathcal{L}(\mathbf{x}|\hat{\theta}_K) + \frac{\nu_K}{2} \log(n)$$

où $\mathcal{L}(\mathbf{x}|\hat{\theta}_K) = \ln f(\mathbf{x}; \hat{\theta}_K)$ est la log-vraisemblance au maximum de vraisemblance et ν_K est la dimension du modèle \mathcal{S}_K (c'est-à-dire le nombre de paramètres libres des mélanges de \mathcal{S}_K). Cependant, le critère BIC ne prenant pas en compte l'objectif de classification mais celui d'estimation de densité, il a tendance à sur-estimer K (Biernacki *et al.*, 2000). Une alternative est le critère ICL (Integrated Complete Likelihood) proposé par Biernacki *et al.* (2000) :

$$\text{ICL}(K) = \text{BIC}(K) + \text{Ent}(K) \text{ avec } \text{Ent}(K) = - \sum_{i=1}^n \sum_{k=1}^K \tau_{ik}(\hat{\theta}_K) \log(\tau_{ik}(\hat{\theta}_K)) \quad (6.5)$$

où le terme $\text{Ent}(\cdot)$ est un terme d'entropie qui tient compte de l'homogénéité de la classification. Une approche alternative pour la sélection de modèles est l'utilisation d'un critère de maximum de vraisemblance pénalisé non asymptotique de la forme

$$\text{CRIT}(K) = -\mathcal{L}(\mathbf{x}|\hat{\theta}_K) + \text{pen}(K). \quad (6.6)$$

Dans [MM11a], il est établi pour les mélanges gaussiens que la pénalité est proportionnelle au nombre de paramètres libres ν_K . On utilise alors l'heuristique de pente de Birgé et Massart (2007) :

$$\text{pen}(K) = 2\hat{\kappa}\nu_K$$

où $\hat{\kappa}$ est une constante calibrée à partir des données correspondant à la pente attendue de la relation linéaire de la log-vraisemblance par rapport à la dimension du modèle pour les modèles les plus complexes (ici les modèles avec de grandes valeurs de K). En pratique, cette calibration est réalisée à l'aide de la procédure *data-driven slope estimation* (DDSE) disponible dans la librairie R `capushe` ([BMM12]). Ainsi le nombre de classes sélectionné \hat{K} minimise le critère pénalisé (6.6). Finalement, les individus i sont classés à l'aide de la règle du MAP (voir Equation (6.3)) en fonction de $\hat{\theta}_{\hat{K}}$.

6.2 Classification à partir des comptages

6.2.1 Modèles de mélange de Poisson

Cette section est dédiée aux travaux publiés dans [RMMC15]. Comme Cai *et al.* (2004) et Kim *et al.* (2007), nous utilisons des modèles log-linéaires de Poisson pour la classification à partir de données RNA-seq ; cependant, plutôt que d'utiliser un tel modèle pour définir une métrique pour un algorithme K -means ou de classification hiérarchique, nous utilisons des mélanges finis de modèles log-linéaires de Poisson. Soit Y_{ij} des variables aléatoires correspondant à la mesure

d'expression du gène i ($i = 1, \dots, n$) pour l'échantillon j ($j = 1, \dots, d$), et y_{ij} la valeur observée de Y_{ij} . On suppose que les Y_{ij} sont distribuées selon un mélange de loi de Poisson :

$$f(\mathbf{y}|\theta_K) = \prod_{i=1}^n \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i|\alpha_{ik}) \quad (6.7)$$

avec $\theta_K = (\pi_1, \dots, \pi_K, \boldsymbol{\alpha})$, $\boldsymbol{\alpha}$ contient tous les paramètres $\{\alpha_{ik}\}_{i,k}$. Bien qu'il existe une version multivariée de la distribution de Poisson (Karlis, 2003), il est difficile de l'implémenter pour des données de grandes dimensions. Pour cette raison, les échantillons sont ici supposés indépendants conditionnellement aux composantes :

$$f_k(\mathbf{y}_i|\alpha_{ik}) = \prod_{j=1}^d \mathcal{P}(y_{ij}; \mu_{ijk}),$$

où $\mathcal{P}(\cdot; v)$ est la fonction de probabilité de la loi de Poisson de paramètre v et $\alpha_{ik} = \{\mu_{ijk}\}_j$. Chaque paramètre de moyenne μ_{ijk} est paramétré par

$$\mu_{ijk} = w_i s_j \lambda_{c(j)k} \quad (6.8)$$

où $w_i = \sum_j y_{ij}$ correspond au niveau d'expression global du gène i et s_j représente la taille normalisée de la librairie pour l'échantillon j , de sorte que $\sum_j s_j = 1$. Ces facteurs de normalisation tiennent compte du fait que le nombre de lectures censées correspondre à un gène dépend non seulement de son niveau d'expression, mais aussi de la taille de la librairie (nombre total de reads alignés) et de la composition globale de la population d'ARN échantillonnée (Dillies *et al.*, 2013). On rappelle que $c(j)$ est la condition expérimentale dont j est un réplicat biologique. Le vecteur de paramètres inconnus $\boldsymbol{\lambda}_k = (\lambda_{1k}, \dots, \lambda_{qk})$ correspond aux paramètres de classification qui définissent les profils des gènes de la classe k à travers toutes les conditions expérimentales.

Pour l'estimation des paramètres à K fixé, remarquons tout d'abord que les $\{s_j\}_j$ sont estimés à partir des données avant l'ajustement du modèle (étape préliminaire de normalisation des données), donc comme les niveaux d'expression globaux w_i , ils sont ensuite considérés comme fixes dans le modèle de mélange de Poisson. Il reste donc à estimer $\boldsymbol{\lambda}_k = (\lambda_{1k}, \dots, \lambda_{qk})$ que l'on effectue à l'aide d'un algorithme EM. Après avoir initialisé les paramètres $\theta_K^{(0)}$ et $\mathbf{z}^{(0)}$, l'étape E à l'itération b correspond au calcul de la probabilité conditionnelle qu'une observation i provienne de la composante k pour la valeur courante des paramètres du mélange :

$$\tau_{ik}^{(b)} = \frac{\pi_k^{(b)} f_k(\mathbf{y}_i; \alpha_{ik}^{(b)})}{\sum_{l=1}^K \pi_l^{(b)} f_l(\mathbf{y}_i; \alpha_{il}^{(b)})}. \quad (6.9)$$

Dans l'étape M, l'estimation des paramètres du mélange est mise à jour pour maximiser la valeur attendue de la vraisemblance complétée, ce qui conduit à pondérer l'observation i pour le groupe k avec la probabilité conditionnelle $\tau_{ik}^{(b)}$. Ainsi ,

$$\pi_k^{(b+1)} = \frac{1}{n} \sum_{i=1}^n \tau_{ik}^{(b)} \quad \text{et} \quad \lambda_{ck}^{(b+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(b)} \sum_{j:c(j)=c} y_{ij}}{\tilde{s}_c \sum_{i=1}^n \tau_{ik}^{(b)} w_i}, \quad \forall c = 1, \dots, q, \quad \forall k = 1, \dots, K$$

où $\tilde{s}_c = \sum_{j;c(j)=c} s_j$. Notons qu'à chaque itération de l'algorithme EM, on obtient que $\sum_{c=1}^q \lambda_{ck}^{(b)} \tilde{s}_c = 1$. Ainsi, $\lambda_{ck} \tilde{s}_c$ peut être interprété comme la proportion d'expression qui est attribuée à la condition expérimentale c dans la classe k , après avoir pris en compte les différences dues à la taille de la librairie (les réplicats pour c sont pris en compte par j ; $c(j) = c$ dans les expressions précédentes).

Pour le choix du nombre de classes, on peut utiliser les critères BIC, ICL ou le critère non asymptotique CRIT calibré par l'heuristique de pente présentés en Section 6.1.3, le nombre de paramètres libres valant $\nu_K = (K - 1) + K \times q$. Pour la mise en oeuvre de l'heuristique de pente, il est nécessaire d'estimer les paramètres pour des grandes valeurs de K (afin d'observer le comportement linéaire pour les modèles complexes) mais l'initialisation de l'algorithme EM est alors un point sensible. Aussi nous avons proposé une variante de la stratégie small-EM (Biernacki *et al.*, 2003) avec la stratégie d'initialisation par division (appelée "*small_EM splitting*") de [PMM13], où la classe de plus grande entropie du modèle à $K - 1$ classes est choisie pour être divisée en deux nouvelles classes.

Toute cette procédure de classification par mélanges de Poisson est implémentée dans la librairie R `HTScluster`.

6.2.2 Illustration

Nous illustrons notre procédure sur des données réelles issues du projet modENCODE, qui vise à fournir une annotation fonctionnelle du génome de *Drosophila melanogaster*. Graveley *et al.* (2011) ont caractérisé la dynamique d'expression au cours de 27 étapes distinctes du développement pendant le cycle de vie de la mouche en utilisant la technologie RNA-seq. On considère ici un sous-ensemble de ces données provenant de 12 échantillons embryonnaires qui ont été collectés toutes les deux heures pendant 24 heures, avec un réplicat biologique pour chaque temps. Les tableaux de phénotypes et les comptages bruts pour les 13 164 gènes ayant au moins un comptage non nul parmi les 12 temps ont été obtenus à partir de ressources en ligne ReCount (Frazee *et al.*, 2011).

Sur trois exécutions indépendantes, nous avons utilisé `HTScluster` avec les paramètres par défaut et la stratégie d'initialisation "*small-EM splitting*" pour ajuster une séquence de modèles de mélange de Poisson avec $K = 1, \dots, 60$ classes; pour chaque nombre de classes, le modèle correspondant à la plus grande log-vraisemblance parmi les trois exécutions a été retenu. Pour s'assurer que l'ensemble des modèles considérés est suffisamment grand pour appliquer l'heuristique de pente, un ensemble supplémentaire de modèles a été ajusté pour $K = 65, \dots, 95$ (par pas de 5) et $K = 100, \dots, 130$ (par pas de 10). En utilisant l'heuristique de pente, le nombre de classes retenu est $\hat{K} = 48$.

La visualisation des résultats d'une analyse de co-expression pour les données RNA-seq peut être quelque peu compliquée en raison de la plage dynamique extrêmement large des comptages et du fait que les gènes les plus fortement exprimés ont tendance à présenter une plus grande variabilité (bien que les coefficients de variation soient beaucoup plus faibles) que les gènes faiblement exprimés. Plutôt que de visualiser directement les données elles-mêmes, nous proposons deux visualisations alternatives. La première permet de visualiser le comportement global de chaque classe, comme le montre la Figure 6.2. Dans ce graphique, la largeur des barres correspond à la proportion estimée de gènes dans chaque classe ($\hat{\pi}_k$), et les proportions de lectures attribuées à chaque temps du développement dans chaque classe $\hat{\lambda}_{ck} \tilde{s}_c$ sont représentées par les segments colorés dans chaque barre. L'avantage d'une telle visualisation est qu'elle permet une comparaison directe des profils typiques entre les classes.

Par exemple, on peut voir que les classes caractérisées par une expression relative plus élevée dans les premiers stades embryonnaires, tels que les classes 6 et 13 (composés de 70 et 60 gènes, respectivement) ont tendance à être beaucoup plus petites que celles avec une expression relative plus élevée dans les stades ultérieurs, par exemple, les classes 4, 18, 19 et 21 (composées de 567, 680, 485 et 475 gènes, respectivement). La seconde représentation consiste à représenter les profils d'expression normalisés $\mathbf{p}_i = (p_{ij})_j$ en fonction de j pour tous les gènes d'une même classe, comme dans la Figure 6.2. On retrouve par exemple que la classe 6 (resp. classe 4) regroupe des gènes dont l'expression est élevée dans les premiers (resp. derniers) stades embryonnaires.

6.2.3 Discussion

Les modèles de mélange de Poisson proposés présentent l'avantage de modéliser directement les données de comptage en prenant en compte le design de l'expérience et la taille variable des bibliothèques. Mais il présente plusieurs limites : (1) l'hypothèse d'indépendance conditionnelle entre les échantillons est assez irréaliste pour les données bulk RNA-seq ; (2) les structures de corrélation par classe ne peuvent pas être incluses dans le modèle ; (3) la distribution de Poisson est probablement trop restrictive, car elle impose une hypothèse de moyennes et de variances égales. Une extension avec des mélanges de binomiales négatives a été envisagée mais les résultats n'étaient pas concluants ; (4) les critères asymptotiques BIC et ICL ont souvent été observés comme ayant un mauvais comportement (sélection de K trop élevée). Pour y remédier, on a considéré un critère de sélection de modèle pénalisé non asymptotique, calibré par heuristique de pente (présenté précédemment) mais cette approche nécessite l'ajustement d'une collection de modèles de mélange pour de très grandes valeurs de K , impliquant un temps de calcul important ainsi que des difficultés pratiques pour l'initialisation et l'estimation des paramètres.

6.3 Classification à partir des profils transformés

Suite aux limitations des mélanges de Poisson mentionnées précédemment, nous avons décidé d'axer la classification des gènes à partir des profils d'expressions \mathbf{p}_i (voir Equation (6.1)). Les profils $\mathbf{p}_i = (p_{i1}, \dots, p_{id}) \in [0, 1]^d$ sont tels que $\sum_{j=1}^d p_{ij} = 1$. On se ramène donc à un problème de classification non supervisée de données compositionnelles (Aitchison, 1986), c'est-à-dire à des données dans le simplexe \mathbb{S}^d . Nous allons dans cette section présenter les contributions publiées dans [RM18] et [GMR19].

6.3.1 Mélanges gaussiens sur profils transformés

Les travaux dans [RM18] ont été motivés par l'utilisation populaire des mélanges gaussiens en classification non supervisée et leur flexibilité, en particulier pour modéliser des structures de corrélation par classe. Mais leur utilisation n'étant pas adaptée pour les données compositionnelles, nous avons étudié des transformations des profils appropriées à notre contexte biologique et permettant l'utilisation des mélanges gaussiens.

