Development of a particle-based model for nonparametric inverse regression

Zacharie Naulet

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'Université Paris-Dauphine

Développement d'un modèle particulaire pour la régression indirecte non paramétrique

École Doctorale de Dauphine — ED 543

Spécialité **Sciences**

COMPOSITION DU JURY :

Eric BARAT
Commissariat à l'Energie Atomique
Co-encadrant

Cristina BUTUCEA
Université Paris-Est Marne-la-Vallée
Présidente du jury

Ismael CASTILLO
Université Pierre et Marie Curie
Rapporteur

Dominique PICARD
Université Paris VII
Membre du jury

Vincent RIVOIRARD
Université Paris-Dauphine
Membre du jury

Judith ROUSSEAU
Université Paris-Dauphine
Directrice de thèse

Surya TOKDAR
Duke University
Rapporteur

Truong TRONG-TUONG
Université de Cergy-Pontoise
Co-Directeur de thèse

Soutenue le 8/11/2016
par Zacharie NAULET

Dirigée par **Judith ROUSSEAU**
Truong TRONG-TUONG

DAUPHINE UNIVERSITÉ PARIS | PSL RESEARCH UNIVERSITY PARIS

Résumé

Cette thèse porte sur les statistiques bayésiennes non paramétriques, avec une attention particulière pour les modèles de mélanges non paramétriques. La thèse est divisée en une introduction générale et trois parties traitant des aspects relativement différents des approches par mélanges (échantillonage, asymptotique, problème inverse). Un résumé des différents chapitres est proposé ci-dessous.

Chapitre 1 (Introduction) Dans un premier temps, nous présentons les concepts clés étudiés dans cette thèse. Particulièrement, nous discutons de l'analyse fréquentiste des méthodes bayésiennes, un des thèmes récurrent au cours des chapitres qui suivent. Dans un second temps, nous nous intéressons à une classe particulière de modèles, à savoir les *modèles de mélange non paramétriques*, thème central de la thèse. Les bases des modèles de mélange sont données, avec un intérêt particulier pour les mélanges par processus Gamma symétriques. La troisème partie de ce chapitre est dédiée à l'*échantillonage*, un concept clé dans le chapitre 2. Finalement, nous introduisons les connaissances nécessaires à la bonne compréhension du chapitre 4, où les outils développés au long de cette thèse sont utilisés pour résoudre le problème inverse mal posé de la *Tomographie Quantique Homodyne*.

Chapitre 2 (co-écrit avec Éric Barat) Dans cet article, nous présentons quelques aspects spécifiques aux mélanges par processus Gamma symétriques pour un usage en régression. Nous proposons un nouvel échantilloneur de Gibbs pour simuler la distribution a posteriori et nous établissons des vitesses de convergence adaptatives pour le modèle de régression moyenne avec bruit Gaussien.

Chapitre 3 (co-écrit avec Judith Rousseau) De nos jours en estimation de densité, les vitesses de contraction de la distribution a posteriori pour les mélanges en translation et en moyenne et variance de Gaussiennes ne sont connus que sous des hypothèses de queues légères; avec de meilleures vitesses obtenues pour les mélanges en translation. Il a été conjecturé par Canale et DeBlasi, mais non prouvé, que la situation devrait être inversée sous des hypothèses de queues lourdes. Cette conjecture est basée sur l'intuition qu'il n'est pas nécessaire d'obtenir un bon ordre d'approximation dans les régions avec peu d'observations (disons les queues), favorisant l'usage des mélanges en moyenne et variance qui peuvent adapter l'ordre d'approximation spatialement. Ici, nous testons l'argument précédent sur le problème de régression gaussienne avec covariables aléatoires. Bien que nous ne puissions pas encore invalider la conjecture, nous trouvons que même avec une hypothèse de queues lourdes, les mélanges en moyenne et variance semblent être moins performants que les mélanges en translation. Cependant, les preuves suggèrent

l'introduction d'un mélange *hybride*, qui semble être toujours plus performant que les deux autres, qu'importe la nature des queues. Finalement, nous montrons que toutes les hypothèses de queues peuvent être supprimées, au prix de rendre la distribution a priori dépendente des covariables.

Chapitre 4 (co-écrit avec Éric Barat) Nous estimons l'état quantique d'un faisceau de lumière à partir des résultats bruités de mesures de tomographie quantique homodyne, réalisées sur des systèmes quantiques identiquement préparés. Nous proposons deux approches Bayésiennes non paramétriques. La première approche est basée sur les modèles de mélanges et est illustrée au travers d'exemples de simulation. La seconde approche est basée sur une expansion sur une base. Nous étudions les performances théoriques de la seconde approche en quantifiant la vitesse de contraction de la distribution a posteriori autour du vrai état quantique dans la métrique L^2.

Summary

This dissertation deals with Bayesian nonparametric statistics, in particular nonparametric mixture models. The manuscript is divided into a general introduction and three parts on rather different aspects of mixtures approaches (sampling, asymptotic, inverse problem). A summary of each chapter is proposed below.

Chapter 1 (Introduction) In a first time, we present the core concepts studied throughout this thesis. Particularly, we discuss frequentist analysis of Bayesian procedures, which constitute one of the governing principle of the next chapters. In a second time, we focus on a particular class of models, namely *nonparametric mixture models*, which is central to the thesis. Basics of mixture models are given, with a stress on Symmetric Gamma Processes Mixtures. The third section of this chapter is dedicated to *posterior sampling*, a key concept in chapter 2. Finally, we introduce the material required for the proper understanding of chapter 4, where the tools developed along the thesis are used to solve the ill-posed inverse problem of *Quantum Homodyne Tomography*.

Chapter 2 (joint work with Éric Barat) In this article, we present some specific aspects of symmetric Gamma process mixtures for use in regression models. We propose a new Gibbs sampler for simulating the posterior and we establish adaptive posterior rates of convergence related to the Gaussian mean regression problem.

Chapter 3 (joint work with Judith Rousseau) Nowadays in density estimation, posterior rates of convergence for location and location-scale mixtures of Gaussians are only known under light-tail assumptions; with better rates achieved by location mixtures. It has been conjectured by Canale and DeBlasi, but not proved, that the situation should be reversed under heavy tails assumptions. The conjecture is based on the feeling that there is no need to achieve a good order of approximation in regions with few data (say, in the tails), favoring location-scale mixtures which allow for spatially varying order of approximation. Here we test the previous argument on the Gaussian errors mean regression model with random design, for which the light tail assumption is not required for proofs. Although we cannot invalidate the conjecture due to the lack of lower bound, we find that even with heavy tails assumptions, location-scale mixtures apparently perform in general worst than location mixtures. However, the proofs suggest to introduce *hybrid* location-scale mixtures that are find to outperform both location and location-scale mixtures, whatever the nature of the tails. Finally, we show that all tails assumptions can be released at the price of making the prior distribution covariate dependent.

Chapter 4 (joint work with Éric Barat) We estimate the quantum state of a light beam from results of quantum homodyne tomography noisy measurements performed on identically prepared quantum systems. We propose two Bayesian nonparametric approaches. The first approach is based on mixture models and is illustrated through simulation examples. The second approach is based on random basis expansions. We study the theoretical performance of the second approach by quantifying the rate of contraction of the posterior distribution around the true quantum state in the L^2 metric.

Remerciements

Je tiens à remercier Judith Rousseau pour m'avoir encadré et transmis une partie de ses connaissances pendant ces trois années et demi de thèse. Merci pour ta patience, en particulier au début de ma thèse, et merci pour toutes les fois où ton intuition géniale m'a évité de perdre des semaines entières de travail. Avec le recul, je mesure la chance que j'ai eu de pouvoir travailler à tes côtés.

J'exprime également toute ma gratitude à Truong Trong-Tuong pour avoir accepté de co-encadrer ce travail, et pour sa bienveillance et son soutient tout au long de cette thèse.

L'essentiel du travail présenté dans ce manuscrit a été réalisé au laboratoire LM2S du Commissariat à l'Energie Atomique (CEA), où j'ai passé trois ans. Un immense merci à Éric Barat pour avoir initié ce sujet de thèse et pour m'avoir choisi comme doctorant. Je te suis infiniment reconnaissant pour la liberté dont j'ai joui durant mes années au CEA. Merci pour avoir eu confiance en moi depuis le début et de m'avoir permis de travailler sur un sujet aussi extraordinaire. Il est clair que sans toi, cette thèse ne serait pas ce qu'elle est aujourd'hui.

Je remercie également toute l'équipe du LM2S, ainsi que tous ceux qui sont passés au laboratoire pendant la période de ma thèse. Merci pour tous les échanges constructifs, les moments partagés autour d'un café, un repas, un billard...

J'adresse mes remerciements à Ismael Castillo et Surya Tokdar pour avoir accepté de rapporter cette thèse. Je suis très honoré de l'intérêt et de l'enthousiasme que vous avez manifesté pour ce travail. Je vous suis reconnaissant pour le temps passé, ainsi que pour vos remarques et suggestions. Merci également à Cristina Butucea, Dominique Picard et Vincent Rivoirard d'avoir accepté de compléter ce jury.

Durant ces trois années, j'ai fait partie de l'ANR BNPSI, une expérience riche qui m'a permis de rencontrer des personnes formidables avec qui nous avons pu échanger sur nos travaux au cours de réunions mémorables. Un grand merci à eux!

En juin 2015, j'ai eu la chance de visiter le département de statistiques de l'Université de Duke. Je tenais à remercier Robert Wolpert et Merlyse Clyde pour leur invitation. Je remercie également tout le département pour leur accueil remarquable.

Une partie de ce manuscrit a été rédigée lors de mon séjour à l'Université d'Oxford en mai-juin dernier. J'en profite pour remercier tous les membres du laboratoire, en particulier François Caron qui est à l'origine de mon séjour là-bas.

Je voudrai également remercier toutes les personnes ayant soutenu et apprécié mon travail, et plus particulièrement ma famille qui m'a toujours soutenu, encouragé et aidé.

Mes derniers remerciements vont à Jessica pour son soutien inconditionnel. Ton support m'a été et me sera toujours très cher. Cette thèse et moi te devons beaucoup.