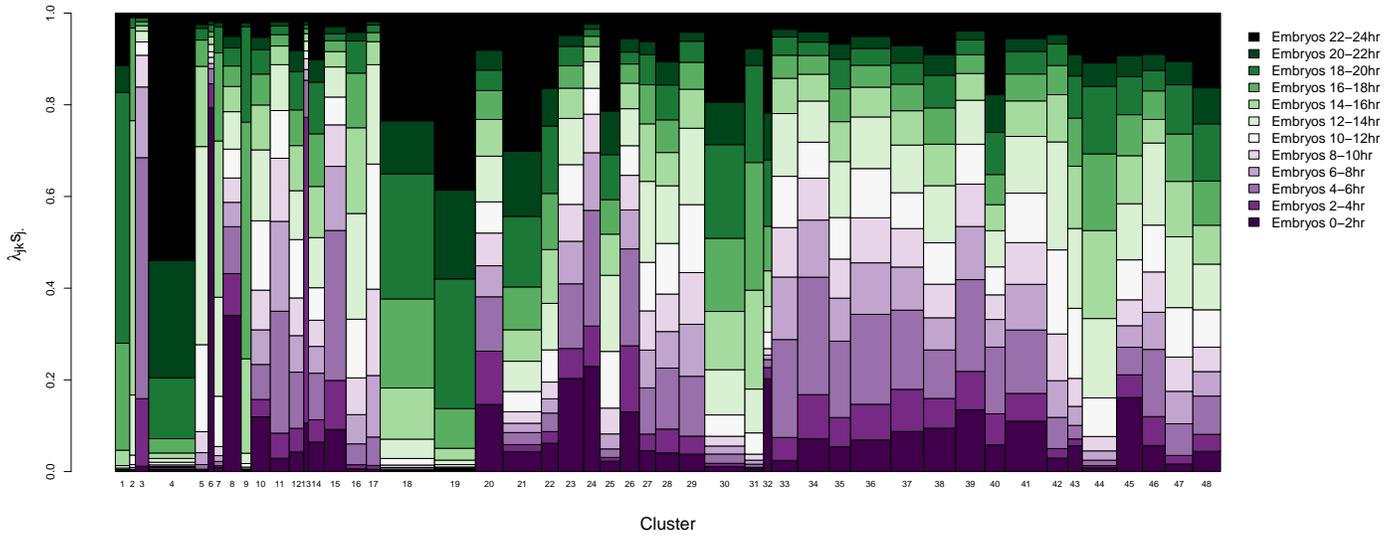


FIGURE 6.2 : Visualisation du comportement global des classes pour les données de développement de *D. melanogaster*. Pour chaque classe, des diagrammes à barres de $\hat{\lambda}_{c_k|\tilde{s}_c}$ sont dessinés pour chaque temps du développement, où la largeur de chaque barre correspond à la proportion estimée $\hat{\pi}_k$.

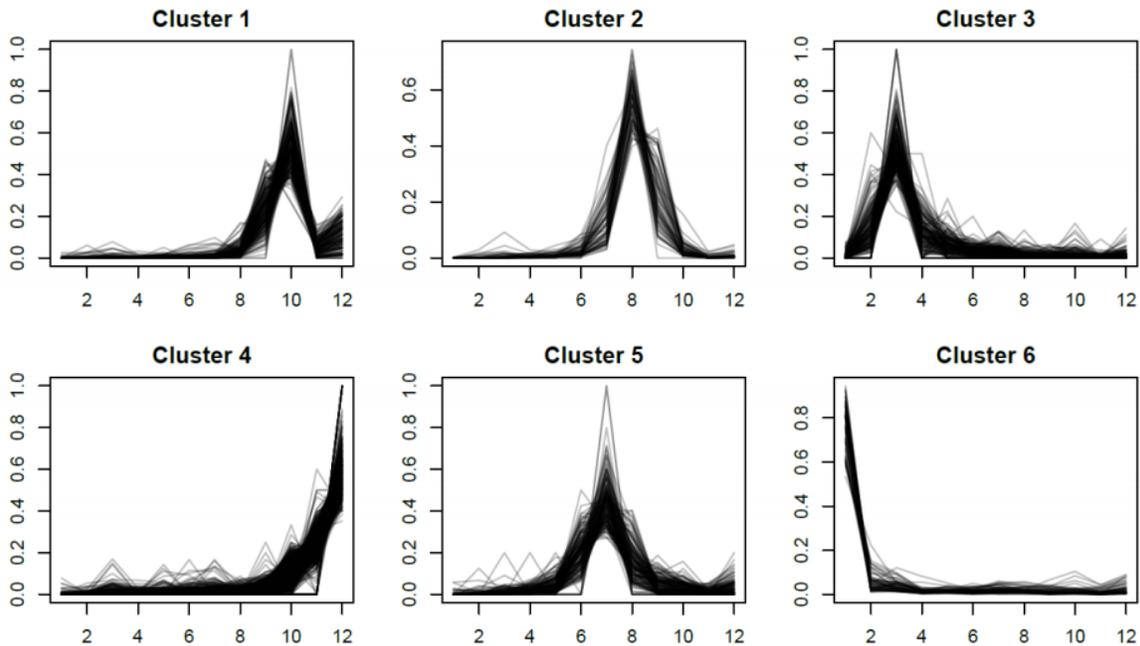


FIGURE 6.3 : Profils d'expression normalisés en fonction des temps de développement pour les gènes assignés aux six premières classes.

6.3.1.1 Mélanges gaussiens

Dans le cas de la classification par mélanges gaussiens, l'ensemble des modèles considéré est défini par $(\mathcal{S}_{(K,v)})_{(K,v) \in \mathcal{M}}$, où

$$\mathcal{S}_{(K,v)} = \left\{ f(\cdot | \theta_{(K,v)}) = \sum_{k=1}^K \pi_{k,v} \phi(\cdot | \mu_k, \Sigma_{k,v}) \right\}, \quad (6.10)$$

avec $\phi(\cdot | \mu_k, \Sigma_{k,v})$ désignant la densité gaussienne d -dimensionnelle de vecteur moyenne μ_k et de matrice de covariance $\Sigma_{k,v}$. L'indice v désigne l'une des formes de mélange gaussien obtenues en contraignant un ou plusieurs des paramètres dans la décomposition suivante de la matrice de covariance de chaque composante du mélange :

$$\Sigma_k = L_k D'_k A_k D_k, \quad (6.11)$$

où $L_k = |\Sigma_k|^{1/d}$, D_k est la matrice des vecteurs propres de Σ_k , et A_k est la matrice diagonale des valeurs propres normalisées de Σ_k . Diverses contraintes sur ces paramètres contrôlent respectivement le volume, l'orientation et la forme de la k -ème classe (Celeux et Govaert, 1995). En permettant de plus aux proportions π_k de varier ou d'être égales pour toutes les classes, nous pouvons définir une collection de 28 modèles de mélange parcimonieux et interprétables, disponibles dans les bibliothèques `Rmixmod` (Langrognet *et al.*, 2020) et `mclust` (Scrucca *et al.*, 2016). On reprend alors la procédure décrite en section 6.1.3 pour l'estimation des paramètres, la sélection de modèle et la classification finale par règle du MAP. Les critères ICL et BIC permettent de sélectionner à la fois K et la forme v . Pour simplifier la présentation, nous allons seulement considérer la forme la plus générale (proportions, volume, orientation et forme variables - appelée $[p_k L_k C_k]$ dans `Rmixmod`) ; la collection de modèles est donc indexée par K dans la suite.

6.3.1.2 Transformation des profils

Il est important de noter que le profil du gène i , $\mathbf{p}_i = (p_{i1}, \dots, p_{id}) \in \mathbb{S}^d$ donc $\sum_{j=1}^d p_{ij} = 1$. Ainsi ses coordonnées sont linéairement dépendantes, ce qui impose des contraintes sur les matrices de covariance Σ_k qui sont problématiques pour le modèle général de mélange gaussien (et en fait pour la plupart des approches statistiques standard). Pour cette raison, nous souhaitons transformer les profils pour briser cette contrainte. De plus, par l'objectif de co-expression, il est attendu que la majorité des gènes ont des profils constants (gènes non différentiellement exprimés dans toutes les conditions) et que nous sommes surtout intéressés par mettre en évidence des variations spécifiques de gènes dans certaines conditions expérimentales. Aussi, nous avons considéré les transformations logit et arcsin suivantes :

$$g_{\arcsin}(p_{ij}) = \arcsin(\sqrt{p_{ij}}) \in [0, \pi/2], \quad (6.12)$$

$$g_{\text{logit}}(p_{ij}) = \log_2 \left(\frac{p_{ij}}{1 - p_{ij}} \right) \in \mathbb{R}. \quad (6.13)$$

Ces deux transformations maintiennent " linéairement " les valeurs intermédiaires des proportions p_{ij} et amplifient les valeurs extrêmes. Cela permet de mettre en avant les profils spécifiques, de maintenir les gènes non différentiellement exprimés ensemble et de ne plus avoir de dépendance linéaire entre coordonnées.

6.3.1.3 Adaptation des critères de sélection de modèle

Les critères pénalisés BIC et ICL peuvent être utilisés pour sélectionner parmi différents modèles ou transformations, comme cela a été suggéré dans Thomas *et al.* (2008) et Gallopin (2015) par exemple. Ce principe permet de s'affranchir du choix arbitraire de la transformation des données en utilisant le cadre de la sélection de modèles. Nous appliquons ici ce principe pour le choix du nombre de classes K et de la transformation des données ; dans un cas plus général, une procédure similaire pourrait être utilisée pour choisir aussi la forme du mélange gaussien (voir Equation (6.11)). Soit $g(\mathbf{x})$ une transformation monotone arbitraire d'un échantillon \mathbf{x} . On suppose que le nouvel échantillon $g(\mathbf{x})$ provient d'une densité de mélange gaussien $f(\cdot|\theta_{(K,g)})$. Si J_g désigne le jacobien de la transformation g et $\hat{\theta}_{(K,g)}$ l'estimation du maximum de vraisemblance obtenue pour le mélange à K classes sur $g(\mathbf{x})$, nous sélectionnons la paire (K, g) conduisant au minimum des critères BIC ou ICL corrigés :

$$\begin{aligned} \text{BIC}^*(K, g) &= -\mathcal{L}(g(\mathbf{x})|\hat{\theta}_{(K,g)}) + \frac{\nu_K}{2n} \ln(n) - \ln[\det(J_g)] \\ \text{ICL}^*(K, g) &= \text{BIC}^*(K, g) + \text{Ent}(K, g). \end{aligned} \quad (6.14)$$

Notez que dans ces expressions, le nombre de paramètres ν_K ne dépend pas de la transformation g . Pour utiliser ces critères corrigés pour comparer les transformations logit et arcsin définies par (6.13) et (6.12) respectivement, on évalue le terme $-\ln[\det(J_g)]$:

$$\begin{aligned} -\ln[\det(J_{\arcsin})] &= nd \ln(2) + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^d \ln[p_{ij}(1-p_{ij})], \\ -\ln[\det(J_{\text{logit}})] &= nd \ln[\ln(2)] + \sum_{i=1}^n \sum_{j=1}^d \ln[p_{ij}(1-p_{ij})]. \end{aligned}$$

6.3.1.4 Illustration

Nous illustrons notre procédure, implémentée dans la librairie Bioconductor `coseq`, sur des données RNA-seq issues d'une étude de l'expansion de 3 régions du néocortex (zone ventriculaire [VZ], zone sous-ventriculaire [SVZ] et la plaque corticale [CP]) chez cinq embryons de souris (Fietz *et al.*, 2012)¹. Les deux transformations logit et arcsin sont considérées et le nombre de classes K varie de 2 à 20. Pour ces données, le critère ICL sélectionne $\hat{K} = 12$ et $\hat{K} = 15$ pour les transformations arcsin et logit respectivement. Et d'après (6.14), le critère ICL corrigé sélectionne la transformation arcsin parmi les deux transformations.

Grâce aux mélanges gaussiens, on peut étudier les structures de covariance par classe estimées pour le modèle sélectionné ($\hat{K} = 12$ et transformation arcsin). Par exemple, les matrices de corrélation estimées pour les classes 2 et 3 sont présentées dans la Figure 6.4. Il est intéressant de noter que, bien que le modèle de mélange gaussien n'incorpore pas explicitement le design de l'expérience (lien entre les réplicats et les conditions expérimentales), les mélanges estimés retrouvent d'importantes corrélations spécifiques entre des réplicats au sein de chaque tissu. De plus, des structures de corrélation spécifiques aux classes entre les régions peuvent être clairement observées ; par exemple la classe 2 est caractérisée par des corrélations négatives

¹Les comptages pour les 5 réplicats pour chacun des 3 tissus ont été téléchargés le 23 décembre 2015 à partir du Digital Expression Explorer (DEE) (Ziemann *et al.*, 2015) en utilisant le numéro d'accèsion SRP013825 associé, et les informations d'exécution ont été téléchargées en utilisant le SRA Run Selector.

très importantes entre les régions CP et SVZ/VZ, tandis que la classe 3 présente au contraire une forte corrélation négative entre les régions VZ et CP/SVZ.

A l'aide de `coseq`, les profils d'expression par classe peuvent être représentés graphiquement de plusieurs façons, selon le type de données (comptages normalisés, profils d'expression normalisés ou profils normalisés transformés), le type de tracé (par exemple, tracés linéaires ou boxplots) et selon que les réplicats dans les conditions expérimentales sont moyennés ou tracés indépendamment. Dans notre exemple, on a opté pour des boxplots des profils par condition expérimentale, l'expression est moyennée sur les réplicats par condition. Les profils d'expression des 12 classes obtenues sont représentés en Figure 6.5. Par exemple, on peut constater que les gènes de la classe 5 s'expriment plus fortement dans la condition VZ, ceux de la classe 2 dans CP et ceux de la classe 7 dans SVZ. Les gènes des classes 4, 6 et 12 sont globalement non différentiellement exprimés dans les trois conditions.

Un autre avantage des modèles de mélange est de pouvoir évaluer la qualité de la classification obtenue en examinant les probabilités conditionnelles maximales d'appartenance pour chaque gène :

$$\tau_{\max}(i) = \max_{1 \leq k \leq K} \tau_{ik} \left(\hat{\theta}_{\hat{K}} \right), \quad i = 1, \dots, n.$$

Les boxplots des probabilités conditionnelles maximales $\tau_{\max}(i)$ par classe pour les données de Fietz *et al.* (2012) sont présentés en Figure 6.6. On peut constater que, dans l'ensemble des classes, la majorité des gènes ont une valeur élevée pour $\tau_{\max}(i)$ (82,4% des gènes ont $\tau_{\max}(i) > 0,8$). Cependant, les boxplots illustrent également que certains gènes ont un $\tau_{\max}(i)$ inférieur à au seuil 0,8; cela indique que pour un petit nombre de gènes, l'attribution des classes est assez ambiguë et que l'attribution à une seule classe est discutable. Dans de tels cas, il peut être prudent de concentrer l'attention sur les gènes dont l'assignation aux classes est très fiable. Enfin, l'examen de la distribution des valeurs $\tau_{\max}(i)$ au sein de chaque classe fournit des informations sur l'homogénéité et la pertinence de chaque classe.