Table of contents

List of symbols

Algebra and others

$a := b$, $b =: a$	a is equal to b by definition
$\lesssim$, $\gtrsim$	Lesser than up to generic constant, greater than up to generic constant
$a \wedge b$, $a \vee b$	The minimum, respectively the maximum, of a and b
$(x)_+$	0 if $x \leq 0$, x otherwise
$X \times Y$	Cartesian product of sets X and Y

Measure theory

$(\Omega, \mathcal{A})$	Measurable space
$(\Omega, \mathcal{A}, \mathbb{P})$	Measured space, probability space if $\mathbb{P}$ is a probability measure
$\mathcal{B}_X$, $\mathcal{B}(X)$	Borel σ-algebra of the topological space X
$\mathbb{E}$	Expectation operator
Pf	Abbreviate notation for $\int f \, dP$
$\Sigma_1 \otimes \Sigma_2$	Product σ-algebra
$\lvert\mu\rvert$	Total variation measure of the signed measure μ
$X \sim \mu$	The random variable X has distribution μ

Function spaces

$\mathcal{S}(X)$	The Schwartz space over X
$L^p(X, \mathcal{X}, \mu)$	The L^p space over $X, \mathcal{X}, \mu$
$\mathtt{C}^\beta(X)$	The β-Hölder space over X
$\lVert \cdot \rVert_p$	The L^p norm

Whenever convenient, we use the abbreviations $\mathcal{S}$, L^p, $L^p(X)$, and $\mathtt{C}^\beta$.

Linear algebra

det	Determinant operator
Tr	Trace operator
$\langle \cdot, \cdot \rangle$	Inner product
$\lVert \cdot \rVert_{p,n}$	For $x \in \mathbb{R}^n$, we define $\lVert x \rVert_{p,n}^p := \sum_{i=1}^n \lvert x_i \rvert^p$.

Asymptotic comparison

$o(n)$	Little-O notation
$O(n)$	Big-O notation
$o_p(n)$	Little-O in probability
$O_p(n)$	Big-O in probability

Special functions

E_1	Exponential integral E_1
Γ	*Gamma* function

CHAPTER 1

Introduction

Résumé

Dans un premier temps, nous présentons les concepts clés étudiés dans cette thèse. Particulièrement, nous discutons de l'analyse fréquentiste des méthodes bayésiennes, un des thèmes récurrent au cours des chapitres qui suivent. Dans un second temps, nous nous intéressons à une classe particulière de modèles, à savoir les *modèles de mélange non paramétriques*, thème central de la thèse. Les bases des modèles de mélange sont données, avec un intérêt particulier pour les mélanges par processus Gamma symétriques. La troisème partie de ce chapitre est dédiée à l'*échantillonage*, un concept clé dans le chapitre 2. Finalement, nous introduisons les connaissances nécessaires à la bonne compréhension du chapitre 4, où les outils développés au long de cette thèse sont utilisés pour résoudre le problème inverse mal posé de la *Tomographie Quantique Homodyne*.

Abstract

In a first time, we present the core concepts studied throughout this thesis. Particularly, we discuss frequentist analysis of Bayesian procedures, which constitute one of the governing principle of the next chapters. In a second time, we focus on a particular class of models, namely *nonparametric mixture models*, which is central to the thesis. Basics of mixture models are given, with a stress on Symmetric Gamma Processes Mixtures. The third section of this chapter is dedicated to *posterior sampling*, a key concept in chapter 2. Finally, we introduce the material required for the proper understanding of chapter 4, where the tools developed along the thesis are used to solve the ill-posed inverse problem of *Quantum Homodyne Tomography*.

1.1. Bayesian statistics

In the Bayesian paradigm, both the parameter and observations are treated as random variables; that is the parameter space Θ is seen as a measurable space (Θ, Σ_Θ). Then we make the following definition of a *Bayesian statistical model*.

Definition 1.1 – *Let Π_n be a probability distribution over $(\mathcal{X}^n \times \Theta, \mathcal{B}_n \otimes \Sigma_\Theta)$. Then we call the probability space $(\mathcal{X}^n \times \Theta, \mathcal{B}_n \otimes \Sigma_\Theta, \Pi_n)$ a Bayesian statistical model. The model is called parametric when Θ is a finite dimensional real vector space, and nonparametric when Θ has infinite dimension.*

The existence of regular conditional probability measures for Π_n is not guaranteed without further assumptions of $\mathcal{X}^n \times \Theta$. In the sequel, we shall always assume that both $\mathcal{X}^n$ and Θ are complete separable metric spaces, which ensure the existence of regular conditional distributions for Π_n (Dudley, 2002, theorem 10.2.2).

Given a model $(\mathcal{X}^n \times \Theta, \mathcal{B}_n \otimes \Sigma_\theta, \Pi_n)$, we define $P_\theta^{(n)}(\cdot) := \Pi_n(\cdot \mid \theta)$ as the conditional distribution of the observations $X^n \in \mathcal{X}^n$ given the parameter $\theta \in \Theta$. The marginal distribution $\Pi_{n,\theta}(\cdot)$ on Θ of Π_n is called the *prior distribution*, and reflects the prior belief in θ in absence of any observation. The marginal distribution P^Π of $\mathcal{X}^n$ is called the *prior predictive distribution*.

The goal of the Bayesian inference is to characterize the *posterior distribution*, defined as $\Pi_{n,\theta}(\cdot \mid X^n)$, the conditional distribution of θ given $X^n \in \mathcal{X}^n$. A common framework is when the family $\{P_\theta^{(n)} : \theta \in \Theta\}$ is dominated by a measure λ_n on $(\mathcal{X}^n, \mathcal{B}_n)$. Then a regular version of the posterior distribution is given by the Bayes rule.

Theorem 1.1 — Bayes' rule. – *Assume that $\{P_\theta^{(n)} : \theta \in \Theta\}$ is dominated by a measure λ_n. Then a regular version of the posterior distribution $\Pi_{n,\theta}(\cdot \mid X^n)$ is given by the following expression. For all $A \in \Sigma_\theta$*

$$\Pi_{n,\theta}(\theta \in A \mid X^n) = \int_A \frac{dP_\theta^{(n)}}{d\lambda_n}(X^n)\, d\Pi_{n,\theta}(\theta) \Big/ \int_\Theta \frac{dP_\theta^{(n)}}{d\lambda_n}(X^n)\, d\Pi_{n,\theta}(\theta).$$

The transition from prior distribution to posterior distribution represents the way in which prior beliefs are turned onto posterior beliefs based on the data. Theorem 1.1 and definition 1.1 emphasize the fact that in the Bayesian paradigm, all values of θ in the support of $\Pi_{n,\theta}$ are possible with more or less likeliness, even in absence of data.

Finally, let mention that although the prior distribution $\Pi_{n,\theta}$ is allowed to depend on n, this will not be the case in this dissertation, except at the end of chapter 3. When the prior distributions considered have some dependency on n, it will be mentioned explicitly; otherwise it is assumed that priors have no dependency on n.

For further discussion on foundations of the Bayesian approach, we refer to Robert (2007); Bernardo and Smith (2001); Berger (2013); Van der Vaart (2000); Le Cam and Yang (2000); Le Cam (1986).

1.1.1. Frequentist validation of Bayes procedures

One of the most frequent criticisms about the Bayesian approach concerns the arbitrariness in the choice of the prior distribution. Of course, when there is concrete prior knowledge about the parameter, the Bayesian paradigm provide a natural way to incorporate it into the model, and this must be done. However, in many cases, prior knowledge is vague, or incomplete, which makes very difficult to specify a well justified prior distribution. Different people, with different opinions, may specify different prior distributions and arrive at different conclusions.

Frequentist study of the posterior distribution is a way to address the above concern. Indeed, the goal is to find conditions on the model under which, with a sufficient amount of data, the sequence of posterior distributions concentrates (in a sense to be clarified below) around the *true* θ_0 in the frequentist sense.

We should note that the argument in the previous paragraph is not the only reason why a Bayesian should take care about frequentist properties of the posterior distribution. See Diaconis and Freedman (1986) for a discussion.

Now we clarify the frequentist properties that we consider here, from the weaker to the stronger notion.

Definition 1.2 — Posterior consistency. – *Let Θ be endowed with the topology $\mathcal{T}_\Theta$. The sequence of posterior distributions $\{\Pi_{n,\theta}(\cdot \mid X^n) : n \geq 1\}$ is said to be consistent for $\mathcal{T}_\Theta$ at θ_0 if for any neighborhood U of θ_0 we have $P_{\theta_0}^{(n)}\Pi_{n,\theta}(U \mid X^n) \to 1$.*

Definition 1.3 — Rates of convergence. – *Let Θ be endowed with the semimetric d_n, and let $(\epsilon_n)_{n\geq 0}$ be a sequence of positive real numbers such that $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$. The sequence of posterior distributions $\{\Pi_{n,\theta}(\cdot \mid X^n) : n \geq 1\}$ is said to contracts around θ_0 at rate ϵ_n if there is a constant $M > 0$ such that $P_{\theta_0}^{(n)}\Pi_{n,\theta}(\theta : d_n(\theta, \theta_0) \geq M\epsilon_n \mid X^n) \to 0$.*

Doob (1949) proposes a general theorem on posterior consistency, which holds under very weak conditions. For a modern exposition of Doob's theorem, we refer to Le Cam and Yang (2000, section 8.2).

Theorem 1.2 — Doob's theorem. – *Let $\mathcal{X}$ be an euclidean space endowed with Borel σ-algebra $\mathcal{B}$. Assume that the observations $X_1, \ldots, X_n \in \mathcal{X}$ are identically and independently distributed with distribution P_{θ_0} on $(\mathcal{X}, \mathcal{B})$, and the map $\theta \mapsto P_\theta$ is injective. Also assume that (Θ, Σ_Θ) is a Borel set in a complete separable metric space with canonical topology $\mathcal{T}_\Theta$. Then the sequence of posterior distributions is consistent for $\mathcal{T}_\Theta$ at Π_θ-almost all θ_0.*

When the parameter space Θ is finite dimensional, there are finer results about the behavior of $\{\Pi_{n,\theta}(\cdot \mid X^n) : n \geq 1\}$, such as Bernstein-von Mises theorem (Le Cam and Yang, 2000) which provide sufficient conditions on the model to ensure the asymptotic normality of the sequence of posterior distributions.

In nonparametric Bayes, even consistency can fails, despite theorem 1.2. In fact, Doob's theorem guarantees the posterior consistency up to null sets of the prior distribution. In infinite dimensional models, it is easy to find examples where null sets of the prior distribution are large in some topological sense (see sections 2.4.2 and 2.4.3 in Hjort et al. (2010), and also Freedman (1963)). If the posterior distribution has an explicit expression, it may be possible to prove consistency or rates of convergence by simple Chebychev-type inequalities. However, examples with explicit expression of the posterior distribution are very special and are not to be expected in all applications. The Schwartz-Ghosal-van-der-Vaart theory proposes a general approach to consistency and rates of convergence.

1.1.2. The Schwartz-Ghosal-van-der-Vaart theory

In nonparametric Bayes, sufficient conditions on the model and prior distribution for posterior consistency have been stated by Schwartz (1965). Her approach has been extended in the seminal papers of Ghosal et al. (2000); Ghosal and Van Der Vaart (2001, 2007a,b) and this to derive posterior concentration rates. Here we present the main ideas behind their approach to posterior rates of convergence. It should be noticed, however, that the conditions on the model and on the prior are less stringent if one seeks only a

posterior consistency result (Ghosh and Ramamoorthi, 2003).

We consider the case of observations coming from a Bayesian model $(\mathcal{X}^n \times \Theta, \mathcal{B}_n \otimes \Sigma_\Theta, \Pi_n)$, where $\Pi_n \equiv P_\theta^{(n)} \times \Pi$; the observations need not to be identically distributed nor independent. We assume that for all $n \geq 1$, the set $\{P_\theta^{(n)} : \theta \in \Theta\}$ is dominated by a measure λ_n. We endow Θ with the semimetric d_n. For two probability measures P_1, P_2, $K(P_1, P_2) := \int \log(dP_1/dP_2)\, dP_1$ stands for the *Kullback-Leibler* divergence. Furthermore, we define the additional discrepancy measure $V_{2,0}(P_1, P_2) := \int |\log(dP_1/dP_2) - K(P_1, P_2)|^2\, dP_1$, and

$$B_n(\theta_0, \epsilon) := \left\{\theta \in \Theta : K(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \leq n\epsilon^2, \quad V_{2,0}(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \leq n\epsilon^2\right\}.$$

We recall that a *test function* $\phi_n : \mathcal{X}^n \to [0,1]$ for testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \in \Theta_1 \subset \Theta$ is a measurable mapping such that we refuse H_0 if $\phi_n(x) = 1$ and we accept H_0 if $\phi_n(x) = 0$; otherwise if $\phi_n(x) \in (0,1)$, we reject H_0 with probability $\phi_n(x)$. Then we state the following proposition.

Proposition 1.1 – *Assume that there are universal constants $C_1, C_2 > 0$, and a sequence $(\epsilon_n)_{n\geq 0}$ with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, such that:*

- *there is a sequence of tests $(\phi_n)_{n\geq 1}$ for testing $H_0 : \theta = \theta_0$ against $H_1 : d_n(\theta, \theta_0) \geq \epsilon_n$, with $P_{\theta_0}^{(n)}\phi_n \to 0$ and $\sup_{\theta\,:\,d_n(\theta,\theta_0)\geq\epsilon_n} P_\theta^{(n)}(1-\phi_n) \leq \exp(-3C_1 n\epsilon_n^2)$, and*
- *$\Pi(B_n(\theta_0, \epsilon_n)) \geq C_2 \exp(-C_1 n\epsilon_n^2)$.*

Then $P_{\theta_0}^{(n)}\Pi(d_n(\theta, \theta_0) \geq \epsilon_n \mid X^n) \to 0$.

The proof of proposition 1.1 indeed rely on a very simple idea, which we sketch now. We may rewrite the posterior as,

$$\begin{aligned}&\Pi(d_n(\theta,\theta_0) \geq \epsilon_n \mid X^n)\\ &\quad = \phi_n(X^n)\Pi(d_n(\theta,\theta_0) \geq \epsilon_n \mid X^n) + (1-\phi_n(X^n))\Pi(d_n(\theta,\theta_0) \geq \epsilon_n \mid X^n). \quad (1.1)\end{aligned}$$

Clearly, $\Pi(d_n(\theta,\theta_0) \geq \epsilon_n \mid X^n) \leq 1$, thus when taking the expectation with respect to $P_{\theta_0}^{(n)}$, the first term of the rhs of equation (1.1) is bounded above by $P_{\theta_0}^{(n)}\phi_n \to 0$. Regarding the second term of the rhs of equation (1.1), we can use theorem 1.1 to rewrite

$$\Pi(d_n(\theta,\theta_0) \geq \epsilon_n \mid X^n) = \int_{\{\theta : d(\theta,\theta_0)\geq\epsilon_n\}} \frac{dP_\theta^{(n)}}{dP_{\theta_0}^{(n)}}(X^n)\, d\Pi(\theta) \Big/ \int_\Theta \frac{dP_\theta^{(n)}}{dP_{\theta_0}^{(n)}}(X^n)\, d\Pi(\theta).$$

Consider the event A_n on which the denominator in the last expression is greater than $\exp(-2C_1 n\epsilon_n^2)$. Then the expectation of the second term in equation (1.1) is bounded by

$$\begin{aligned}&P_{\theta_0}^{(n)}\left[(1-\phi_n(X^n))\Pi(d_n(\theta,\theta_0) \geq \epsilon_n \mid X^n)\, \mathbb{1}_{A_n}\right] + P_{\theta_0}^{(n)}(A_n^c)\\ &\qquad \leq e^{2C_1 n\epsilon_n^2} P_{\theta_0}^{(n)} \int_{\{\theta : d(\theta,\theta_0)\geq\epsilon_n\}} (1-\phi_n(X^n)) \frac{dP_\theta^{(n)}}{dP_{\theta_0}^{(n)}}(X^n)\, d\Pi(\theta) + P_{\theta_0}^{(n)}(A_n^c).\end{aligned}$$

By Fubini's theorem, the first term in the rhs of the last equation is bounded above by $\sup_{\theta\in\{\theta : d(\theta,\theta_0)\geq\epsilon_n\}} P_\theta^{(n)}(1-\phi_n)\exp(2C_1 n\epsilon_n^2)$. To finish the proof of proposition 1.1, it

remains to show that $P_{\theta_0}(A_n^c) \to 0$, which is the case if $\Pi(B_n(\theta_0, \epsilon_n)) \geq C_2 \exp(-C_1 n\epsilon_n^2)$ as proved in Ghosal and Van Der Vaart (2007a, lemma 10).

A common approach to construct the tests ϕ_n is to find test functions $\varphi_n^{\theta_1}$ for testing balls, *i.e.* $H_0 : \theta = \theta_0$ against $H_1 : d_n(\theta, \theta_1) \leq r\epsilon_n$ for some $0 < r < 1$ and $d_n(\theta_1, \theta_0) \geq \epsilon_n$, that satisfy

$$P_{\theta_0}^{(n)} \varphi_n^{\theta_1} \leq e^{-3C_1 n\epsilon_n^2}, \qquad \sup_{\theta\,:\, d_n(\theta,\theta_1) \leq r\epsilon_n} P_\theta^{(n)}(1 - \varphi_n^{\theta_1}) \leq e^{-3C_1 n\epsilon_n^2}. \tag{1.2}$$

For instance, in density estimation existence of such tests for the Hellinger distance and L^1 distance follows from Le Cam (1986, lemma 4, p. 478), whereas in Gaussian regression (see also chapters 2 and 3) this follows from Birgé (2006, proposition 4) for the $\|\cdot\|_{2,n}$ distance. If there exists a finite covering of Θ with $N(r\epsilon_n, d_n, \Theta) \lesssim \exp(2C_1 n\epsilon_n^2)$ balls of radius $r\epsilon_n$ and centers $\{\theta_i \,:\, 1 \leq i \leq N(r\epsilon_n, d_n, \Theta)\}$, then ϕ_n can be constructed as

$$\phi_n = \max\Big\{\varphi_n^{\theta_i} \,:\, d(\theta_0, \theta_i) \geq \epsilon_n \text{ and } 1 \leq i \leq N(r\epsilon_n, d_n, \Theta)\Big\}.$$

If Θ cannot be covered with $N(r\epsilon_n, d_n, \Theta) \lesssim \exp(2C_1 n\epsilon_n^2)$ balls of radius $r\epsilon_n$, it is enough to find $\Theta_n \subset \Theta$ such that $N(r\epsilon_n, d_n, \Theta_n) \lesssim \exp(2C_1 n\epsilon_n^2)$ and $P_{\theta_0}^{(n)} \Pi(\Theta \backslash \Theta_n \mid X^n) \to 0$. Then we have the following theorem.

Theorem 1.3 – *Let $C_1, C_2 > 0$ as previously, $0 < r < 1$, and $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$. Suppose that $\Theta_n \subset \Theta$ is such that $\Pi(\Theta \backslash \Theta_n) \leq \exp(-3C_1 n\epsilon_n^2)$ and $\log N(r\epsilon_n, d_n, \Theta_n) \leq 2C_1 n\epsilon_n^2$. Furthermore assume that there is a sequence of tests $(\varphi_n^{\theta_1})_{n\geq 1}$ satisfying equation (1.2), and assume that $\Pi(B_n(\theta_0, \epsilon_n)) \geq C_2 \exp(-C_1 n\epsilon_n^2)$. Then $P_{\theta_0}^{(n)} \Pi(d_n(\theta, \theta_0) \geq \epsilon_n \mid X^n) \to 0$.*

The condition $\Pi(\Theta \backslash \Theta_n) \leq \exp(-3C_1 n\epsilon_n^2)$ is sufficient to ensure that $P_{\theta_0}^{(n)} \Pi(\Theta \backslash \Theta_n \mid X^n) \to 0$; this follows from a similar argument that we used in the sketch of the proof of proposition 1.1.

1.1.3. Minimax theory

A classical way to assess the asymptotic performances of nonparametric frequentists estimators is to compare the risk they achieve over a class $\Theta_0 \subset \Theta$ to the optimal risk, the so-called *minimax risk*. Following Tsybakov (2009), given a semimetric d_n, we let the performance of an estimator $\widehat{\theta}_n$ of θ_0 be measured by the maximum risk of this estimator on Θ_0 : $\sup_{\theta\in\Theta_0} P_\theta^{(n)}[d_n(\widehat{\theta}_n, \theta)^2]$. Then we have the following definition of the *minimax risk* and the *optimal rates of convergence* over Θ_0, respectively.