6.3.2 Classification par K -means de profils logCLR-transformés

6.3.2.1 Motivations

Dans cette section, on revient sur les travaux publiés dans [GMR19]. On est ici reparti des pratiques usuelles de classification non supervisée pour des données compositionnelles. Une possibilité est l'utilisation de l'algorithme des K -means (MacQueen *et al.*, 1967) en modifiant le concept de proximité entre deux vecteurs du simplexe (Aitchison, 1986; Pawlowsky-Glahn et Egozcue, 2001). Une des distances les plus connues dans ce contexte est la distance d'Aitchison (Aitchison, 1982), qui revient à la distance euclidienne usuelle sur des vecteurs du simplexe transformés par le rapport logarithmique centré (CLR) :

$$\text{CLR} : \mathbf{p}_i = (p_{i1}, \dots, p_{id}) \in \mathbb{S}^d \mapsto \left(\ln \left(\frac{p_{i1}}{g_m(\mathbf{p}_i)} \right), \dots, \ln \left(\frac{p_{id}}{g_m(\mathbf{p}_i)} \right) \right) \in \mathbb{R}^d \quad (6.15)$$

où $g_m(\mathbf{p}_i)$ est la moyenne géométrique de \mathbf{p}_i .

Pour les données RNA-seq, il existe un grand nombre de gènes dont l'expression reste inchangée dans différentes conditions (c'est-à-dire dont les profils sont proches du centre du simplexe), et un nombre relativement faible de gènes dont l'expression est hautement spécifique à une seule condition (c'est-à-dire dont les profils sont proches d'un sommet du simplexe).

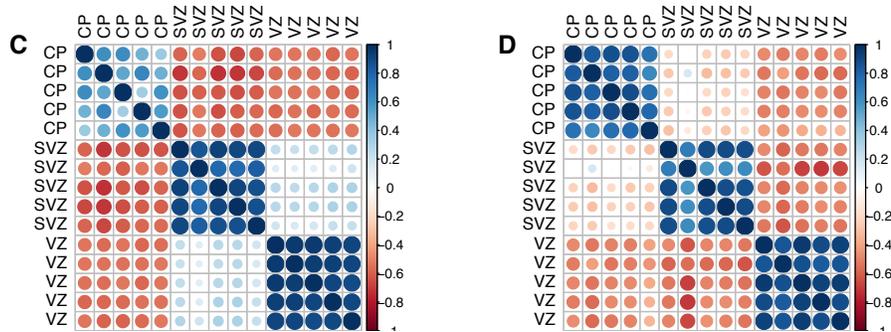


FIGURE 6.4 : Matrices de corrélation des gènes pour la classe 2 (à gauche) et la classe 3 (à droite) des données de Fietz et al. (2012). Le bleu foncé et le rouge foncé représentent des corrélations proches de 1 et de -1 respectivement, et les aires des cercles correspondent à la valeur absolue des coefficients de corrélation.

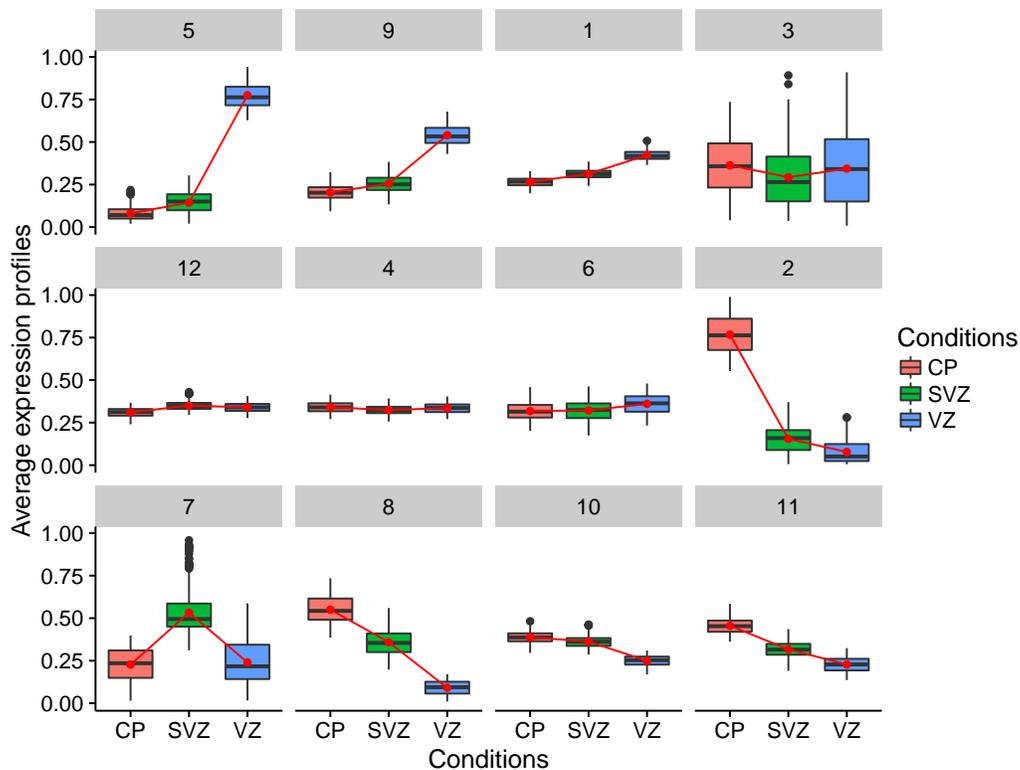


FIGURE 6.5 : Profils d'expression par classe pour les données de Fietz et al. (2012). Les lignes rouges connectées correspondent au profil d'expression moyen pour chaque classe. Les classes sont organisées de sorte que celles qui ont des vecteurs moyennes similaires (selon la distance euclidienne) soient tracées les unes à côté des autres.

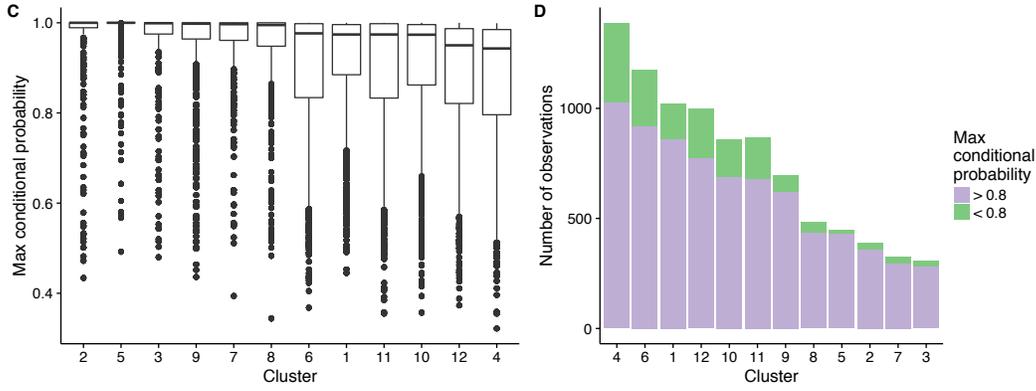


FIGURE 6.6 : Évaluation de la qualité de la classification pour les données de Fietz et al. (2012). A gauche : Probabilités conditionnelles maximales $\tau_{\max}(i)$ pour chaque classe, triées par ordre décroissant de la médiane de la classe. A droite : Barplot de l'effectif des classes, en fonction des $\tau_{\max}(i)$ supérieurs ou inférieurs à 0,8, triés en fonction du nombre de gènes avec $\tau_{\max}(i) > 0,8$.

Pour le problème de la co-expression, ce sont ces profils hautement spécifiques que nous souhaitons déterminer par une procédure de classification non supervisée. Or pour les données de dimension modérée, lorsqu'un grand nombre de coordonnées ont des proportions très faibles, la transformation CLR a tendance à être assez sensible aux petites fluctuations proches de zéro (voir une illustration sur des profils jouets en Figure 6.7). Cela peut avoir un effet indésirable important sur les résultats de classification lorsqu'un petit nombre d'observations ont des profils très spécifiques. Aussi une modification de cette transformation est envisagée dans notre contexte.

6.3.2.2 Procédure de classification

Pour tenir compte du phénomène évoqué précédemment, on souhaite modifier la transformation CLR afin de donner plus d'importance aux coordonnées ayant de grandes valeurs relatives. Nous avons proposé le rapport log centré (logCLR) défini par : pour tout $\mathbf{p}_i \in \mathbb{S}^d$, $\log\text{CLR}(\mathbf{p}_i) := (\log\text{CLR}(p_{i1}), \dots, \log\text{CLR}(p_{id}))$ où

$$\log\text{CLR}(p_{ij}) := \begin{cases} -[\ln(1 - \ln[p_{ij}/g_m(\mathbf{p}_i)])]^2 & \text{si } p_{ij}/g_m(\mathbf{p}_i) \leq 1, \\ (\ln[p_{ij}/g_m(\mathbf{p}_i)])^2 & \text{sinon.} \end{cases}$$

Le terme logarithmique supplémentaire lorsque $p_{ij}/g_m(\mathbf{p}_i) \leq 1$ accorde moins d'importance aux individus dont les proportions sont relativement faibles, tandis que le terme au carré facilite la concentration des profils proches du centre du simplexe $(\frac{1}{d}, \dots, \frac{1}{d})$ (voir à droite en Figure 6.7).

Pour la classification, nous allons ici considérer l'algorithme des K -means. De façon générale, pour des profils $(\mathbf{p}_i)_i$ transformés en utilisant une transformation $h(\cdot)$ et à K fixé, l'algorithme revient à minimiser

$$\text{SSE}_h(\mathcal{P}_K) := \sum_{k=1}^K \sum_{i \in \mathcal{C}_k} \|h(\mathbf{p}_i) - \mu_{k,h}\|_2^2,$$

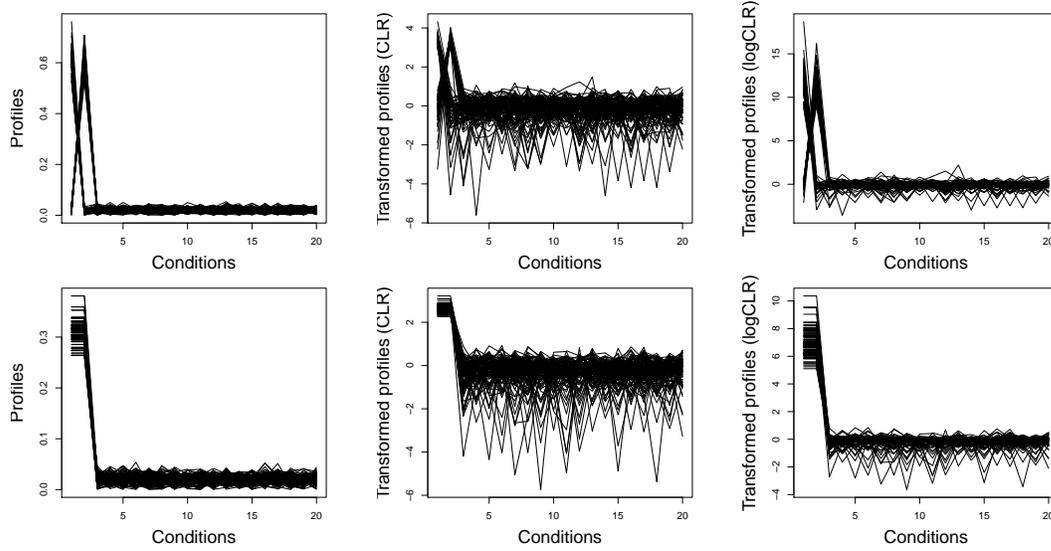


FIGURE 6.7 : *Profils jouets sur le simplexe pour illustrer le comportement des transformations CLR et logCLR. Les profils non transformés sont représentés à gauche et après transformation à l'aide de la CLR (au milieu) ou de la logCLR (à droite). Dans la première ligne, deux groupes de profils spécifiques à un seul échantillon (soit le premier, soit le second) sont considérés ; dans la seconde ligne, un seul groupe de profils spécifiques à la première paire d'échantillons est considéré.*

où $\mu_{k,h}$ est la moyenne des profils transformés appartenant à la classe \mathcal{C}_k :

$$\mu_{k,h} := \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} h(\mathbf{p}_i).$$

Pour sélectionner le nombre de classes pour l'algorithme des K -means, de nombreux critères existent (Caliński et Harabasz, 1974; Krzanowski et Lai, 1988), les plus courants sont la statistique Gap (Tibshirani *et al.*, 2001) et la maximisation du critère Silhouette (Kaufman et Rousseeuw, 2009). Cependant, dans notre cas, ces méthodes ont tendance à conduire au choix d'un modèle trop parcimonieux en pratique. Nous nous concentrons plutôt sur la méthode de sélection introduite par Fischer (2011), qui est basée sur la minimisation d'un critère pénalisé non asymptotique défini pour tout $K \in \mathbb{N}^*$ par

$$\text{CRIT}(K) = \sum_{i=1}^n \min_{k=1, \dots, K} \|h(\mathbf{p}_i) - \mu_{k,h}\|_2^2 + \text{pen}(K), \quad (6.16)$$

où h est la transformation utilisée. La fonction pénalité $\text{pen} : \mathbb{N} \rightarrow \mathbb{R}_+$ est définie par

$$\text{pen}(K) := a_h \sqrt{Knd}.$$

Le terme de pénalité est connu à une constante multiplicative près a_h , qui est calibrée en pratique par heuristique de pente implémentée dans la librairie `capushe` ([BMM12]).

En s'appuyant sur le fait que l'algorithme des K -means peut être vu comme un algorithme de classification EM sur des mélanges gaussiens sphériques, on peut par mimétisme définir des

probabilités conditionnelles que chaque gène i appartient à une classe \mathcal{C}_k de la façon suivante

$$\tau_{ik,h} := \frac{|\mathcal{C}_k| \phi \left(h(\mathbf{p}_i); \mu_{k,h}, \sigma_{k,h}^2 I_d \right)}{\sum_{k'=1}^K |\mathcal{C}_{k'}| \phi \left(h(\mathbf{p}_i); \mu_{k',h}, \sigma_{k',h}^2 I_d \right)},$$

où $\phi \left(\cdot; \mu_{k,h}, \sigma_{k,h}^2 I_d \right)$ est la densité de la loi gaussienne de vecteur moyenne $\mu_{k,h}$ et de matrice de covariance $\sigma_{k,h}^2 I_d$ avec $\sigma_{k,h}^2$ la variance intra-classe :

$$\sigma_{k,h}^2 = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \|h(\mathbf{p}_i) - \mu_{k,h}\|_2^2.$$

L'ensemble de cette procédure des K-means, avec la possibilité de choisir la transformation h parmi l'identité, CLR ou logCLR, est implémentée dans la librairie Bioconductor `coseq`. On retrouve les outils graphiques de visualisation des classes de profils présentés en section 6.3.1.4.

6.3.2.3 Illustration

Pour illustrer notre procédure, nous reprenons le jeu de données RNA-seq de Fietz *et al.* (2012) présenté en section 6.3.1.4. Nous appliquons l'algorithme des K-means avec la distance euclidienne sur les profils non transformés (h est l'identité) ou transformés par CLR ($h = \text{CLR}$) ou logCLR ($h = \text{logCLR}$) en faisant varier le nombre de classes entre 2 et 40. Le nombre de classes sélectionné par notre critère non asymptotique (voir Equation (6.16)) vaut 22 dans le cas de profils non transformés et 20 pour les transformations CLR et logCLR. En comparaison, le Gap statistic et le critère Silhouette retiennent 2 ou 3 classes, ce qui est non satisfaisant en pratique.

En examinant les groupes de profils identifiés pour chaque transformation, nous pouvons noter des différences qualitatives, en particulier pour les gènes ayant des profils très spécifiques. Les comparaisons entre les classes de profils analogues pour les données non transformées, transformées par CLR et par logCLR (voir Figure 6.8, à gauche) indiquent que les données non transformées et transformées par CLR ont tendance à produire un plus grand nombre de petites classes peu variables situées au centre du simplexe, correspondant à plusieurs classes regroupant des gènes globalement non différentiellement exprimés. Ces deux stratégies tendent également à donner des classes plus variables pour les profils proches des sommets du simplexe, notamment pour les gènes dont l'expression est élevée dans le tissu SVZ par rapport aux autres. Au contraire, l'utilisation de la transformation logCLR tend à produire des classes stables sur les bords et aux sommets du simplexe (c'est-à-dire, où l'expression dans un ou plusieurs des tissus est proche de zéro), et un plus petit nombre de classes variables au centre du simplexe. On retrouve cette stabilité des classes spécifiques et beaucoup plus d'incertitude dans les classes au centre du simplexe au travers des probabilités conditionnelles maximales (voir Figure 6.8, à droite). Cela suggère que la transformation logCLR conduit à une grande classe de gènes non différentiellement exprimés et privilégie l'identification de plusieurs classes de gènes au comportements très spécifiques.

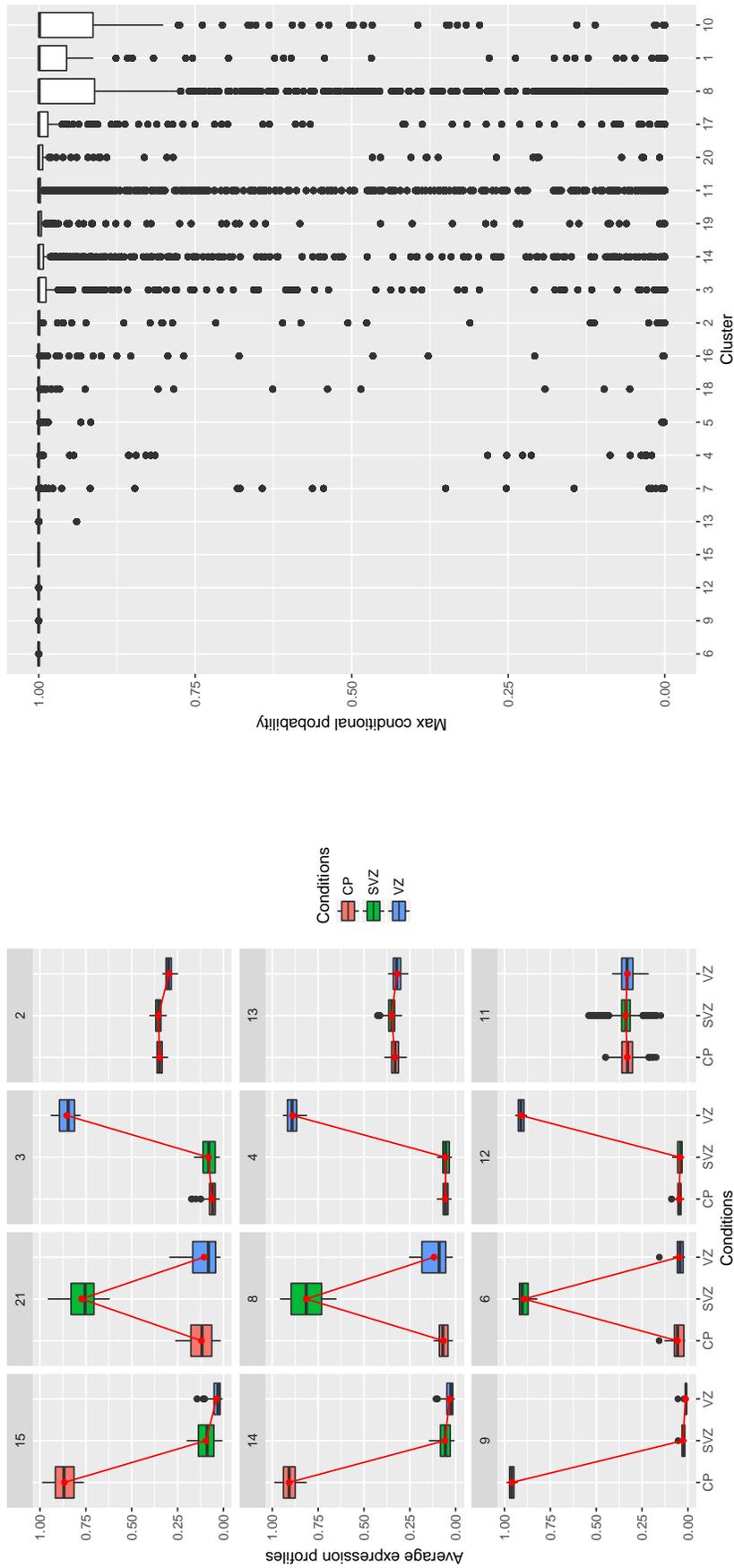


FIGURE 6.8 : A gauche : Exemples de classes de gènes issus des données RNA-seq de Fietz et al. (2012) : les profils d'expression normalisés de classes sélectionnées obtenues avec l'algorithme K-means sans transformation (en haut), et avec les transformations CLR (au milieu) et logCLR (en bas) sont présentés. Les classes analogues pour chaque approche ont été alignées verticalement. Les lignes rouges connectées correspondent aux profils moyens pour chaque condition. A droite : Probabilités conditionnelles maximales par classe des données RNA-seq de Fietz et al. (2012) obtenues avec l'algorithme K-means et la transformation logCLR. Les classes ont été triées par les probabilités conditionnelles médianes par classe.

6.4 Conclusions et perspectives

Les travaux résumés dans ce chapitre sont issus de nombreuses réflexions et tentatives de modélisation pour répondre à cette question de classification non supervisée à partir de données RNA-seq en s'adaptant à la nature de ces données, leur sparcité, le design des expériences... Dans ces tentatives, on atteint certaines limites des modèles de mélange pour ce type de données de séquençage. Il est difficile de choisir et d'estimer la distribution sous-jacente des données, cela impactant en particulier un mauvais comportement des critères asymptotiques usuels de sélection du nombre de classes. On a alors pris le parti de reformuler le problème en la classification de données compositionnelles et de choisir des transformations afin d'accentuer les profils atypiques. En pratique, on ne cherche par nécessairement à classer tous les gènes mais à mettre en lumière des classes de gènes co-exprimés homogènes et caractéristiques. La modélisation par modèle de mélange permet d'avoir accès aux probabilités conditionnelles d'appartenance qui sont un outil intéressant d'analyse de l'homogénéité des classes.

Plusieurs problèmes en traitement de données omiques peuvent être reformulés en l'étude de données compositionnelles, par exemple les données de microbiome. Au delà des questions biologiques posées, le problème de classification non supervisée sur des données compositionnelles pourrait être étudié plus en profondeur et surtout dans le contexte de matrices de données creuses.

Et d'un point de vue très macroscopique, l'évolution des données biologiques, leur complexité, l'intégration de différentes sources d'informations... nous amène à devoir aborder des problèmes de classification non supervisée dans des espaces de plus en plus vastes, avec des géométries particulières... On abordera dans le chapitre 7 des problèmes de classification pour des données de séquençage à haut-débit sur cellule unique. Dans le cadre du projet DEFIANT, on souhaite classer des conformations de molécules à partir de données angulaires et de positions, en prenant en compte des invariants géométriques et des contraintes dues à la structure du support.

Quelques challenges pour l'analyse de données single-cell RNA-seq

Résumé Des perspectives de recherche dans le cadre de l'analyse de données single-cell RNA-seq sont présentées dans ce chapitre. Ces réflexions ont été impulsées par le (co)-encadrement de plusieurs stages et de la thèse en cours de Nicolas Enjalbert-Courrech (voir page 89) sur cette thématique, la coordination du projet TTIL SingleCell et l'implication dans le projet DDisc.

Sommaire

7.1. Contexte	75
7.2. Classification non supervisée de cellules	77
7.3. Détection des gènes marqueurs	80
7.4. Analyse différentielle	83

7.1 Contexte

Les technologies de mesure d'expression à haut débit ont progressé ces dernières années. Alors que les techniques de type bulk-RNA-seq (voir Section 6.1.1) donnent seulement accès à des mesures moyennes d'expression des gènes sur l'ensemble des cellules d'un échantillon biologique considéré, les techniques plus récentes appelées single-cell RNA-seq (scRNAseq) permettent de mesurer l'expression des gènes à l'échelle de la cellule unique (voir Figure 7.1). Le principe du single-cell RNA-seq consiste à isoler les différentes cellules d'un échantillon biologique et à les séquencer séparément ce qui permet par exemple de différencier les cellules à partir des ARN présents dans chacune d'elles. Comme pour le bulk RNA-seq, on se retrouve alors avec un tableau de comptages d'expression mais en ajoutant ici la "dimension cellules" : on n'a pas une colonne de valeurs par condition biologique mais un tableau gènes par cellules (donc de grande dimension). Ce tableau est globalement sparse (environ 80% des données sont des zéros et ces zéros sont principalement des non-détections d'expression de gènes (dropouts)), la taille de la librairie (somme des comptages pour chaque cellule) est très variable d'une cellule à l'autre... Les méthodes statistiques développées pour l'étude des données bulk RNAseq

ne sont pas adaptées à la prise en compte de l'hétérogénéité due aux cellules, de nouveaux développements statistiques sont donc nécessaires pour l'analyse des données scRNAseq.

Le nombre de publications et de logiciels dédiés au traitement des données scRNAseq croît exponentiellement depuis ces sept dernières années (voir par exemple sur le site scRNA-tools¹). Les méthodes proposées couvrent toutes les étapes usuelles d'analyse dont

- la *normalisation* des données : elle permet de corriger les biais techniques et rendre comparable les expressions des différentes cellules. On peut citer entre autre les méthodes proposées dans Bacher et Kendzioriski (2016), Ding *et al.* (2017), McCarthy *et al.* (2017) et Vallejos *et al.* (2017). Cette étape est liée à la façon dont les données sont générées, il est donc important d'instaurer en amont des protocoles rigoureux sur les plateformes de séquençage et dans le traitement bio-informatique, ainsi que d'adapter cette normalisation aux spécificités de chaque plan d'expériences. Il est d'usage en pré-traitement de filtrer les gènes peu/pas exprimés dans les cellules et de privilégier les gènes hautement variants afin de réduire en amont le jeu de données.
- la *réduction de dimension* : elle permet la visualisation des données et c'est une étape souvent requise préliminairement par d'autres méthodes. L'ACP n'ayant pas de très bons comportements sur ces données (en particulier en raison de la répartition et la proportion des comptages faibles), la méthode t-SNE (*t-distributed Stochastic Neighbor Embedding*) (Maaten et Hinton, 2008) a été privilégiée dans un premier temps avant d'être concurrencée par l'UMAP (*Uniform Manifold Approximation and Projection*) (Becht *et al.*, 2018) en autre. Un exemple de représentation de données scRNAseq en t-SNE et en UMAP est donné en Figure 7.2. Pour une plus large étude sur le sujet, on peut se reporter aux articles de Sun *et al.* (2019a) et Raimundo *et al.* (2020) en autre.
- la *classification non supervisée* des cellules : C'est une étape fondamentale dans l'étude des données scRNAseq. On ne connaît pas a priori la nature des cellules isolées et séquencées. Dans une condition expérimentale donnée, un premier challenge est de déterminer les types cellulaires en présence à partir de leur profil d'expression. On reviendra sur cette question de clustering des cellules en Section 7.2.
- la *détection des gènes marqueurs* pour chaque classe de cellules : Une fois des classes de cellules obtenues, il est nécessaire de déterminer des gènes dont l'expression se différencie pour un groupe de cellules par rapport aux autres afin de caractériser le groupe cellulaire. Mais cette détection s'effectue à partir du même jeu de données qui a servi à construire le clustering de cellules. Ce problème de test post-clustering sera abordé en Section 7.3.

On retrouve ces différentes étapes dans les logiciels d'analyse pour les données scRNAseq dont les plus utilisés sont Seurat² (Butler *et al.*, 2018) et Scanpy³ (Wolf *et al.*, 2018). Selon les projets biologiques, d'autres questions peuvent être abordées comme

- *l'analyse différentielle* entre plusieurs conditions expérimentales : l'objectif est ici de détecter les gènes qui ont une différence d'expression entre plusieurs conditions biologiques comme en bulk RNA-seq mais en tenant compte ici de l'hétérogénéité due aux cellules en scRNAseq. Cette question sera développée en Section 7.4.

¹scRNA-tools : <https://www.scrna-tools.org/>

²<https://satijalab.org/seurat/>

³<https://scanpy.readthedocs.io/>

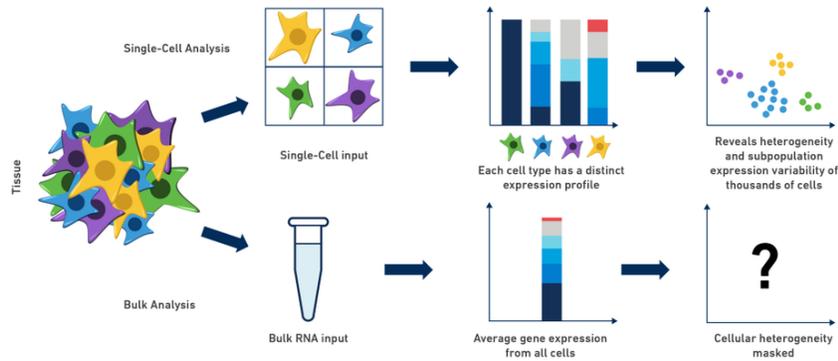


FIGURE 7.1 : Schéma résumant la différence entre des données bulk RNA-seq et des données single-cell RNA-seq. Le single-cell RNA-seq permet d'avoir accès aux profils d'expression hétérogènes des cellules alors que l'on en observe qu'un profil moyen en bulk RNA-seq. (Source : site web de 10X Genomics)

- *l'inférence de trajectoires* (appelée *pseudotime reconstruction*) : l'objectif est d'inférer une trajectoire entre groupes de cellules afin de reconstruire une dynamique cellulaire. Cette problématique ne sera pas abordée dans ce chapitre, on peut se référer aux articles de Cannoodt *et al.* (2016) et Saelens *et al.* (2019) proposant un état de l'art sur le sujet.
- *la détection d'abondance différentielle de groupes cellulaires* : l'objectif est de déterminer les groupes cellulaires dont l'abondance varie entre plusieurs conditions biologiques. Cette problématique ne sera pas abordée dans ce chapitre, on peut citer les travaux récents de Cao *et al.* (2019) et Zhao *et al.* (2021) entre autre.