Definition 1.4 — Minimax risk. – *The minimax risk over $\Theta_0 \subset \Theta$ associated with a model $\{P_\theta^{(n)} \,:\, \theta \in \Theta\}$ and with a semimetric d_n is $\mathcal{R}_n^* := \inf_{\widehat{\theta}_n} \sup_{\theta\in\Theta_0} P_\theta^{(n)}[d_n(\widehat{\theta}_n, \theta)^2]$, where the infinimum is taken over all the estimators.*

Definition 1.5 — Optimal rate of convergence. – *A positive sequence $(\epsilon_n)_{n\geq 1}$ is called an optimal rate of convergence of estimators on (Θ_0, d_n) if there are constants $C_1, C_2 > 0$ such that $\limsup_n \epsilon_n^{-2} \mathcal{R}_n^* \leq C_1$ and $\liminf_n \epsilon_n^{-2} \mathcal{R}_n^* \geq C_2$. An estimator $\widehat{\theta}_n$ satisfying $\sup_{\theta\in\Theta_0} P_\theta^{(n)}[d_n(\widehat{\theta}_n, \theta)^2] \leq C_3 \epsilon_n^2$, where $(\epsilon_n)_{n\geq 1}$ is the optimal rate of convergence on (Θ_0, d_n) is called a rate optimal estimator on (Θ_0, d_n).*

Regarding Bayesian procedures, it is proved in Ghosal et al. (2000, theorem 2.5) that if a sequence of posterior distributions $\{\Pi(\cdot \mid X^n) : n \geq 1\}$ converges at rate ϵ_n at all $\theta \in \Theta_0$ for the semimetric d_n — under weak conditions on d_n — then there exists at least one Bayesian point estimator based on $\Pi(\cdot \mid X^n)$ that achieve the maximum risk ϵ_n (up to constants). In other words, the rate of contraction of $\{\Pi(\cdot \mid X^n) : n \geq 1\}$ over (Θ_0, d_n) cannot be faster than the frequentist optimal rate of convergence for (Θ_0, d_n), otherwise it would imply the existence of a point estimator performing better than the best possible estimator.

Often, the parameter is assumed to belong to a class $\Theta_0 \equiv \mathcal{A}(\beta, L)$ where β relates to a notion of smoothness or sparsity and L controls the "size" of the class in a certain sense. As we in general do not know in advance the value of β for θ_0, we want procedures that achieve optimal rates simultaneously for all value of β; such an estimator is called *adaptive* over $\mathcal{A}(\beta, L)$.

1.2. Nonparametric mixtures

So far, we have introduced the Bayesian paradigm from a general angle. The object of study of this thesis is essentially concerned with the case when the parameter of interest to be estimated is a function from $\mathbb{R}^d$ to $\mathbb{R}$ (or $\mathbb{C}$). Popular prior distributions over function spaces are *Gaussian processes* (Rasmussen, 2004), and in the special situation when θ is a probability density function, *Dirichlet Process Mixtures* (Antoniak, 1974; Escobar and West, 1995).

The use of mixtures has received much less attention outside the density estimation framework, apart from Abramovich et al. (2000); Wolpert et al. (2011); De Jonge and Van Zanten (2010). Before clarifying the meaning of a *mixture*, we need the following definition of a *random measure*. We will consider random probability measures and random (positive) measures as particular cases of the more general *random signed measures*.

Definition 1.6 — Random signed measure. – *Let $(\Omega, \mathcal{E}, \mathbb{P})$ be a probability space and $(\mathcal{G}, \Sigma)$ be a measurable space. We call a mapping $Q : \Omega \times \Sigma \to \mathbb{R} \cup \{\pm\infty\}$ a random signed measure if $\omega \mapsto Q(\omega, A)$ is a random variable for each $A \in \Sigma$ and if $A \mapsto Q(\omega, A)$ is a signed measure for each $\omega \in \Omega$.*

From now, without explicit mention, measures refer to signed measures. Random measures may be either viewed as stochastic processes indexed by a σ-algebra $\{Q(\cdot, A) : A \in \Sigma\}$, or as a probability distribution over the set of measures. The latter interpretation is very popular in Bayesian nonparametrics, especially for random probability measures (Ferguson, 1973; Lijoi and Prünster, 2010). From random measures, it is rather easy to construct random functions, which we call *mixtures*.

Definition 1.7 – *Given a jointly measurable kernel $K : \mathcal{G} \times \mathbb{R}^d \to \mathbb{R}$ and a random measure $Q : \Omega \times \Sigma \to \mathbb{R} \cup \{\pm\infty\}$, we define the mixture of K by Q as the random function $\theta : \Omega \times \mathbb{R}^d \to \mathbb{R}$ such that $\theta(\omega, y) = \int_{\mathcal{G}} K(x; y)\, Q(\omega, dx)$ for all $(\omega, y) \in \Omega \times \mathbb{R}^d$.*

It is unclear for which kernels K definition 1.7 is well-defined. Assume that the *mean measure* $\mu(A) = \int_\Omega Q(\omega, A)\, \mathbb{P}(d\omega)$ exists and is bounded for all boundet set $A \in \Sigma$. Consider simple functions $f = \sum_{i=1}^n c_i \mathbb{1}_{A_i}$ with $n \in \mathbb{N}$, $c_i \in \mathbb{R}$ and $|\mu|(A_i) < +\infty$. Then $\omega \mapsto \int f(x)\, Q(\omega, dx) = \sum_{i=1}^n c_i\, Q(\omega, A_i) =: X_f(\omega)$ is measurable, with $\mathbb{E}X_f = \int_{\mathcal{G}} f\, d\mu$

and $\mathbb{E}|X_f| \leq \int_{\mathcal{G}} |f|\, d|\mu|$. Now for any $f \in L^1(|\mu|)$ we can construct a sequence (f_n) of simple functions such that $f_n \to f$ pointwise and in $L^1(|\mu|)$ with $|f_n| \leq |f|$. It follows for any $n, m \in \mathbb{N}$ that $\mathbb{E}|X_{f_n} - X_{f_m}| \leq \int_{\mathcal{G}} |f_n - f_m|\, d|\mu|$. Hence (X_{f_n}) is Cauchy in $L^1(\mathbb{P})$, by completeness the limit X_f exists and is measurable. Moreover, X_f does not depend on the approximating sequence (f_n), *i.e.* $\omega \mapsto \int f(x)\, Q(\omega, dx)$ is measurable and belongs to $L^1(\mathbb{P})$. Therefore, definition 1.7 is at least satisfactory if the mean measure μ of Q exists and $K(\cdot; y) \in L^1(|\mu|)$ for all $y \in \mathbb{R}^d$; otherwise we should pay attention to the measurability of $\omega \mapsto \theta(\omega, y)$.

1.2.1. Dirichlet Process Mixtures

Since Ferguson (1973), Dirichlet Processes are very popular in Bayesian nonparametrics, with a use across a wide variety of a applications, such as model validation, clustering or density estimation. The Dirichlet Process is a probability distribution over the set of probability measures, defined as follows.

Definition 1.8 — Dirichlet Process. – *Let α be a non null finite positive measure on $(\mathcal{G}, \Sigma)$. We say that a random measure $P : \Omega \times \Sigma \to [0, 1]$ follows a Dirichlet Process with base measure α, abbreviated* $\mathrm{DP}(\alpha)$*, if for all $k \in \mathbb{N}^*$ and all measurable partition $(B_1, \ldots, B_k)$ of $\mathcal{G}$*

$$(P(B_1), \ldots, P(B_k)) \sim \mathrm{Dirichlet}(\alpha(B_1), \ldots, \alpha(B_k)).$$

Beside this formal definition, Dirichlet Processes can be characterized in a series of different ways:

- Let $X_1 \sim \alpha(\cdot)/\alpha(\mathcal{G})$; and for any $k = 1, \ldots, n-1$ let $X_{k+1} \sim \alpha_k(\cdot)/\alpha_k(\mathcal{G})$, with $\alpha_k(\cdot) = \alpha(\cdot) + \sum_{i=1}^{k} \delta_{X_i}$. Then $\{X_i : i \geq 1\}$ is a *Pólya urn sequence* with base distribution α. The sequence $\{X_i : i \geq 1\}$ is *exchangeable*. Therefore deFinetti's theorem (Kallenberg, 2006, theorem 9.6) implies that there is a random probability distribution P such that conditionally on P, the random variable $X_1, X_2, \ldots$ are independent. It can be shown that P follows a $\mathrm{DP}(\alpha)$ (Blackwell and MacQueen, 1973).
- The *stick-breaking* construction, due to Sethuraman (1994) is the following. Let $\overline{\alpha} := \alpha(\mathcal{G})$ and $V_1, V_2, \ldots$ be identically and independently distributed (iid) with distribution $\mathrm{Beta}(1, \overline{\alpha})$. Set $W_1 = V_1$, and for all $k \geq 2$ set $W_k = V_k \prod_{j=1}^{k-1}(1 - V_j)$. Finally let $X_1, X_2, \ldots$ be iid with distribution $\alpha(\cdot)/\overline{\alpha}$, independently of the V_k's. Then $P = \sum_{k=1}^{\infty} W_k \delta_{X_k}$ converges almost-surely toward a $\mathrm{DP}(\alpha)$ distributed random measure. Note that this construction emphasize the almost-sure discreteness character of the DP.
- If $Q : \Omega \times \Sigma \to \mathbb{R}^+$ is a (nonsymmetric) Gamma random measure with base measure α (see below), then $P : \Omega \times \Sigma \to [0, 1]$ such that $P(A) := Q(A)/Q(\mathcal{G})$ for all $A \in \Sigma$ follows a $\mathrm{DP}(\alpha)$ (Ferguson, 1973).

Now let $\{K(x; \cdot) : x \in \mathcal{G}\}$ be a family of probability density functions indexed by $\mathcal{G}$. Then the mixture of K by a $\mathrm{DP}(\alpha)$ (in the sense of definition 1.7) is almost-surely a probability density function.

Finally, note that the conditions on K established after definition 1.7 to ensure that $(\omega, y) \mapsto \int K(x; y)\, P(dx)$ is a well-defined random variable at all $y \in \mathbb{R}^d$ can be slightly

weakened when $P \sim \mathrm{DP}(\alpha)$. Indeed, from Feigin and Tweedie (1989, theorem 4), it suffices to have $\int \log(1 + |K(x;y)|)\,\alpha(dx) < +\infty$ for all $y \in \mathbb{R}^d$.

1.2.2. Symmetric Gamma Process Mixtures

A common theme of this dissertation is the use of mixture models outside the popular framework of density estimation with Dirichlet Process Mixtures, in particular in direct or inverse regression. In that case, Dirichlet Processes are no longer a natural choice of random measures, and we shall in general need *random signed measures*. *Lévy Random Measures* (LRM)[1] (Wolpert et al., 2011), seem then to be an ideal choice for our purpose, since they are easy to construct and amenable to posterior computations. However, the study of posterior distributions for general LRM is still challenging, so that we restrict ourselves to *symmetric Gamma random measures*, defined as follows.

Definition 1.9 — Symmetric Gamma random measure. – *Let α be a non null finite positive measure on $(\mathcal{G}, \Sigma)$ and $\eta > 0$. We say that a random signed measure $Q : \Omega \times \Sigma \to \mathbb{R}$ is a symmetric Gamma random measure with base measure α and scale parameter η if, for all $k \in \mathbb{N}^*$ and all disjoint measurable sets $A_1, \dots, A_k \in \Sigma$ the random variables $Q(A_1), \dots, Q(A_k)$ are independent and distributed as the difference of two independent* $\mathrm{Gamma}(\alpha(A_i), \eta)$ *random variables for all $i = 1, \dots, k$.*

Remark 1.1 – In some problems, it is not worth having recourse to symmetrized random measure; for instance in chapter 4. Obviously, when we are in such situation, we should rely on (nonsymmetric) Gamma random measures, with obvious definition. Moreover, all the results of this section translates to the Gamma process in a straightforward fashion.

Now we devote the following sections to constructive definitions of symmetric Gamma random measures and symmetric Gamma process mixtures in terms of the *Poisson Process*.

Poisson random measures

For a general treatment of Poisson random measures, we refer to Çınlar (2011); Kingman (1992). Let $(E, \mathcal{U})$ be a measurable, locally compact and separable space. Let ν be a σ-finite positive measure on $(E, \mathcal{U})$. A random measure $\mathcal{N} : \Omega \times \mathcal{U} \to \mathbb{N} \cup \{+\infty\}$ is a Poisson random measure with mean ν if

(1) For each $A \in \mathcal{U}$, $\mathcal{N}(\cdot, A) \sim \mathrm{Po}(\nu(A))$

(2) If $A_1, \dots, A_n$ are pairwise disjoint sets in $\mathcal{U}$ then $\mathcal{N}(\cdot, A_1), \dots, \mathcal{N}(\cdot, A_n)$ are independent Poisson distributed random variables.

If ν is a σ-finite but not finite measure on $(E, \mathcal{U})$ the above definition still makes sense (because the definition below is consistent with the infinite limit of the Poisson law) if we define,

(3) For all sets $A \in \mathcal{U}$ with $\nu(A) = +\infty$, then $\mathcal{N}(\cdot, A) = +\infty$ almost surely.

[1] One find the name *Infinitely divisible and independently scattered random measures* in Rajput and Rosinski (1989), or also *Lévy bases* in Barndorff-Nielsen and Schmiegel (2004).

Let $\mathcal{N}$ be a PRM with mean ν and $A \in \mathcal{U}$ a Borel set, the characteristic function of the random variable $\mathcal{N}(\cdot, A)$ is,

$$\mathbb{E}[e^{it\mathcal{N}(\cdot,A)}] = \exp\left\{-\nu(A)(1-e^{it})\right\}. \tag{1.3}$$

When $\nu(E) < +\infty$, there is a convenient way to construct and interpret PRM. Indeed, start with the probability measure $\pi(\cdot) = \nu(\cdot)/\nu(E)$ on $(E,\mathcal{U})$ and let $K \sim \text{Po}(\nu(E))$, $X_k \overset{iid}{\sim} \pi(\cdot)$ for $1 \leq k \leq K$. From this sample form the following measure,

$$\mathcal{N}(A) = \sum_{i=1}^{K} \mathbb{1}_A(X_k), \quad A \in \mathcal{U}. \tag{1.4}$$

An easy computation shows that the characteristic function of equation (1.4) is nothing else but equation (1.3) and thanks to the uniqueness of the Fourier transform, the random measure defined in equation (1.4) is a PRM with mean ν. When $\nu(E) = +\infty$ we can find a disjoint partition $\{E_i\}_{i=1}^{\infty}$ of E such that $\nu(E_i) < +\infty$ for all $i = 1, \ldots, \infty$. Now define $\mathcal{N}_i$ the PRM with mean $\nu(E_i)$ on the subset E_i. By the above discussion $\mathcal{N}_i$ is almost-surely purely atomic. Let $\mathcal{N}(A) = \sum_{i=1}^{\infty} \mathcal{N}_i(A \cap E_i)$ for all $A \in \mathcal{U}$. As $\mathcal{N}_i(A \cap E_i)$ are independent Poisson random variables with means $\nu(E_i)$, it follows that $\mathcal{N}(A)$ is a Poisson random variable with mean $\nu(E) = +\infty$ and hence $\mathcal{N}$ is a Poisson random measure with mean ν. As a consequence we have the following proposition.

Proposition 1.2 – *Let ν be a σ-finite measure on $(E,\mathcal{U})$ and $\mathcal{N}$ a PRM with mean ν. Then,*

(1) $\mathcal{N}$ is almost surely purely atomic.

(2) If $\nu(E) < +\infty$, then $\mathcal{N}$ as almost-surely a finite number of atoms.

(3) If $\nu(E) = +\infty$, then $\mathcal{N}$ as almost-surely a countably infinite number of atoms, but for all compact $A \in \mathcal{U}$ with $\nu(A) < +\infty$, $\mathcal{N}$ has almost-surely a finite number of atoms in A.

Poisson representation of symmetric Gamma process mixtures

Let $\eta > 0$, $H_\eta(du) := |u|^{-1}e^{-|u|\eta}du$, α be a non null finite positive measure on $(\mathcal{G}, \Sigma)$ and $\mathcal{N}$ be the PRM with mean $H_\eta \times \alpha$ on $(\mathbb{R} \times \mathcal{G}, \mathcal{B}_\mathbb{R} \otimes \mathcal{G})$. Then (see for instance Wolpert et al. (2011)) the random measure Q such that

$$Q(\cdot, A) := \int_{\mathbb{R}\times A} u\mathcal{N}(\cdot, dudx), \quad A \in \Sigma$$

has the distribution of a symmetric Gamma random measure with base measure α and scale parameter η. Moreover, given a jointly measurable kernel $K : \mathcal{G} \times \mathbb{R}^d \to \mathbb{R}$, the mixture of K by Q can be represented as

$$\theta(\cdot, y) = \int_{\mathbb{R}\times\mathcal{G}} uK(x;y)\,\mathcal{N}(\cdot, dudx) \equiv \sum_{i=1}^{\infty} u_i K(x_i; y),$$

where $(u_i, x_i)_{i=1}^{\infty}$ are the atoms of a Poisson random measure with mean $H_\eta \times \alpha$; the series has almost-surely an infinite number of terms, and converges (pointwise) almost-surely if $\int_{\mathbb{R}\times\mathcal{G}} \min(1, |uK(x;y)|)H_\eta(du)\alpha(dx) < +\infty$ for all $y \in \mathbb{R}^d$. For other modes of convergence, see Wolpert et al. (2011, section 4).

Other representations and connections

Just to mention a few, other representations of the symmetric Gamma random measures, based on the Poisson process, may be found on Rosiński (2001).

Also, it is well known (see for instance Tsilevich et al., 2000, lemma 2) that if $Q : \Omega \times \Sigma \to \mathbb{R}^+$ is a (nonsymmetric) Gamma random measure with base measure α on $(\mathcal{G}, \Sigma)$ and scale parameter $\eta > 0$, then the random variable $\omega \mapsto Q(\omega, \mathcal{G})$ and the random variables $\omega \mapsto Q(\omega, A)/Q(\omega, \mathcal{G})$ for all $A \in \Sigma$ are independent. Hence, the construction of the Dirichlet random measure from Gamma random measures can be easily inverted, allowing to represent Gamma random measures in term of the Dirichlet Process.

The latter connection between Dirichlet Process and Gamma random measures is exploited in section 1.3.2 below to design an algorithm for simulating the posterior of Gamma process mixtures, and in a lesser extent has been one of the motivation of the algorithm proposed in chapter 2.

1.2.3. Concrete examples of mixtures

So far we discussed nonparametric mixture models in all their generality. Here we present some concrete examples with related existing asymptotic results.

In the sequel for any random measure $Q : \Omega \times \Sigma \to \mathbb{R} \cup \{\pm\infty\}$, $Q(A)$ is understood as $Q(\cdot, A)$. Similarly, we make implicit the dependence on $\omega \in \Omega$ of all functional of Q.

Location and Location-scale mixtures

Given a measurable *mother* function $g : \mathbb{R}^d \to \mathbb{R}$, we define the *location-scale* kernel $K(A, \mu; y) \equiv K_A(x - \mu) := \det(A)^{-1/p} g(A^{-1}(x - \mu))$ for all $x \in \mathbb{R}^d$ and all $A \in \mathcal{E}$, where $\mathcal{E}$ is the set of all $d \times d$ positive define real matrices; it may be convenient to adapt the power factor $1/p$ depending on the statistical model considered. In probability density estimation, g is a probability density and $p = 1$, while this constraint is not necessary in regression.

We recall the following definitions of a location mixture and, respectively, of a location-scale mixture prior.

Definition 1.10 – *Let $\mathcal{E}_0 \subset \mathcal{E}$ and $\mathbb{R}^d$ be endowed with Borel σ-algebra. Let Π_A be a prior on $\mathcal{E}_0$, and $Q : \Omega \times \mathcal{B}(\mathbb{R}^d) \to \mathbb{R} \cup \{\pm\infty\}$ be a random measure. Then $\theta(x) := \int K_A(x - \mu)\, Q(d\mu)$, where $A \sim \Pi_A$, is a location mixture of g by Q (provided it is a well-defined random variable at all $x \in \mathbb{R}^d$).*

Definition 1.11 – *Let $\mathcal{E}_0 \subset \mathcal{E}$ and let $\mathcal{E}_0 \times \mathbb{R}^d$ be endowed with Borel σ-algebra. Let $Q : \Omega \times \mathcal{B}(\mathcal{E}_0 \times \mathbb{R}^d) \to \mathbb{R} \cup \{\pm\infty\}$ be a random measure. Then $\theta(x) := \int K_A(x - \mu)\, Q(dA d\mu)$ is a location-scale mixture of g by Q (provided it is a well-defined random variable at all $x \in \mathbb{R}^d$).*

Posterior asymptotics for location and location-scale mixtures is not fully understood yet, but a lot of things — especially for density estimation — are known.