Dans la suite, on va se focaliser sur quelques challenges liés à l'étude des données single-cell RNA-seq. Pour illustrer certains propos, on considère les données IFNB issues du package SeuratData (Satija *et al.*, 2019) de R. Ces données sont issues de prélèvements sur 8 patients atteints de lupus. Les PBMC⁴s4 de certains de ces patients ont été activées par une molécule d'IFN- et d'autres non, et les cellules ont été collectées au bout de 6h. On a donc ici 2 matrices de données scRNAseq X^{CTRL} et X^{STIM} qui correspondent à l'expression des cellules sous contrôle (CTRL) et des cellules stimulées (STIM) respectivement. L'étape de contrôle de la qualité des cellules a été faite et les données ont été ensuite normalisées à l'aide du package Seurat. On conserve la même notation pour les matrices après normalisation pour simplifier. Au final, les données normalisées contiennent $G = 14053$ gènes et $C = 13999$ cellules en tout, dont $C_1 = 7451$ cellules pour STIM et $C_2 = 6548$ pour CTRL.

7.2 Classification non supervisée de cellules

Dans un premier temps, on se place dans une condition biologique et on observe donc un tableau de données single-cell RNA-seq (par exemple les données X^{STIM} pour la condition STIM). L'objectif consiste à faire une classification non supervisée des cellules afin de construire des groupes cellulaires de profil d'expression similaire. De nombreuses méthodes ont été vite proposées car c'est une étape clé pour l'analyse de données scRNAseq. On peut citer parmi les

⁴PBMC : Peripheral Blood Mononuclear Cell

articles faisant un état de l'art ceux de Duò *et al.* (2018), Freytag *et al.* (2018) et Zhang *et al.* (2020). Pour montrer la variété des méthodes proposées, on peut citer SC3 (Kiselev *et al.*, 2017) qui construit une classification hiérarchique à partir d'une matrice consensus obtenue via plusieurs clusterings de type K-means effectués sur les différents plans de l'ACP. Cette méthode n'accepte que 5000 cellules au maximum, elle réaffecte les cellules restantes a posteriori dans les classes construites. Dans le package Seurat, l'étape de clustering est basée sur un graphe SNN (Shared Nearest Neighbor) à partir des K plus proches voisins selon la distance euclidienne dans l'espace de l'ACP. Le groupement des cellules se fait itérativement par un algorithme de Louvain en optimisant une fonction de modularité, cet algorithme nécessite de préciser un paramètre de résolution (et non le nombre de classes). Parmi les méthodes de clustering hiérarchiques, la méthode DendroSplit (Zhang *et al.*, 2018) construit une hiérarchie à partir du lien complet et une distance basée sur la corrélation de Pearson. Le découpage en classes est alors piloté par un score de séparation comme métrique entre deux classes. On trouve également des méthodes bayésiennes comme BAMMSC (Sun *et al.*, 2019b) pour "Bayesian hierarchical Dirichlet multinomial mixture model", des méthodes de deep clustering comme scDeepCluster (Tian *et al.*, 2019) qui utilise un auto-encodeur avec une fonction de perte basée sur la vraisemblance d'une distribution ZINB (distribution binomiale-négative avec prise en compte des dropout) et une fonction de perte pour la partie clustering basée sur la divergence de Kullback-Leibler.

Lors du stage de Master 1 de Fanny Mollandin (voir page 89), plusieurs méthodes de clustering (SC3, pcreduce, DendroSplit, Seurat et scDeepCluster) ont été testées et comparées sur un même jeu de données scRNAseq pour l'étude de cellules stromales mésenchymateuses chez la souris, issu d'une collaboration avec Marielle Ousset (ReStore). La calibration de certaines méthodes et la sélection du nombre de classes étaient très sensibles, les classifications obtenues étaient parfois très différentes (grande variabilité des valeurs de l'Adjusted Rand Index), la méthode SC3 ne retrouvait même pas un groupe de cellules spécifiques facilement détectable dans les données. En Figure 7.3, on peut observer comment les cellules des 13 classes obtenues par la méthode Seurat (au centre) se répartissent dans les autres classifications. Ainsi, les cellules de la classe n°12 (qui était facilement détectable dans les données) restent classées ensemble par les autres méthodes sauf SC3 qui les répartit dans deux classes (et avec d'autres cellules ce que l'on ne voit pas sur cette représentation). Au contraire d'autres classes comme la n° 10 sont moins stables dans les autres classifications. Afin d'aider les biologistes à comparer et interpréter les classifications obtenues, l'application shiny ASTEC-sc⁵ a été développée par Nicolas Enjalbert-Courrech lors de son stage de Master 1 (voir page 89).

Le problème se complexifie lorsque l'on doit étudier plusieurs jeux de données scRNAseq issus de plusieurs conditions biologiques (avec de potentiels réplicats). La stratégie la plus largement utilisée actuellement consiste à débiter par une étape d'intégration des jeux de données, au sens de les plonger dans un espace commun pour ensuite faire une méthode de clustering sur le mélange de toutes les cellules dans les différentes conditions. C'est par exemple la stratégie proposée dans Seurat, où les jeux de données sont intégrés deux à deux. Après une analyse canonique des corrélations, des "anchors" (paires de cellules, une de chaque condition) sont recherchées par mutual NN afin de recalibrer un jeu de données sur l'autre. On procède alors à la classification de toutes les cellules dans le sous-espace commun. Dans la même idée, on peut aussi citer les méthodes scmap (Kiselev *et al.*, 2018), Harmony (Korsunsky *et al.*, 2019), Scanorama (Hie *et al.*, 2019). La difficulté de ces méthodes est de trouver un bon compromis de

⁵<https://cmaugis.github.io/ASTEC-sc/>

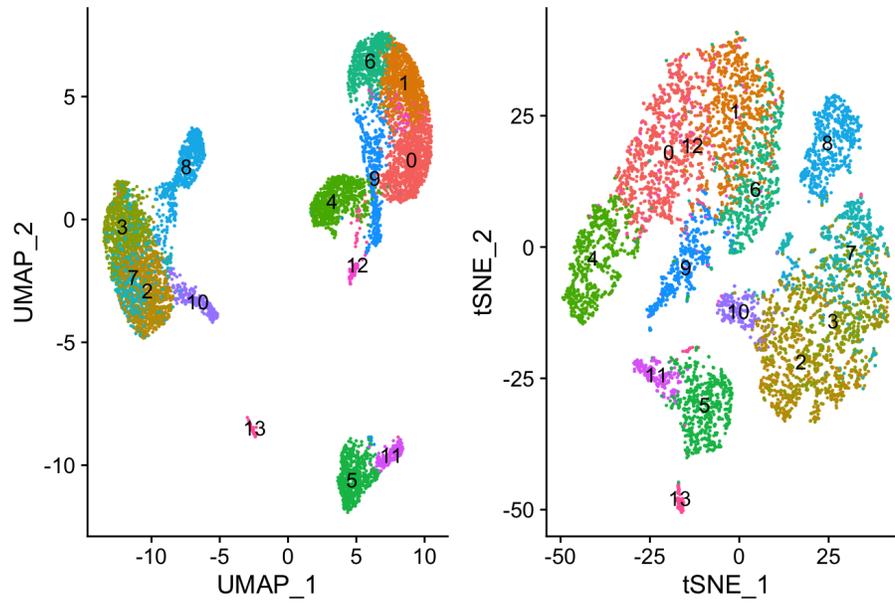


FIGURE 7.2 : Représentation des cellules du jeu de données IFNB dans la condition STIM dans le premier plan de la réduction UMAP (à gauche) et t-SNE (à droite).

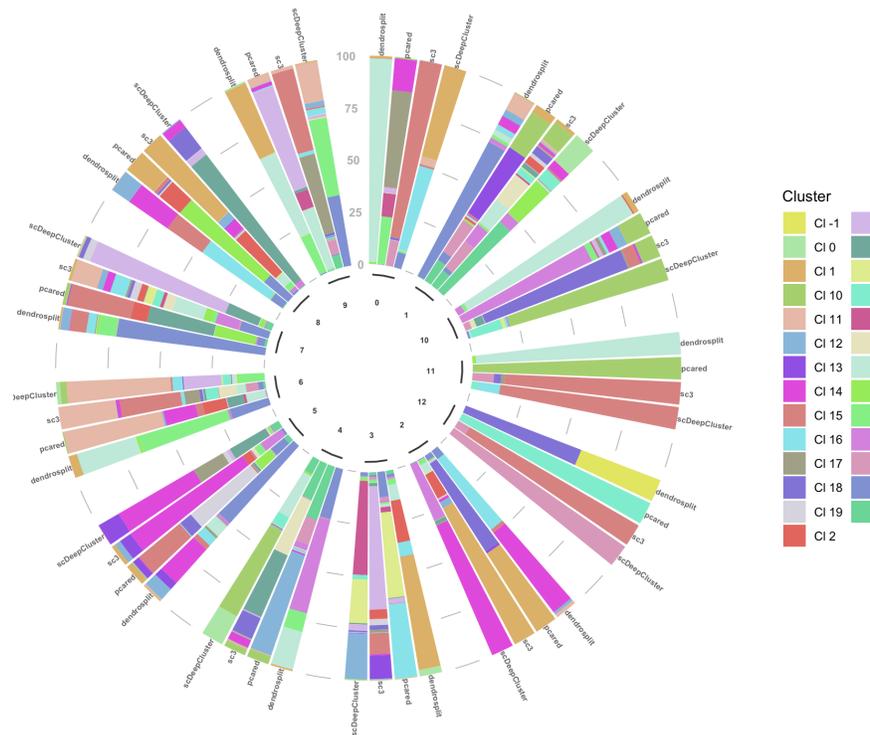


FIGURE 7.3 : Pour chacune des 13 classes obtenues avec Seurat pour les données stromales mésenchymateuses, on observe comment les cellules se répartissent dans les autres classes des autres méthodes de clustering.

recalage pour atténuer les effets techniques tout en maintenant la variabilité biologique que l'on souhaite étudier. Une seconde stratégie consiste à étudier séparément les jeux de données et l'on cherche ensuite à retrouver des groupes cellulaires identiques dans les différentes conditions. C'est la stratégie de ClusterMap (Gao *et al.*, 2019) qui fait une classification hiérarchique sur les groupes cellulaires des différentes conditions avec un lien moyen et une distance de Jaccard à partir d'une matrice binaire fondée sur les gènes marqueurs des différentes classes de cellules. Mais comme on le verra en Section 7.3, la détection des gènes marqueurs est également un point délicat.

Sur cette question de la classification non supervisée des cellules, plusieurs pistes de travail sont envisagées :

- Développer des méthodes de clustering soft dans le but de contrôler (par des poids, le filtrage de probabilités conditionnelles d'appartenance, ...) le niveau de classification des cellules et qui prennent en compte les caractéristiques des données scRNAseq (la dimensionnalité, la sparsité de l'information, ...). L'idée est de privilégier des classes de cellules plus stables et homogènes, quitte à ne pas toutes les classer ou de prendre en compte le poids des cellules dans la suite des analyses ;
- Envisager une méthode de consensus de clusterings afin de privilégier des classes cellulaires stables ;
- Développer des méthodes de clustering qui intègrent le design des données (les différentes conditions biologiques, les potentiels réplicats) afin d'éviter l'utilisation d'une méthode de recalage préliminaire à l'étape de clustering ;
- Creuser le délicat choix du nombre de classes qui est actuellement peu abordé dans les méthodes les plus utilisées dans la communauté mais qui impacte fortement la suite de l'analyse de ces données scRNAseq.

7.3 Détection des gènes marqueurs

Partant d'une classification des cellules (voir Section 7.2), on souhaite déterminer les gènes dont l'expression diffère entre une (un groupe de) classe(s) \mathcal{C}_1 et une autre (un autre groupe de) classe(s) \mathcal{C}_2 pour "identifier" le type cellulaire en présence. On souhaite donc tester pour chaque gène g si

$$\mathcal{H}_0^{(g)} : \bar{X}_g^{[1]} = \bar{X}_g^{[2]} \text{ contre } \mathcal{H}_1^{(g)} : \bar{X}_g^{[1]} \neq \bar{X}_g^{[2]} \text{ avec } \bar{X}_g^{[j]} = \frac{1}{|\mathcal{C}_j|} \sum_{c \in \mathcal{C}_j} X_{cg}, \quad j = 1, 2$$

où X_{cg} est la mesure d'expression du gène g pour la cellule c .

Actuellement, on trouve dans les routines d'analyse de données scRNAseq des tests adaptés de l'analyse différentielle en bulk RNA-seq comme DESeq2, des tests basés sur des modèles linéaires généralisés de Poisson ou de binomiale négative, le test des rangs de Wilcoxon... et les p -valeurs sont corrigées par une procédure de Bonferroni. Dans notre exemple, on a obtenu une classification des cellules de la condition STIM en 14 classes (voir Figure 7.2) avec le package Seurat pour un paramètre de résolution de 1. Si l'on s'intéresse à la détection des gènes marqueurs pour la classe n°13 (\mathcal{C}_1) par rapport à toutes les autres classes (\mathcal{C}_2), on obtient 885 gènes ayant une p -valeur issue du test de Wilcoxon corrigée par Bonferroni inférieure à 0.05. En

Figure 7.4, la distribution des p -valeurs et l'évolution du nombre de gènes déclarés marqueurs en fonction du risque montrent une détection trop importante, phénomène récurrent dans les analyses de données scRNAseq. On est en présence d'un biais de sélection, dû au fait que le clustering des cellules et les tests statistiques sont effectués sur les mêmes données, conduisant à un risque accru de faux positifs (Lähnemann *et al.*, 2020).

Face à cette difficulté, les routines d'analyse proposent des indicateurs complémentaires comme le log fold-change des expressions moyennes entre \mathcal{C}_1 et \mathcal{C}_2 , la différence des pourcentages de cellules où le gène g s'exprime dans \mathcal{C}_1 et dans \mathcal{C}_2 , la valeur de la courbe ROC (AUC)... Mais leur utilisation nécessite de faire un choix de seuil pour chacun et aucune garantie statistique n'est possible. Une visualisation interactive est proposée dans l'application shiny ASTEC-sc afin de sensibiliser les biologistes à l'impact du choix des différents seuils sur la liste de gènes marqueurs. Dans notre exemple, l'utilisation du critère AUC avec un seuil $\tau = 0.8$ amène à la sélection de 8 gènes marqueurs mais on peut voir en Figure 7.4 la sensibilité du choix du seuil.