In Ghosal et al. (1999); weak and L^1 consistency for Dirichlet process location and location-scale mixtures of Gaussians — that is $g(x) = \exp(-x^2/2)$ — is investigated in univariate ($d = 1$) density estimation. They established conditions on the prior to give pos-

terior consistency at θ_0 when $\theta_0(x) = \int K_{\sigma_0}(x-\mu)\, dP_0(\mu)$ with some compactly supported probability measure P_0, or θ_0 is compactly supported and $\lim_{\sigma\to 0} \int \theta_0 \log(\theta_0/K_\sigma * \theta_0) = 0$.

In the meantime, still in univariate density estimation, Genovese and Wasserman (2000) provided a first attempt to rates of convergence for location-scale mixtures of Gaussians, although with a frequentist sieve estimator $\widehat{\theta}_n$. Under the assumption that $\theta_0(x) = \int K_\sigma(x-\mu)\, dP_0(\mu)$, they found $(t_n)_{n\geq 1}$ such that $P^n_{\theta_0}(d(\theta_0, \widehat{\theta}_n) > t_n) = o(1)$, where d is the Hellinger distance. The rates t_n crucially depend on the tails of P_0; the best rate is achieved in the compactly-supported case with $t_n = (\log n/n)^{1/4}$. Their rate is clearly suboptimal, as we expect to achieve a nearly parametric rate of $n^{-1/2}$ (up to a power of $\log n$) in this *supersmooth* situation.

Ghosal and Van Der Vaart (2001) improved the preceding result and recovered a rate of $(\log n)^\kappa/\sqrt{n}$, where κ depends on the tails of P_0. Also, assuming that P_0 has sub-Gaussian tails, and that the bandwidth σ_0 is bounded from above and below, they obtained sufficient conditions to achieve the posterior contraction rate of $(\log n)^\kappa/\sqrt{n}$ for the Hellinger distance with a Dirichlet process location mixture of Gaussians prior.

Tokdar (2006) established weak and L^1 consistency for Dirichlet process location-scale mixtures of Gaussians (again in univariate density estimation) under mild assumptions on θ_0. Particularly, θ_0 is not assumed to be itself a mixture and can be heavy-tailed.

Considering univariate density estimation with Dirichlet process location mixtures, assuming θ_0 twice continuously differentiable with compact support, Ghosal and Van Der Vaart (2007b) established a nearly minimax contraction rate of $n^{-2/5}$ (up to a power of $\log n$) for the Hellinger distance. This is the first time that posterior rates are established without assuming θ_0 supersmooth. However, if θ_0 is more than twice continuously differentiable, their rate does not improve and becomes suboptimal.

The main difficulty in improving the previous rates rely on the ability of approximating a smooth density θ_0 with a convex mixture of densities, when θ_0 is not itself a mixture[2]. With the assumption $\log \theta_0 \in \mathtt{C}^\beta$, Kruijer et al. (2010) used an idea from Rousseau (2010) to circumvent the issue, leading the way to rate adaptive posterior contraction rates of $n^{-2\beta/(2\beta+1)}$ (up to a power of $\log n$) over the whole classes of β-log-Hölder densities with exponential tails, $\beta > 0$, using Dirichlet process location mixtures of generalized Gaussians and either Hellinger or L^1 distance.

The first asymptotic result for multivariate location mixtures in regression appeared in De Jonge and Van Zanten (2010). The authors considered the Gaussian errors regression with fixed design on $[0,1]^d$ for a finite mixture model, although with a random number of components. Their result is near optimal and rate adaptive over the Hölder classes $\mathtt{C}^\beta$, $\beta > 0$, for the $\|\cdot\|_{2,n}$ distance, and holds for a quite general family of mother functions.

Scricciolo (2011) investigated univariate density estimation with Dirichlet process mixtures of exponential power densities, assuming that the true density is itself a mixture with mixing measure having sub-exponential tails. She found nearly parametric posterior rates of convergence when the mother function is differentiable at 0, and slower rates otherwise.

Regarding multivariate density estimation with Dirichlet process location mixtures of Gaussians, Shen et al. (2013) established near optimal and adaptive rates over anisotropic Hölder classes for the Hellinger distance, provided exponential tails assumptions on θ_0.

[2] Whereas approximating a smooth density with a non convex mixture is relatively easy, it is not as simple to achieve a nice order of approximation with convex mixtures.

Their approach allowed to release the assumption $\log \theta_0 \in \mathtt{C}^\beta$ used in Kruijer et al. (2010) onto $\theta_0 \in \mathtt{C}^\beta$, provided supplementary conditions on ratio of the form $D^k f / f^p$.

Canale and De Blasi (2017) extended the previous result to multivariate density estation with Dirichlet process location-scale mixtures of Gaussians. They obtained suboptimal rates for the Hellinger distance over the Hölder classes, assuming exponential tails for the true density. Moreover, they conjectured that under weaker tail assumptions their bound on the rate remains true and should be near optimal. We believe today that this conjecture is wrong (see below).

Recently, Scricciolo (2014) investigated univariate density estimation with Pitman-Yor process and normalized Inverse-Gamma process location mixtures. The novelty here is that she looked at contraction rates under L^p metric, for $1 \leq p \leq \infty$. For a class of analytic functions, she found that the procedure is nearly optimal, provided θ_0 has exponential tails. However, regarding at β-Sobolev classes, $\beta < \infty$, the rates are optimal only when $1 \leq p \leq 2$, and deteriorate by a genuine power of n otherwise.

Some questions remain open about location and location-scale mixtures. Why do the rates found in Scricciolo (2014) for the L^p distance deteriorate when $p \geq 2$ and the true function is not analytic ? We even don't know if there is a Bayesian procedure that achieve the optimal rates in this situation (Giné and Nickl, 2011).

Another interesting issue concerns the exponential tail assumption on θ_0. Currently it appears difficult to get rates of contraction for density estimation with location or location-scale mixture without light tail assumption on θ_0. Under light-tail hypothesis, everything happens as if θ_0 is compactly supported, so that it is not surprising that location mixtures perform better than location-scale mixtures (Shen et al., 2013; Canale and De Blasi, 2017). It is known from the frequentist litterature (Juditsky and Lambert-Lacroix, 2004; Reynaud-Bouret et al., 2011; Goldenshluger and Lepski, 2014) that without tails assumptions, the optimal rates for β-Hölder classes, for the L^p distance, deteriorate when $1 \leq p \leq 2 + 1/\beta$ in comparison to the optimal rates under tail condition. For $p = 1$, and hence for the Hellinger distance, the rates deteriorate to $\epsilon_n^2 = 1$ if no tails assumptions are made, and may take a bunch of values between 1 and $n^{-2\beta/(2\beta+1)}$ depending on the heaviness of the tail. This explains why Canale and De Blasi (2017) conjectured that the suboptimal rates they found with location-scale mixtures of Gaussians under light tail assumptions may become optimal if they could release the tail condition. We do not think that this conjecture is correct (see also chapter 3), and we present here an argument to explain why we think the conjecture is false.

Assume $d = 1$ for simplicity. Define the pointwise Hölder regularity as follows. Let $\theta : \mathbb{R} \to \mathbb{R}$ be a function, $\alpha > 0$, $\alpha \notin \mathbb{N}$ and $x_0 \in \mathbb{R}$. Then $\theta \in C^\alpha(x_0)$ if and only if there exists a real $\eta > 0$, a polynomial P with degree less than α and a constant $C > 0$ such that $|f(x) - P(x - x_0)| \leq C|x - x_0|^\alpha$ for all $|x - x_0| \leq \eta$. The pointwise Hölder exponent of θ at x_0 is $\beta(x_0) = \sup\{\alpha : \theta \in C^\alpha(x_0)\}$. Returning to (global) Hölder spaces $\mathtt{C}^\beta$, there are functions for which $\beta(x) = \beta$ for all $x \in \mathbb{R}$. This seemlingly innocuous fact is indeed crucial, because for these functions, there is no reason for location-scale mixtures to perform better order of approximation than location mixtures. One could argue that there is no need to approximate θ_0 with the same order of approximation in zones with few data as in zone of dense data, encouraging location-scale mixtures. This argument is hard to implement in density estimation, but could be rather easily applied in regression with Gaussian errors (see chapter 3), still leading to suboptimal rates for location-scale

mixtures.

In chapter 3 for random design regression, we obtain rates for both location and location-scale mixtures under various tail constraints, which are always worst for location-scale mixtures than for location mixtures. The novelty here is that, under heavy tail assumptions, both location and location-scale appear to be suboptimal and we can find hybrid procedure that performs always better than both types of mixtures. That said, to be able to accept or reject definitively Canale and De Blasi's conjecture, it is necessary to establish lower bounds on the rate of contraction of the posterior (in the sense of Castillo (2008)), which remains a challenge for mixture models.

Finally, we believe that although location-scale mixtures are suboptimal over Hölder classes, they should outperform location mixtures if looking at classes of locally varying smoothness, making them useful in practice.

Mixtures of Betas and Gammas

Another common examples are mixtures of Beta densities with respect to a random probability measure. Although this prior makes sense only in density estimation over $[0, 1]$, we would like to mention a few asymptotic results, in particular because the tools developed in the next references have been useful in the study of posterior distribution for general mixtures.

Consistency for Dirichlet process mixtures of Bernstein polynomials goes back to Petrone and Wasserman (2002) under very mild assumptions on the true density. For some prior distributions, Ghosal (2001) established nearly parametric rates of convergence for the Hellinger distance under the assumption that the true density is itself of mixture of Bernstein polynomials, and provided suboptimal rates otherwise.

Regarding at Dirichlet process mixtures of Beta densities, Kruijer and van der Vaart (2008) established posterior rates over β-Hölder classes for $0 < \beta \leq 2$, in Hellinger distance, achieving optimal adaptive rate over $0 < \beta \leq 1$, and suboptimal otherwise. For the L^1 distance, Rousseau (2010) is the first article in which mixture priors are shown to yield nearly optimal and adaptive rates over the whole β-Hölder classes for any $\beta > 0$, paving the way for the results of De Jonge and Van Zanten (2010); Kruijer et al. (2010); Shen et al. (2013) in location mixtures.

Recently, Bochkina and Rousseau (2016) investigated posterior contraction rates for density estimation on the positive real line based on mixtures of Gamma distributions. They required less stringent conditions on the tails of the true density than the usual exponential tail condition required for location and location-scale mixtures. Together with the result of Canale and De Blasi (2017), this paper was a supplementary motivation for the work presented in chapter 3.

1.2.4. Contributions

This section presents the main contribution of this manuscript to theoretical results about symmetric Gamma process mixtures.

Fixed design regression with symmetric Gamma process mixtures

In chapter 2, we are concerned with the problem of a random response Y corresponding to a deterministic covariate vector x taking values in $[-S,S]^d$ for some $S>0$. We aim at estimating the regression function $f:[-S,S]^d\to\mathbb{R}$ such that $f(x_i)=\mathbb{E}\,Y_i$, based on independent observations of Y. More precisely, the nonparametric regression model we consider is the following,

$$\begin{aligned} Y_i|\epsilon_i &= f(x_i)+\epsilon_i, \quad i=1,\dots,n,\\ \epsilon_1,\dots,\epsilon_n|\sigma^2 &\overset{\text{i.i.d}}{\sim}\mathcal{N}(0,\sigma^2), \quad \text{independently of } (f,\sigma),\\ (f,\sigma) &\sim \Pi, \end{aligned}$$

with Π the distribution on an abstract space Θ, given by $\sigma\sim P^\sigma$ independently of f drawn from the distribution of a symmetric Gamma process mixture.

We consider two families of mixtures. First, the location-scale mixtures, as defined in definition 1.11; and second we introduce a new type of mixtures, which we call *location-modulation* mixtures, defined as follow.

For a measurable mother function $g:\mathbb{R}^d\to\mathbb{R}$, we define the *location-modulation* kernel $K_{\xi,\phi}(x):=g(x)\cos(\sum_{i=1}^d\xi_i x_i+\phi)$, for all $x\in\mathbb{R}^d$, all $\xi\in\mathbb{R}^d$ and all $\phi\in[0,\pi/2]$. Then, a mixture of $K_{\xi,\phi}$ by a symmetric Gamma random measure $Q:\Omega\times\mathcal{B}(\mathbb{R}^d\times\mathbb{R}^d\times[0,\pi/2])\to[-\infty,\infty]$ is the distribution of the random function (for those g for which it is well-defined)

$$f(\cdot,x):=\int_{\mathbb{R}^d\times\mathbb{R}^d\times[0,\pi/2]}K_{\xi,\phi}(x-\mu)\,Q(\cdot,d\xi d\mu d\phi),\qquad \forall x\in\mathbb{R}^d\,.$$

Our main result concerns the posterior rates of convergence for the two families of prior distributions. Assuming that the true regression function f_0 belongs to the Hölder space $\mathtt{C}^\beta[-S,S]^d$, we find that the parameters of the mixing symmetric Gamma random measure and g can be chosen so that the following posterior rates of convergence hold for these models, with the empirical metric $d_n((f_1,\sigma_1),(f_2,\sigma_2)):=(n^{-1}\sum_{i=1}^n|f_1(x_i)-f_2(x_i)|^2)^{1/2}+|\log\sigma_1-\log\sigma_2|$:

- Location-scale mixtures: there is $t>0$ such that $\epsilon_n^2\lesssim n^{-\beta/(2\beta+d+\kappa/2)}(\log n)^t$, where $\kappa>0$ depends on the parameters of the prior.
- Location-modulation mixtures: there is $t>0$ such that $\epsilon_n^2\lesssim n^{-\beta/(2\beta+d)}(\log n)^t$.

Regarding location-scale mixtures, the same phenomena as in density estimation with DPM take place here (Canale and De Blasi, 2017), yielding to suboptimal rates. However, it seems like location-modulation behave better over the β-Hölder classes as they achieve the minimax rate of convergence (up to a power of $\log n$ factor).

On the tail conjecture

In chapter 3, we investigate the conjecture of Canale and De Blasi (2017) about the robustness of location-scale mixtures to heavy tailed distributions. To better understand the ability of mixture models to capture heavy tails we study their use in nonparametric

univariate regression models:

$$Y_i = f(X_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} N(0, s^2), \quad i = 1, \dots, n,$$
$$X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} Q_0, \quad f \in L^2(Q_0).$$

We assume that s is known, which is just a matter of convenience for proofs, and we wish to estimate the parameter f using location or location-scale mixtures of Gaussians by symmetric Gamma random measures. Our aim is to study posterior concentration rates around the true regression function f_0 under the empirical ℓ_2 distance of the covariates, defined as $d_n(f, f_0)^2 := n^{-1} \sum_{i=1}^n |f(x_i) - f_0(x_i)|^2$. We assume that $f_0 \in L^1$ and belongs to a Hölder ball with smoothness β. The tail condition are then on the design distribution and written as $\int_{\mathbb{R}} |x|^p dQ_0(x) < +\infty$, $p \geq 0$.

We show in chapter 3, that in most cases location mixtures have a better posterior concentration rate than location-scale mixtures and unless p goes to infinity the posterior concentration rates is not as good as the usual $n^{-\beta/(2\beta+1)}$. This rate is suboptimal for light tail design points, since in this case the minimax posterior concentration rate is given by $n^{-\beta/(2\beta+1)}$. To improve on this rate we propose a new version of location-scale mixture models, which we call the hybrid location-scale mixture and we show that this nonparametric mixture model leads to better posterior concentration rates than the location mixture (and thus than the location-scale mixture). All these results are up to $\log n$ terms. The results are summarized in table 1.1 which displays the value q defined by $\epsilon_n^2 = n^{-q}$.

Table 1.1: Summary of posterior rates of convergence for different types of mixtures. The rates are understood to be in the form $\epsilon_n^2 = n^{-q}$, up to powers of $\log n$ factors, where q is given below.

	$0 < p < 2$			$p \geq 2$		
	$0 < p \leq \frac{2\beta}{\beta+1}$	$\frac{2\beta}{\beta+1} < p \leq 2\beta$	$p > 2\beta$	$0 < p \leq \frac{2\beta}{\beta+1}$	$\frac{2\beta}{\beta+1} < p \leq 2\beta$	$p > 2\beta$
Location	$\frac{2\beta}{3\beta+1}$	$\frac{2\beta}{3\beta+1}$	$\frac{2\beta}{3\beta+1}$	$\frac{2\beta}{2\beta+1+2\beta/p}$	$\frac{2\beta}{2\beta+1+2\beta/p}$	$\frac{2\beta}{2\beta+1+2\beta/p}$
Location-scale	$\frac{2\beta}{3\beta+2}$	$\frac{2\beta}{2\beta+1+2\beta/p}$	$\frac{2\beta}{2\beta+2}$	$\frac{2\beta}{3\beta+2}$	$\frac{2\beta}{2\beta+1+2\beta/p}$	$\frac{2\beta}{2\beta+2}$
Hybrid	$\frac{2\beta}{3\beta+1}$	$\frac{p}{p+1}$	$\frac{2\beta}{2\beta+1}$	$\frac{2\beta}{3\beta+1}$	$\frac{p}{p+1}$	$\frac{2\beta}{2\beta+1}$

Although the results are presented in the regression model, we believe that similar phenomena should take place in the density estimation problem.

Finally, we also compute the rates for location mixtures when $p = 0$ and the parameters of the mixing symmetric Gamma random measures are dependent of the covariates. The strategy can be easily translated to location-scale and hybrid mixtures.

1.3. Posterior sampling

Quite generally, in Bayesian nonparametrics, the posterior distribution is either analytically intractable, or difficult to compute. A popular way to circumvent the issue is to construct a Markov chain whose transition kernel admits the posterior distribution as

invariant distribution; such algorithms are called *Markov Chain Monte Carlo* algorithms (MCMC). Then from the samples of the Markov chain, one can (approximately) evaluate all functionals of the posterior distribution.

Although Dirichlet Process Mixtures (DPM) is not the object of this dissertation is, we first review an algorithm for sampling the posterior distribution of DPM (in density estimation model); the interest is two fold. First, it will highlight the difficulties encountered for sampling mixture models in general models. Second, the algorithm presented in the next section has been a great source of inspiration for the design of the algorithm proposed in chapter 2 (because of the close relationship between Dirichlet Processes and Gamma Processes).

We will conclude this section with an overview of the problems encountered with general mixture models, followed by a presentation of previsouly existing algorithm for posterior simulation, which could be used as alternative to the algorithm of chapter 2.

1.3.1. Gibbs sampling Dirichlet Process Mixtures

Gibbs methods for sampling the posterior distribution of DPM go back to Escobar's unpublished thesis; published later in Escobar (1994). The emergence of these algorithms signed the boom of Bayesian nonparametrics, which had a solid theoretical framework for a while, but lacked to be used effectively in practice. Gibbs algorithms for sampling the posterior of DPM may be classified in two categories.

- *Conditional samplers.* These algorithms use an explicit representation of the measure $P \sim \mathrm{DP}(\alpha)$, mainly the *stick-breaking* representation (Ishwaran and James, 2011; Papaspiliopoulos and Roberts, 2008; Walker, 2007; Kalli et al., 2011). We will not review these methods here. For an exhaustive review (among other cool Bayesian nonparametrics stuff) we recommend the unpublished thesis of Fall (2012).
- *Collapsed samplers.* These algorithms marginalize the distribution of $P \sim \mathrm{DP}(\alpha)$ in the posterior distribution, and are the first techniques developed for DPM sampling. Escobar's algorithm falls into that category (Escobar, 1994; Escobar and West, 1995).

However, Escobar's algorithm suffers from two shortcomings. First, the mixing time of Escobar's algorithm turns out to be rather huge and inefficient. Second, the algorithm works only when the likelihood of the parameter is conjugate to the base distribution α of the mixing Dirichlet Process. The first issue is addressed for instance in Neal (1992); West and Escobar (1993); MacEachern (1994), while the second issue is addressed in MacEachern and Müller (1998); Walker and Damien (1998); Neal (2000).