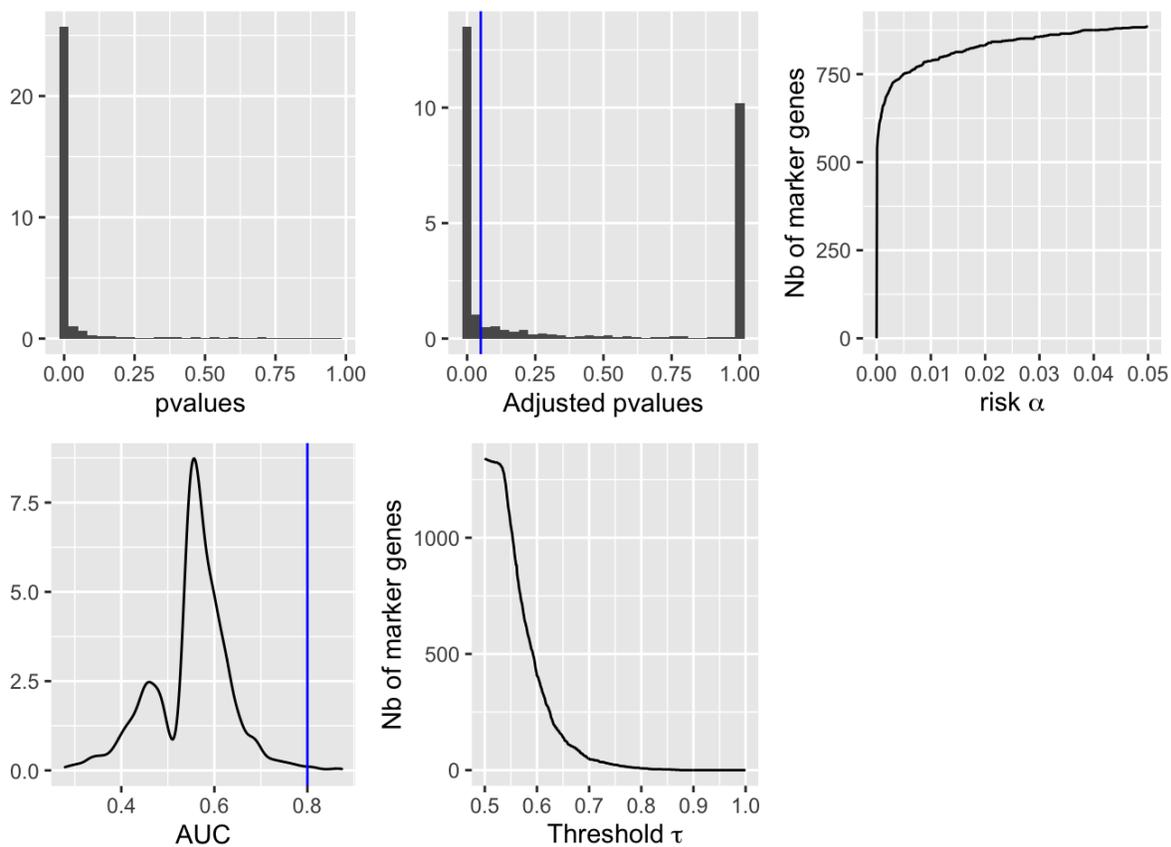


FIGURE 7.4 : Détermination des gènes marqueurs de la classe n°13 par rapport aux autres classes obtenues avec Seurat. En haut, distribution des p -valeurs du test de Wilcoxon (à gauche), ajustées (p_g) par Bonferroni avec barre verticale bleue pour marquer le risque $\alpha = 0.05$ (au milieu), et nombre de gènes déclarés marqueurs ($p_g < \alpha$) en fonction de la valeur du risque α (à droite). En bas, distribution des valeurs de l'AUC avec barre verticale bleue pour le seuil $\tau = 0.8$ (à gauche) et le nombre de gènes déclarés marqueur ($AUC_g > \tau$) en fonction du seuil τ (à droite).

Le challenge est donc ici de corriger les p -valeurs des procédures de test pour tenir compte de ce phénomène de test post-clustering. Du point de vue statistique, ce problème "classification + test" rentre dans le cadre de l'inférence post sélection de modèle. C'est un domaine de la statistique étudié depuis une dizaine d'années, principalement pour des problèmes de régression (inférence conditionnelle, voir Lee *et al.* (2016); méthodes PoSI, voir Berk *et al.* (2013)) ou de tests multiples (Goeman et Solari, 2011). En particulier, dans la recherche de gènes différentiellement exprimés entre deux conditions à partir de données de puces à ADN ou de bulk RNA-seq, on peut utiliser les travaux de Blanchard *et al.* (2020) pour contrôler le nombre de faux positifs dans toute liste de gènes choisis par l'utilisateur. Cependant, ces méthodes ne s'appliquent que dans le cas où les groupes comparés sont connus à l'avance. L'extension des méthodes d'inférence conditionnelle au cadre test post-clustering requiert des développements statistiques dédiés.

Du côté des méthodes cherchant à répondre à ce problème pour les données scRNAseq, peu de choses existent à ma connaissance. Zhang *et al.* (2019) proposent le "TN-test". L'idée est de couper le jeu de données en deux parties. La première sert à estimer un hyperplan pour séparer les données en deux classes. La labélisation des cellules de la seconde partie est faite vis-à-vis de l'hyperplan estimé. Un test basé sur une loi gaussienne tronquée est alors construit pour détecter les gènes ayant une différence d'expression entre les deux classes. Cette méthode a pour limitations qu'elle n'est valable que pour deux classes, que le découpage aléatoire initial n'assure pas en scRNAseq que les groupes cellulaires soient bien représentés dans les deux sous-ensembles de données et des premiers essais sur données simulées gaussiennes (car la procédure est basée sur cette hypothèse) n'ont pas été convaincants. Dans un contexte de validation de classes, Gao *et al.* (2020) proposent une construction de p -valeurs corrigées pour tester si deux classes issues d'une procédure de clustering sont bien distinctes. Les données de départ sont supposées gaussiennes et les auteurs évaluent des p -valeurs avec un conditionnement plus fort faisant intervenir le clustering obtenu mais aussi la direction portée par les deux centres des classes comparées et la projection orthogonale des données à cette droite. Cela permet de reformuler la p -valeur corrigée en une probabilité sur une variable aléatoire du chi-deux tronquée sur un ensemble \mathcal{S} qui regarde le maintien des deux classes par la procédure de clustering choisie sur les données translatées. Ils obtiennent une expression explicite dans quelques cas de classification hiérarchique. Pour d'autres procédures de clustering, il est nécessaire de faire une approximation de type Monte-Carlo ce qui implique de relancer la méthode de clustering sur plusieurs versions translatées de données. Cette procédure a été testée lors de deux stages de Master 1 sur des données scRNAseq mais c'est computationnellement intenable si on n'a pas une expression explicite de la p -valeur corrigée. Il est aussi nécessaire d'avoir accès au choix du nombre de classes du clustering (ce qui n'est pas le cas dans la procédure de clustering de Seurat par exemple). Par contre, cette stratégie est mathématiquement adaptable à la question de la détection de gènes marqueurs.

Face à ce challenge, nous souhaitons exploiter les progrès récents en inférence post sélection de modèle (Lee *et al.*, 2016; Kuchibhotla *et al.*, 2020) pour proposer des méthodes d'identification de gènes marqueurs offrant des garanties statistiques. Une première perspective consiste à se placer dans le cadre de la classification convexe (Pelckmans *et al.*, 2005) où l'on souhaite résoudre le problème d'optimisation général suivant

$$\widehat{B}^{(\lambda)} = \underset{B \in \mathbb{R}^{C \times G}}{\operatorname{argmin}} \frac{1}{2} \sum_{c=1}^C \|X_c - B_c\|_2^2 + \lambda \sum_{c < c'} \omega_{cc'} \|B_c - B_{c'}\|_q$$

où X_c est le vecteur des mesures d'expression pour la cellule c . Deux cellules c et c' sont alors dans la même classe si $\widehat{B}_c^{(\lambda)} = \widehat{B}_{c'}^{(\lambda)}$. Dans le cas de la pénalité l_1 ($q = 1$), ce problème se reformule en un problème de régression lasso généralisée. On souhaite alors s'inspirer des travaux de Tibshirani et Taylor (2011) et Lee *et al.* (2016) pour proposer une procédure de test adaptée. D'un point de vue computationnel, il faut réussir à résoudre le problème de clustering convexe à l'échelle des données scRNAseq. Ensuite les travaux de Tibshirani et Taylor (2011) s'appuie sur le lemme polyhedral qui est fondé sur une hypothèse forte de loi gaussienne. Le défi sera alors de combler le gap entre les résultats théoriques et l'application aux données scRNAseq. Le travail sur les log-comptages ne sera certainement pas suffisant, la prise en compte des nombreux zéros devra en particulier être envisagée. Et à plus long terme, trouver des stratégies pour répondre à cette question, adaptables à tout type de test et sans hypothèse forte sur la distribution des données. Un autre axe pourrait être de traiter un problème de biclustering et de tester ensuite la cohérence des blocs avec une correction adaptée.

7.4 Analyse différentielle

L'analyse différentielle a pour objectif de détecter les gènes différentiellement exprimés entre plusieurs conditions biologiques. Cette question a déjà été adressée avec les données d'expression issues des puces à ADN puis de données bulk RNA-seq. Dans le cadre des données scRNAseq, le défi supplémentaire est de tenir compte de l'hétérogénéité cellulaire. En effet, en scRNAseq les données sont multimodales avec un nombre élevé de zéros. Aussi une distribution différentielle (DD) pour un gène peut être issue de différentes situations (voir Figure 7.5) : une expression différentielle "classique" (DE), une proportion différentielle de cellules dans chaque condition (DP), un différentiel de modalité (DM), une expression différentielle et un différentiel de modalité en même temps (DB), une proportion différentielle de zéros (DZ).

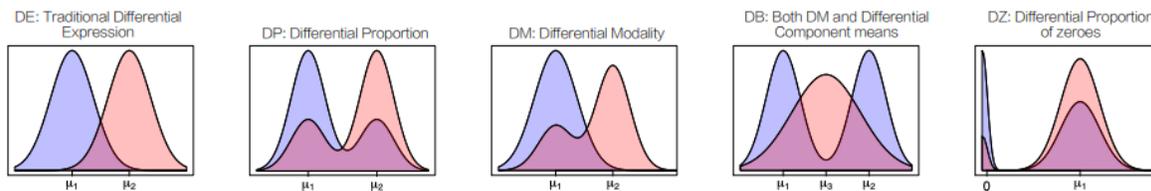


FIGURE 7.5 : Schéma des modèles de distribution différentielle possibles : expression différentielle (DE), une proportion différentielle de cellules dans chaque condition (DP), un différentiel de modalité (DM), DM et DE en même temps (DB), une proportion différentielle de zéros (DZ) .

Plusieurs méthodes ont été proposées ces dernières années pour l'analyse différentielle sur des données scRNAseq pour essayer d'intégrer certains aspects parmi les expressions multimodales, la proportion de zéros, les réplicats par condition, la comparaison de plus de deux conditions, ... Parmi elles, SCDE (Kharchenko *et al.*, 2014) et MAST (Finak *et al.*, 2015) s'appuient sur une modélisation en deux parties afin d'isoler la contribution des comptages nuls. DEsingle (Miao *et al.*, 2018) utilise un modèle de régression ZINB (binomiale négative avec zéro inflaté) et classe les gènes DE en trois catégories. scDD (Korthauer *et al.*, 2016) utilise une modélisation bayésienne (processus de Dirichlet à mélange de gaussiennes) pour déterminer les gènes DD et les classer selon les catégories DE, DP, DM ou DB. EMDomics

(Nabavi *et al.*, 2016) identifie les gènes DE en utilisant une distance entre les distributions dans deux conditions pour les gènes. SigEMD (Wang et Nabavi, 2018) est une extension intégrant une estimation de la distribution des comptages nuls. IDEAS (Zhang *et al.*, 2022) estime la distribution d’expression d’un gène par une méthode paramétrique (ZINB ou NB) ou non-paramétrique et construit un test à partir d’une distance de Wasserstein entre les distributions de deux conditions biologiques. Les méthodes distinct (Tiberi *et al.*, 2021) et ccdf (Gauthier *et al.*, 2021) se basent sur les écarts des fonctions de répartitions empiriques. La première construit un test à partir des écarts des valeurs moyennes des fonctions de répartition empirique en plusieurs points entre deux conditions, nécessitant donc des réplicats par condition. La seconde reformule le problème en un test de nullité de paramètres d’un modèle linéaire généralisé, permettant la prise en compte de covariables et de plusieurs (> 2) conditions mais pas de réplicats. Plusieurs études ont comparé des méthodes d’analyse différentielle dédiées au scRNAseq entre elles et aussi avec des méthodes bien connues pour les données bulk-RNAseq. On peut citer en outre Miao et Zhang (2016), Jaakkola *et al.* (2016), Sonesson et Robinson (2018) et Wang *et al.* (2019).

Ces différentes méthodes pour la détection de gènes différentiellement exprimés tiennent compte chacune de certains aspects parmi la multimodalités, l’hétérogénéité et/ou les nombreux comptages nuls dans les données, de designs complexes (nombre de conditions à comparer et réplicats potentiels). Du développement méthodologique est encore nécessaire pour couvrir plus largement ces aspects globalement. Cette question est au coeur du projet DDisc puisque les trois projets biologiques impliqués nécessitent des comparaisons de cinétiques avec parfois des réplicats et chacun étudie un organisme différent : cinétique de réparation tissulaire chez la souris ; développement embryonnaire du porc ; influence de micro-organismes sur le développement de la luzerne tronquée. Se pose aussi la question du bénéfice ou non d’intégrer la connaissance d’un clustering obtenu préliminairement sur les cellules. Cela renvoie à la problématique de tests post-clustering comme expliquée pour la détection de gènes marqueurs (voir Section 7.3).

Liste des contributions scientifiques

Articles dans revues internationales à comité de lecture

- [DGMM21] De Castro, Y., Gadat, S., Marteau, C. and **Maugis-Rabusseau, C.** (2021) Supermix : sparse regularization for mixtures. *Annals of Statistics*, 49(3), 1779-1809.
- [A21] Arabaciyan, S. and Saint-Antoine, M. and **Maugis-Rabusseau, C.** and François, J.-M. and Singh, A. and Parrou, J.-L. and Capp, J.-P. (2021). Insights on the control of yeast single-cell growth variability by members of the Trehalose Phosphate Synthase (TPS) complex. *Frontiers in cell and developmental biology*, (9).
- [GMR20] Godichon-Baggioni, A., **Maugis-Rabusseau, C.** and Rau, A. (2020) Multi-view cluster aggregation and splitting with an application to multi-omic breast cancer data. *Annals of Applied Statistics*, 14(2), 752-767.
- [GKMM20] Gadat, S., Kahn, J., Marteau, C. and **Maugis-Rabusseau, C.** (2020) Parameter recovery in two-component contamination mixtures : the L2 strategy. *Annales de l'IHP*, 56(2), 1391-1418.
- [GMR19] Godichon-Baggioni, A., **Maugis-Rabusseau, C.** and Rau, A. (2019) Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. *Journal of Applied Statistics*, 46(1), 47-65.
- [CMS19] Celeux, G., **Maugis-Rabusseau, C.** and Sedki, M. (2019). Variable selection in model-based clustering and discriminant analysis with regularization approach. *Advances in Data Analysis and Classification*, 13 (1), 259-278.
- [LMM18] Laurent, B., Marteau, C. and **Maugis-Rabusseau, C.** (2018). Multidimensional two-component Gaussian mixtures detection. *Annales de l'IHP, Probabilités et Statistiques*, 54 (2), 842–865.
- [RM18] Rau, A. and **Maugis-Rabusseau, C.** (2018) Transformation and model choice for RNA-seq co-expression analysis. *Briefings in bioinformatics*, 19(3), 425-436.
- [R18] Rigauill, G., Balzergue, S., Brunaud, V., Blondet, E., Rau, A., Rogier, O., Caius, J., **Maugis-Rabusseau, C.**, Soubigou-Taconnat, L., Aubourg, S., Lurin, C., Martin-Magniette, M.-L. and Delannoy, E. (2018) Synthetic datasets for the identification of key ingredients for RNA-seq differential analysis. *Briefings in bioinformatics*, 19(1), 65-76.
- [LMM16] Laurent, B., Marteau, C. and **Maugis-Rabusseau, C.** (2016) Non-asymptotic detection of two-component mixtures with unknown means. *Bernoulli*, 22 (1), 242-274.
- [PMM16] Papastamoulis, P., Martin-Magniette, M.-L. and **Maugis-Rabusseau, C.** (2016) On the estimation of mixtures of Poisson regression models with large numbers of components. *Computational Statistics and Data Analysis*, Volume 93, 97-106.

- [RMMC15] Rau, A., **Maugis-Rabusseau, C.**, Martin-Magniette, M.-L. and Celeux, G. (2015) Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 31 (9), 1420-1427.
- [MM13] **Maugis-Rabusseau, C.** and Michel, B. (2013) Adaptive density estimation using finite Gaussian mixtures. *ESAIM : P&S*, 17, 698-724.
- [BMM12] Baudry, J.-P., **Maugis, C.** and Michel, B. (2012) Slope Heuristics : overview and implementation. *Statistics and Computing*, 22(2), 455-470.
- [MCM11] **Maugis, C.**, Celeux, G. and Martin-Magniette, M.-L. (2011) Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, 102, 1374-1387.
- [MM11b] **Maugis, C.** and Michel, B. (2011) Data-driven penalty calibration : A case study for Gaussian mixture model selection. *ESAIM : P&S*, 15, 320-339.
- [MM11a] **Maugis, C.** and Michel, B. (2011) A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM : P&S*, 15, 41-68.
- [MCM09b] **Maugis, C.**, Celeux, G. and Martin-Magniette, M.-L. (2009) Variable selection in model-based clustering : A general variable role modeling. *Computational Statistics and Data Analysis*, 53, 3872-3882.
- [MCM09a] **Maugis, C.**, Celeux, G. and Martin-Magniette, M.-L. (2009) Variable selection for Clustering with Gaussian Mixture Models. *Biometrics*, 65, 701-709.

Articles dans revues nationales à comité de lecture

- [CMMR14] Celeux, G., Martin-Magniette, M.-L., **Maugis-Rabusseau, C.** and Raftery, A. E. (2014) Comparing model selection and regularization approaches to variable selection in model-based clustering. *Journal de la SFdS*, 155 (2), 57-71.
- [MMP12] **Maugis-Rabusseau, C.**, Martin-Magniette, M.-L. and Pelletier, S. (2012) Selvar-ClustMV : Variable selection approach in model-based clustering allowing for missing values. *Journal de la SFdS*, 153 (2), 21-36.
- [M09] **Maugis, C.**, Martin-Magniette, M.-L., Tamby, J.-P., Renou, J.-P., Lecharny, A., Aubourg, S. and Celeux, G. (2009) Sélection de variables pour la classification par mélanges gaussiens pour prédire la fonction des gènes orphelins. *MODULAD*, 40, 69-80.

Chapitres d'ouvrages

- [BM17] Biernacki, C. and **Maugis-Rabusseau, C.** Chapter 9 : "High-dimensional clustering". Model choice and model aggregation, sous la direction de F. BERTRAND, J.-J. DROESBEKE, G. SAPORTA, C. THOMAS-AGNAN. Edition Technip, septembre 2017.
- [MMR17] Martin-Magniette, M.-L., **Maugis-Rabusseau, C.** and Rau, A. Chapter 10 : "Clustering of co-expressed genes". Model choice and Model aggregation, sous la direction de F. BERTRAND, J.-J. DROESBEKE, G. SAPORTA, C. THOMAS-AGNAN. Edition Technip, septembre 2017.

Autres

- [MM12] Meynet, C. and **Maugis-Rabuseau, C.** (2012) A sparse variable selection procedure in model-based clustering. Préprint INRIA, hal-00734316.
- [CMMR11] Celeux, G., Martin-Magniette, M.-L., **Maugis, C.** and A. E. Raftery (2011) Letter to the Editor : "A framework for feature selection in clustering". Journal of the American Statistical Association, 106, 383.
- [M08] **Maugis, C.** Sélection de variables pour la classification non supervisée par mélanges gaussiens. Application à l'étude de données transcriptomes, Thèse de l'Université Paris-Sud 11, Novembre 2008.

Logiciels

- ASTEC-sc** Application shiny <https://cmaugis.github.io/ASTEC-sc/>
ASTEC-sc : A Shiny application To Explore Clusterings of single cell RNA seq data.
This application allows to visualize cells in dimensionality reduction methods and expressed genes, to compare cell clusterings and to determine marker genes.
- maskmeans** Librairie R <https://github.com/andreamrau/maskmeans>
Multi-view agglomeration/splitting K-means clustering algorithm.
- coseq** Librairie Bioconductor <https://bioconductor.org/packages/release/bioc/html/coseq.html>
This package is devoted to the co-expression analysis of sequencing data.
- HTScluster** Librairie R <https://cran.r-project.org/web/packages/HTScluster/index.html>
This package implements two parameterizations of a Poisson mixture model to cluster observations (e.g., genes) in high throughput sequencing data.
- capushe** Librairie R <https://cran.rstudio.com/web/packages/capushe/index.html>
Calibration of penalized criteria for model selection. The available calibration methods are based on the slope heuristics and described in [BMM12].
- SelvarMix** Librairie R <https://cran.r-project.org/web/packages/SelvarMix/index.html>
This package implements a procedure proposed in [CMS19] for variable selection in model-based clustering and discriminant analysis with a regularization approach.
- SelvarClustIndep** Logiciel en C++.
It is dedicated to the so-called SRUW method of variable selection for model-based clustering proposed in [MCM09b]
- SelvarClust** Logiciel en C++
It is dedicated to the so-called SR method of variable selection for model-based clustering proposed in [MCM09a]

Liste de direction et codirection de thèses, post-doctorats et stagiaires

Thèses en cours

- Nicolas Enjalbert-Courrech [10/2021 -]
Post-clustering inference for single-cell RNA sequencing data.
Encadré par C. M.-R. et Pierre Neuvial (CNRS / IMT)
- William Margerit [10/2021 -]
An interdisciplinary approach to the design of effective nanoparticle-based antimicrobials.
Encadré par Juan Cortès (LAAS), C. M.-R. et Nathalie Tarrat (CEMES).

Co-encadrement de post-doctorat

- Antoine Godichon-Baggioni [09/2016-08/2017]
Sujet : Clustering of co-expressed genes based on RNA-seq data.
Encadré par C. M.-R. et Andréa Rau (INRA Jouy-en-Josas)
- Mohammed Sedki [11/2013-08/2014]
Sujet : Sélection de variables en classification supervisée et non supervisée.
Encadré par Gilles Celeux et C. M.-R.
- Andréa Rau [09/2010-09/2011]
Sujet : Clustering RNA sequencing data.
Encadrée par G. Celeux, M.-L. Martin-Magniette et C. M.-R.

Encadrement / co-encadrement de stage

- Stages niveau M2
 - ▶ Mouad Abouqateb [02/2020 - 07/2020]
Sujet : Considering symmetries in clustering methods : Application to the study of biomolecules.
Encadré avec Juan Cortès (LAAS/CNRS) et C. M.-R.
 - ▶ Fanny Mollandin [02/2019 - 07/2019]
Sujet : Analyses statistiques de données single-cell RNA-seq : application à l'étude de cellules stromales mésenchymateuses.
Encadrée par C. M.-R., en collaboration avec Marielle Ousset (Restore).

- Stages niveau M1
 - ▶ Markéta Krupova [06/2021-09/2021]
Sujet : Prédiction multi-label d'agents pathogènes présents dans un échantillon clinique.
Encadrée par Mathilde Plinet (Dendris) et C. M.-R.
 - ▶ Elisa Escanez [06/2021-09/2021]
Sujet : Analyses statistiques de données single-cell RNA-seq.
Encadrée par C. M.-R.
 - ▶ Nicolas Enjalbert Courrech [04/2020-07/2020]
Sujet : Développement d'une application Shiny pour l'exploitation de résultats de classification non supervisée pour des données single-cell RNA-seq.
Encadré par Sébastien Déjean (IMT) et C. M.-R.
 - ▶ Tha Nguyen Khoa Nguyen [06/2019 - 09/2019]
Sujet : Etude de la co-expression des gènes du Medicago Truncatula à partir de données RNA-seq.
Encadrée par Sandrine Laguerre (INRA/TBI) et C. M.-R., en collaboration avec Sandra Bensmihen (LIPM)
 - ▶ Vincent Brault [03/2011 - 07/2011]
Sujet : Programmation en R du logiciel Capushe.
Encadré par J.-P. Baudry, C. M.-R. et B. Michel
- Stage niveau L3
 - ▶ Tuvana Satar [06/2015 - 07/2015]
Sujet : Analyse différentielle de données RNA-seq.
Encadrée par C. M.-R.

Bibliographie

- ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. et LUGOSI, G. (2010). On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092.
- AITCHISON, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177.
- AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.
- ANDERS, S. et HUBER, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1.
- ARIAS-CASTRO, E., CANDES, E. et DURAND, A. (2011). Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39:278–304.
- AUER, P. L. et DOERGE, R. W. (2011). A two-stage Poisson model for testing RNA-seq data. *Statistical Applications in Genetics and Molecular Biology*, 10(26):1–26.
- AZAÏS, J.-M., DE CASTRO, Y. et GAMBOA, F. (2015). Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195.
- AZAÏS, J.-M., GASSIAT, É. et MERCADIER, C. (2009). The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM. Probability and Statistics*, 13:301–327.
- BACHER, R. et KENDZIORSKI, C. (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17(1):63.
- BALAKRISHNAN, S., WAINWRIGHT, M. J. et YU, B. (2017). Statistical guarantees for the EM algorithm : From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.
- BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.
- BECHT, E., MCINNES, L., HEALY, J., DUTERTRE, C.-A., KWOK, I. W. H., NG, L. G., GINHOUX, F. et NEWELL, E. W. (2018). Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37:38 EP –.

- BERK, R., BROWN, L., BUJA, A., ZHANG, K. et ZHAO, L. (2013). Valid post-selection inference. *The Annals of Statistics*, pages 802–837.
- BHASKAR, B. N., TANG, G. et RECHT, B. (2013). Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999.
- BIERNACKI, C., CELEUX, G. et GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.
- BIERNACKI, C., CELEUX, G. et GOVAERT, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575.
- BIRGÉ, L. (2001). An alternative point of view on Lepski’s method. *In State of the art in probability and statistics (Leiden, 1999)*, volume 36 de *IMS Lecture Notes Monogr. Ser.*, pages 113–133. Inst. Math. Statist., Beachwood, OH.
- BIRGÉ, L. et MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73.
- BLANCHARD, G., NEUVIAL, P. et ROQUAIN, E. (2020). Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, 48(3):1281–1303.
- BORDES, L., MOTTELET, S. et VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model. *The Annals of Statistics*, 34(3):1204–1232.
- BOUYEVYRON, C., CELEUX, G., MURPHY, T. B. et RAFTERY, A. E. (2019). *Model-based clustering and classification for data science : with applications in R*, volume 50. Cambridge University Press.
- BUNEA, F., TSYBAKOV, A. B., WEGKAMP, M. H. et BARBU, A. (2010). Spades and mixture models. *The Annals of Statistics*, 38(4):2525–2558.
- BUTLER, A., HOFFMAN, P., SMIBERT, P., PAPALEXI, E. et SATIJA, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411.
- BUTUCEA, C. et INGSTER, Y. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688.
- BUTUCEA, C. et VANDEKERKHOVE, P. (2014). Semiparametric mixtures of symmetric distributions. *Scandinavian Journal of Statistics. Theory and Applications*, 41(1):227–239.
- CAI, L. *et al.* (2004). Clustering analysis of SAGE data using a Poisson approach. *Genome Biology*, 5:R51.
- CAI, T. T., JENG, X. J. et JIN, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 73(5):629–662.
- CAI, T. T., JIN, J. et LOW, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics*, 35(6):2421–2449.

- CAI, T. T. et WU, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory*, 60(4):2217–2232.
- CALIŃSKI, T. et HARABASZ, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.
- CANDÈS, E. J. et FERNANDEZ-GRANDA, C. (2014). Towards a Mathematical Theory of Super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956.
- CANNOODT, R., SAELENS, W. et SAEYS, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *European journal of immunology*, 46(11):2496–2506.
- CAO, Y., LIN, Y., ORMEROD, J. T., YANG, P., YANG, J. Y. et LO, K. K. (2019). scDC : single cell differential composition analysis. *BMC bioinformatics*, 20(19):1–12.
- CELEUX, G. et GOVAERT, G. (1995). Gaussian parsimonious clustering models. *Pattern recognition*, 28(5):781–793.
- CHARNIGO, R. et SUN, J. (2004). Testing homogeneity in a mixture distribution via the L^2 distance between competing models. *Journal of the American Statistical Association*, 99(466):488–498.
- CHEN, H., CHEN, J. et KALBFLEISCH, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):pp. 19–29.
- CHEN, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, pages 221–233.
- CHEN, J. et LI, P. (2009). Hypothesis test for normal mixture models : the EM approach. *The Annals of Statistics*, 37(5A):2523–2542.
- CHENEY, W. et LIGHT, W. (2009). *A course in approximation theory*, volume 101 de *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.
- CHERNOFF, H. et LANDER, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference*, 43(1-2):19–40.
- CHIZAT, L. (2021). Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, pages 1–46.
- DACUNHA-CASTELLE, D. et GASSIAT, E. (1999). Testing the order of a model using locally conic parametrization : population mixtures and stationary ARMA processes. *The Annals of Statistics*, 27(4):1178–1209.
- DE CASTRO, Y. et GAMBOA, F. (2012). Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and Applications*, 395(1):336–354.
- DE CASTRO, Y., GAMBOA, F., HENRION, D. et LASSERRE, J.-B. (2017). Exact solutions to Super Resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630.

- DEMPSTER, A. P., LAIRD, N. M. et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 39(1):1–38.
- DENOYELLE, Q., DUVAL, V., PEYRÉ, G. et SOUBIES, E. (2019). The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*.
- DILLIES, M.-A., RAU, A., AUBERT, J., HENNEQUET-ANTIER, C., JEANMOUGIN, M., SERVANT, N., KEIME, C., MAROT, G., CASTEL, D., ESTELLE, J. *et al.* (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683.
- DING, B., ZHENG, L. et WANG, W. (2017). Assessment of single cell rna-seq normalization methods. *G3 : Genes, Genomes, Genetics*, 7(7):2039–2045.
- DONOHO, D. et JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994.
- DUÒ, A., ROBINSON, M. D. et SONESON, C. (2018). A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7.
- DUVAL, V. et PEYRÉ, G. (2015). Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, pages 1–41.
- DWIVEDI, R., HO, N., KHAMARU, K., WAINWRIGHT, M. J., JORDAN, M. I. et YU, B. (2020). Singularity, misspecification and the convergence rate of EM. *The Annals of Statistics*, 48(6):3161–3182.
- EISEN, M. B. *et al.* (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868.
- FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19:1257–1272.
- FIETZ, S. A., LACHMANN, R., BRANDL, H., KIRCHER, M., SAMUSIK, N., SCHRÖDER, R., LAKSHMANAPERUMAL, N., HENRY, I., VOGT, J., RIEHN, A. *et al.* (2012). Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal. *Proceedings of the National Academy of Sciences*, 109(29):11836–11841.
- FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., MCEL RATH, M. J., PRLIC, M. *et al.* (2015). Mast : a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13.
- FISCHER, A. (2011). On the number of groups in clustering. *Statistics & Probability Letters*, 81(12):1771–1781.
- FRAZEE, A. C., LANGMEAD, B. et LEEK, J. T. (2011). ReCount : a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12.

- FREYTAG, S., TIAN, L., LÖNNSTEDT, I., NG, M. et BAHLO, M. (2018). Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research*, 7.
- FROMONT, M. et LAURENT, B. (2006). Adaptive goodness-of-fit tests in a density model. *The Annals of Statistics*, 34(2):1–45.
- GALLOPIN, M. (2015). *Classification et inférence de réseaux pour les données RNA-seq*. Thèse de doctorat, Université Paris-Saclay.
- GAO, L. L., BIEN, J. et WITTEN, D. (2020). Selective inference for hierarchical clustering. *Preprint arXiv :2012.02936*.
- GAO, X., HU, D., GOGOL, M. et LI, H. (2019). Clustermap : compare multiple single cell rna-seq datasets across different experimental conditions. *Bioinformatics*, 35(17):3038–3045.
- GAREL, B. (2007). Recent asymptotic results in testing for mixtures. *Computational Statistics & Data Analysis*, 51(11):5295–5304.
- GAUTHIER, M., AGNIEL, D., THIÉBAUT, R. et HEJBLUM, B. P. (2021). Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis. *bioRxiv*.
- GHOSAL, S. et van der VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35:697–723.
- GHOSAL, S., VAN DER VAART, A. W. *et al.* (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263.
- GOEMAN, J. J. et SOLARI, A. (2011). Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597.
- GRAVELEY, B. R., BROOKS, A. N., CARLSON, J. W., DUFF, M. O., LANDOLIN, J. M., YANG, L., ARTIERI, C. G., van BAREN, M. J., BOLEY, N., BOOTH, B. W. *et al.* (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature*, 471(7339):473–479.
- HANGELBROEK, T. et RON, A. (2010). Nonlinear approximation using Gaussian kernels. *Journal of Functional Analysis*, 259(1):203–219.
- HEINRICH, P. et KAHN, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870.
- HIE, B., BRYSON, B. et BERGER, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, 37(6):685–691.
- HO, N. et NGUYEN, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307.
- HO, N., NGUYEN, X. *et al.* (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726–2755.
- HUNTER, D. R., WANG, S. et HETTMANSPERGER, T. (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics*, 35(1):224–251.

- INGSTER, Y. (1999). Minimax detection of a signal for l^n -balls. *Mathematical Methods of Statistics*, 7(4):401–428.
- JAAKKOLA, M. K., SEYEDNASROLLAH, F., MEHMOOD, A. et ELO, L. L. (2016). Comparison of methods to detect differentially expressed genes between single-cell populations. *Briefings in Bioinformatics*, 18(5):735–743.
- JAGER, L. et WELLNER, J. A. (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5):2018–2053.
- JIANG, D. *et al.* (2004). Cluster analysis for gene expression data : A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386.
- KARLIS, D. (2003). An EM algorithm for multivariate poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77.
- KAUFMAN, L. et ROUSSEEUW, P. J. (2009). *Finding groups in data : an introduction to cluster analysis*, volume 344. John Wiley & Sons.
- KHARCHENKO, P. V., SILBERSTEIN, L. et SCADDEN, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742.
- KIM, K., ZHANG, S., JIANG, K., CAI, L., LEE, I.-B., FELDMAN, L. J. et HUANG, H. (2007). Measuring similarities between gene expression profiles through new data transformations. *BMC Bioinformatics*, 8(29).
- KISELEV, V. Y., KIRSCHNER, K., SCHAUB, M. T., ANDREWS, T., YIU, A., CHANDRA, T., NATARAJAN, K. N., REIK, W., BARAHONA, M., GREEN, A. R. et HEMBERG, M. (2017). SC3 : consensus clustering of single-cell rna-seq data. *Nature Methods*, 14:483 EP –.
- KISELEV, V. Y., YIU, A. et HEMBERG, M. (2018). scmap : projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362.
- KLAR, B. et MEINTANIS, S. G. (2005). Tests for normal mixtures based on the empirical characteristic function. *Computational Statistics & Data Analysis*, 49(1):227–242.
- KORSUNSKY, I., MILLARD, N., FAN, J., SLOWIKOWSKI, K., ZHANG, F., WEI, K., BAGLAENKO, Y., BRENNER, M., LOH, P.-r. et RAYCHAUDHURI, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296.
- KORTHAUER, K. D., CHU, L.-F., NEWTON, M. A., LI, Y., THOMSON, J., STEWART, R. et KENDZIORSKI, C. (2016). A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome biology*, 17(1):1–15.
- KRUIJER, W., ROUSSEAU, J., VAN DER VAART, A. *et al.* (2010). Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.
- KRZANOWSKI, W. J. et LAI, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34.
- KUCHIBHOTLA, A. K., BROWN, L. D., BUJA, A., CAI, J., GEORGE, E. I. et ZHAO, L. H. (2020). Valid post-selection inference in model-free linear regression. *The Annals of Statistics*, 48(5):2953–2981.

- LABAJ, P. P., LEPARC, G. G., LINGGI, B. E., MARKILLIE, L. M., WILEY, H. S. et KREIL, D. P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, 27(13):i383–i391.
- LÄHNEMANN, D., KÖSTER, J., SZCZUREK, E., MCCARTHY, D. J., HICKS, S. C., ROBINSON, M. D., VALLEJOS, C. A., CAMPBELL, K. R., BEERENWINKEL, N., MAHFOUZ, A. *et al.* (2020). Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35.
- LANGROGNET, F., LEBRET, R., POLI, C., IOVLEFF, S., AUDER, B. et IOVLEFF, S. (2020). *Rmixmod : Classification with Mixture Modelling*. R package version 2.1.5.
- LE CAM, L. (1990). Maximum likelihood : an introduction. *International Statistical Review*, 58(2):153–171.
- LEE, J. D., SUN, D. L., SUN, Y., TAYLOR, J. E. *et al.* (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.
- MAATEN, L. v. d. et HINTON, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.
- MACQUEEN, J. *et al.* (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.
- MASSART, P. (2000). About the constants in Talagrand’s concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884.
- MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *École d’été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics*. Springer.
- MCCARTHY, D. J., CAMPBELL, K. R., LUN, A. T. et WILLS, Q. F. (2017). Scater : pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186.
- MCLACHLAN, G. et PEEL, D. (2000). *Finite Mixture Models*. Wiley series in Probability and Statistics.
- MCLACHLAN, G. J. et KRISHNAN, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.
- MIAO, Z., DENG, K., WANG, X. et ZHANG, X. (2018). Desingle for detecting three types of differential expression in single-cell rna-seq data. *Bioinformatics*, 34(18):3223–3224.
- MIAO, Z. et ZHANG, X. (2016). Differential expression analyses for single-cell rna-seq : old questions on new data. *Quantitative Biology*, 4(4):243–260.
- MORTAZAVI, A., WILLIAMS, B., MCCUE, K., SCHAEFFER, L. et WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628.
- NABAVI, S., SCHMOLZE, D., MAITITUOHETI, M., MALLADI, S. et BECK, A. H. (2016). Emdomics : a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics*, 32(4):533–541.

- NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400.
- OSHLACK, A. et WAKEFIELD, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(14).
- PATRA, R. K. et SEN, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 78(4):869–893.
- PAWLOWSKY-GLAHN, V. et EGOZCUE, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5):384–398.
- PELCKMANS, K., DE BRABANTER, J., SUYKENS, J. A. et DE MOOR, B. (2005). Convex clustering shrinkage. In *PASCAL workshop on statistics and optimization of clustering workshop*.
- POON, C., KERIVEN, N. et PEYRÉ, G. (2019). Support localization and the fisher metric for off-the-grid sparse regularization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1341–1350. PMLR.
- RAIMUNDO, F., VALLOT, C. et VERT, J.-P. (2020). Tuning parameters of dimensionality reduction methods for single-cell rna-seq analysis. *Genome Biology*, 21(1):1–17.
- ROBINSON, M. D. et OSHLACK, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(R25).
- RUDIN, W. (1991). *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, Inc., New York.
- SAELENS, W., CANNODT, R., TODOROV, H. et SAEYS, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554.
- SATIJA, R., HOFFMAN, P. et BUTLER, A. (2019). *SeuratData : Install and Manage Seurat Datasets*. <http://www.satijalab.org/seurat>, <https://github.com/satijalab/seurat-data>.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- SCRUCCA, L., FOP, M., MURPHY, T. B. et RAFTERY, A. E. (2016). mclust 5 : clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317.
- SONESON, C. et ROBINSON, M. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15:255–261.
- SPOKOINY, V. G. (1996). Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6):2477–2498.
- SRIPERUMBUDUR, B. K., FUKUMIZU, K. et LANCKRIET, G. R. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410.

- SUN, S., ZHU, J., MA, Y. et ZHOU, X. (2019a). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell rna-seq analysis. *Genome biology*, 20(1):1–21.
- SUN, Z., CHEN, L., XIN, H., JIANG, Y., HUANG, Q., CILLO, A. R., TABIB, T., KOLLS, J. K., BRUNO, T. C., LAFYATIS, R. *et al.* (2019b). A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. *Nature communications*, 10(1):1649.
- THOMAS, I., FRANKHAUSER, P. et BIERNACKI, C. (2008). The morphology of built-up landscapes in wallonia (belgium) : A classification using fractal indices. *Landscape and urban planning*, 84(2):99–115.
- TIAN, T., WAN, J., SONG, Q. et WEI, Z. (2019). Clustering single-cell rna-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4):191.
- TIBERI, S., CROWELL, H. L., WEBER, L. M., SAMARTSIDIS, P. et ROBINSON, M. D. (2021). distinct : a novel approach to differential distribution analyses. *bioRxiv*, pages 2020–11.
- TIBSHIRANI, R., WALTHER, G. et HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 63(2):411–423.
- TIBSHIRANI, R. J. et TAYLOR, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.
- TSYBAKOV, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York.
- VALLEJOS, C. A., RISSO, D., SCIALDONE, A., DUDOIT, S. et MARIONI, J. C. (2017). Normalizing single-cell rna sequencing data : challenges and opportunities. *Nature methods*, 14(6):565–571.
- Van der VAART, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *The Annals of Statistics*, 24(2):862–878.
- VERZELEN, N. et ARIAS-CASTRO, E. (2017). Detection and feature selection in sparse mixture models. *The Annals of Statistics*, 45(5):1920–1950.
- WANG, T., LI, B., NELSON, C. E. et NABAVI, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC bioinformatics*, 20(1):1–16.
- WANG, T. et NABAVI, S. (2018). Sigemd : A powerful method for differential gene expression analysis in single-cell rna sequencing data. *Methods*, 145:25–32.
- WOLF, F. A., ANGERER, P. et THEIS, F. J. (2018). SCANPY : large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15.
- WOLFOWITZ, J. (1950). Minimax estimation of the mean of a normal distribution with known variance. *Annals of Mathematical Statistics*, 21:218–230.

- WU, C. J. *et al.* (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103.
- XU, J., HSU, D. J. et MALEKI, A. (2016). Global analysis of expectation maximization for mixtures of two gaussians. *In Advances in Neural Information Processing Systems*, pages 2676–2684.
- ZHANG, J. M., FAN, J., FAN, H. C., ROSENFELD, D. et DAVID, N. T. (2018). An interpretable framework for clustering single-cell rna-seq datasets. *BMC bioinformatics*, 19(1):93.
- ZHANG, J. M., KAMATH, G. M. et DAVID, N. T. (2019). Valid post-clustering differential analysis for single-cell rna-seq. *Cell systems*, 9(4):383–392.
- ZHANG, M., LIU, S., MIAO, Z., HAN, F., GOTTARDO, R. et SUN, W. (2022). Ideas : individual level differential expression analysis for single-cell rna-seq data. *Genome Biology*, 23(1):1–17.
- ZHANG, S., LI, X., LIN, Q. et WONG, K.-C. (2020). Review of single-cell RNA-seq Data Clustering for Cell Type Identification and Characterization. *arXiv preprint arXiv :2001.01006*.
- ZHAO, J., JAFFE, A., LI, H., LINDENBAUM, O., SEFIK, E., JACKSON, R., CHENG, X., FLAVELL, R. A. et KLUGER, Y. (2021). Detection of differentially abundant cell subpopulations in scrna-seq data. *Proceedings of the National Academy of Sciences*, 118(22).
- ZIEMANN, M., KASPI, A., LAZARUS, R. *et al.* (2015). Digital Expression Explorer : A user-friendly repository of uniformly processed RNA-seq data. *In ComBio2015*, volume POSTUE-099, Melbourne.