We now review briefly Neal (2000, algorithm 8), which is of particular interest for the comprehension of chapter 2. We let α be a finite positive measure over $(\mathcal{G}, \Sigma)$, $\overline{\alpha} = \alpha(\mathcal{G})$, and $K : \mathcal{G} \times \mathbb{R}^d \to \mathbb{R}^+$ be a jointly measurable kernel such that $K(x;\cdot)$ is a probability density function for all $x \in \mathcal{G}$. Neal's algorithm is based on the following hierarchical representation of DPM.

$$\begin{aligned} y_i|x_i &\sim K(x_i;\cdot) \quad \forall i = 1,\dots,n \\ x_1,\dots,x_n|P &\overset{\text{i.i.d}}{\sim} P \\ P &\sim \mathrm{DP}(\alpha). \end{aligned} \tag{1.5}$$

The samples $y_1, \dots, y_n$ from the model of equation (1.5) are identically distributed with distribution density $y \mapsto \int K(x; y)\, dP(y)$, where $P \sim \mathrm{DP}(\alpha)$. If we marginalize P in the equation (1.5), then $x_1, \dots, x_n$ has the distribution of a Pólya urn sequence with base measure α (see section 1.2.1); this fact is crucial and is one of the foundation of the algorithm developed later in chapter 2.

Going back to Neal's algorithm and equation (1.5), each sample y_i is associated with a component x_i of the mixture. However, because of the almost-sure discreteness of the Dirichlet Process, there might be some identical values of x_i. We let $x_1^\star, x_2^\star, \dots$ denote the unique values of $x_1, \dots, x_n$, and we introduce the latent clustering variables $c_1, \dots, c_n$ such that $c_i = k$ if $x_i = x_{c_i}^\star = x_k^\star$. Observation y_i with $c_i = k$ is said to belong to cluster k. Neal's algorithm produces a Markov chain whose invariant distribution is the posterior distribution of $(c_1, \dots, c_n, x_1^\star, x_2^\star, \dots)$. The algorithm draws samples by successively sampling from

- $c_1, \dots, c_n | y_1, \dots, y_n, x_1^\star, x_2^\star, \dots$ using the auxiliary variables method, as described below;
- $x_1^\star, x_2^\star, \dots | y_1, \dots, y_n, c_1, \dots, c_n$ using the Metropolis-Hastings algorithm (Hastings, 1970), or any other method which allows to draw samples from this distribution.

We now briefly describe the *auxiliary variables method*. Assume that you want to draw samples from a distribution π_x. Then you can perform the following update scheme for x. First choose a joint distribution π_{xy} for (x, y) that admits π_x as marginal. Then draw a sample of y from the conditional distribution of $y|x$ followed by a update of (x, y) that leave π_{xy} invariant. Finally discard y and keep only x. This update for x will keep π_x invariant.

Neal (2000) uses the auxiliary variable methods to update $(c_i)_{i=1}^n$. He uses $m \in \mathbb{N}^*$ auxiliary variables representing new potential clusters. The choice of m is governed by a trade-off between computation cost and mixing efficiency.

Each c_i is updated using Gibbs sampling. Since the observations $(y_i)_{i=1}^n$ are exchangeable, we can assume we are updating c_i for the last observation. We let k^- denote the number of clusters that contain at least one observation y_i for $i = 1, \dots, n-1$, and we label these clusters with distinct values in $\{1, \dots, k^-\}$. The prior probability of c_i to be equal to $1 \le l \le k^-$ is $n_{-i,l}/(n-1+\overline{\alpha})$, where $n_{-i,l}$ denote the number of observations y_j with $j \neq i$ in cluster l. The probability of c_i having other value is $\overline{\alpha}/(n-1+\overline{\alpha})$, which is split among the m auxiliary clusters that are associated with parameters $\overline{x}_1, \dots, \overline{x}_m$ drawn independently from $\alpha(\cdot)/\overline{\alpha}$. As shown in Neal (2000), this representation of the prior translates to the posterior distribution, yielding the algorithm 1.

1.3.2. Sampling symmetric Gamma process mixtures

When dealing with other statistical models than density estimation with DPM, in general we cannot associate an observation to a unique component of the mixture, as in equation (1.5). In chapter 2, we will introduce a representation of the symmetric Gamma process which allows to adapt algorithm 1 to symmetric Gamma process mixtures. Meanwhile, we present here the two main existing alternatives to the algorithm we propose, taken respectively from Wolpert et al. (2002, 2011) and Erhardsson (2008). The goal of these algorithms is the use of mixtures outside the density estimation framework.

Algorithm 1 Neal (2000, algorithm 8)

Let $m \in \mathbb{N}^*$ and the state of the Markov Chain consist of $(c_1, \dots, c_n)$ and $(x_1^\star, x_2^\star, \dots)$. Repeatedly sample as follows.

- For all $i = 1, \dots, n$: Let k^- denote the number of clusters that contain at least one observation y_j for $j \neq i$, and k denote the total number of clusters.
  - If $c_i \neq c_j$ for all $j \neq i$, then $k^- = k - 1$ and do
    - $x^\star \leftarrow x_{c_i}^\star$
    - $(x_1^\star, x_2^\star, \dots, x_{c_i-1}^\star, x_{c_i+1}^\star, \dots) \leftarrow (x_1^\star, x_2^\star, \dots, x_{c_i-1}^\star, x_{c_i}^\star, x_{c_i+1}^\star, \dots)$
    - $x_k^\star \leftarrow x^\star$ and for all k with $c_k \geq c_i : c_k \leftarrow c_k - 1$
    - Draw values independently from $\alpha(\cdot)/\alpha(\mathcal{G})$ for those $x_l^\star$ for which $k+1 \leq l \leq k+m-1$.
  - Else if there is one $j \neq i$ such that $c_i = c_j$, then $k^- = k$ and do
    - Draw values independently from $\alpha(\cdot)/\alpha(\mathcal{G})$ for those $x_l^\star$ for which $k+1 \leq l \leq k+m$.

  Draw a new value for c_i from $\{1, \dots, k^- + m\}$ using the following probabilities

$$\mathbb{P}(c_i = l \mid c_{-i}, y_i, x_1^\star, x_2^\star, \dots) \propto \begin{cases} \frac{n_{-i,l}}{n-1+\alpha(\mathcal{G})} K(x_l^\star; y_i) & 1 \leq l \leq k^- \\ \frac{\alpha/m}{n-1+\alpha(\mathcal{G})} K(x_l^\star; y_i) & k^- < l \leq h. \end{cases}$$

  Change the state to contain only those $x_1^\star, x_2^\star, \dots$ that are associated to at least one observation (relabel the clusters if necessary).
- Sample from $(x_1^\star, x_2^\star, \dots)|c_1, \dots, c_n, y_1, \dots, y_n$.

Note that Wolpert et al.'s algorithm serves as a reference to assess the performance of our algorithm in chapter 2.

Wolpert's Birth-Death algorithm

Wolpert et al.'s algorithm is based on a finite approximation of the Poisson representation of symmetric Gamma process mixtures (see section 1.2.2). Note that the Poisson representation is not restricted to symmetric Gamma Process mixtures, and so is Wolpert et al.'s algorithm.

Let α be a finite positive measure on $(\mathcal{G}, \Sigma)$, $\eta > 0$, $\epsilon > 0$ and define $H_{\eta,\epsilon}(du) := \mathbb{1}_{|u\eta|>\epsilon}(u)|u|^{-1}e^{-|u|\eta}du$. Let $K : \mathcal{G} \times \mathbb{R}^d \to \mathbb{R}$ be a jointly measurable kernel such that $\int \min(1, |uK(x;y)|)\, H_{\eta,0}(du)\alpha(dx) < +\infty$, and $\mathcal{N}_\epsilon$ be the Poisson random measure with mean $H_{\eta,\epsilon} \times \alpha$. For all $y \in \mathbb{R}^d$, the mixture $\theta_\epsilon(\cdot, y) := \int uK(x;y)\mathcal{N}_\epsilon(\cdot, dudx)$ converges in probability to a mixture of K by a symmetric Gamma random measure with parameters (α, η) as $\epsilon \to 0$ (Wolpert et al., 2011).

Let E_1 be the exponential integral E_1 function defined by $E_1(x) := \int_x^\infty t^{-1}e^{-t}\,dt$ for all $x > 0$ and let $P_{\epsilon,u}$ be the distribution which has density $\mathbb{1}_{|u\eta|>0}(u)|u|^{-1}e^{-|u|\eta}/(2E_1(\epsilon))$ with respect to the Lebesgue measure on $\mathbb{R}$. By virtue of section 1.2.2, $\theta_\epsilon(y)$ has the almost-sure representation $\theta_\epsilon(y) = \sum_{j=1}^J u_j K(x_j; y)$, where $J \sim \text{Po}(2\alpha(\mathcal{G})E_1(\epsilon)))$, $u_j \overset{\text{i.i.d}}{\sim} P_{\epsilon,u}$ for

$j = 1, \dots, J$, and $x_j \overset{\text{i.i.d}}{\sim} \alpha(\cdot)/\alpha(\mathcal{G})$ for all $j = 1, \dots, J$, independently of $(u_j)_{j=1}^J$.

Using the previous approximation of symmetric Gamma process mixtures, Wolpert et al. (2002, 2011) implement a *reversible jumps MCMC algorithm* (Green, 1995) with birth-and-death move proposals (Cappé et al., 2003), which we summarize in algorithm 2 (see also the manuscript thesis of Tu (2006) for complements).

Algorithm 2 Wolpert et al. (2011, Birth-Death algorithm)

Let $p_b, p_d \in (0,1)$ with $p_b + p_d < 1$ be parameters of the algorithm.

- At $t = 0$ initialize the Markov chain at S^0.
- At iteration $t \geq 1$:
  - With probability p_b: $S^t \leftarrow \texttt{BIRTH}(S^{t-1})$.
  - With probability $1 - p_b$: Select randomly an index j uniformly in $\{1, \dots, J^{t-1}\}$.
    - With probability $1 - p_b - p_d$: Draw $u_j^* = u_j^{t-1} + N(0, \sigma_u^2)$. If $|u_j^* \eta| \leq \epsilon$ then $S^t \leftarrow \texttt{DEATH}(j, S^{t-1})$. Otherwise $S^t \leftarrow \texttt{UPDATE}(j, u_j^*, S^{t-1})$.
    - With probability p_d: $S^t \leftarrow \texttt{DEATH}(j, S^{t-1})$.

We now clarify the notations used in algorithm 2, as well as details of the three `BIRTH`, `DEATH` and `UPDATE` moves involved in the algorithm. We use the vector notations $u := (u_1, \dots, u_J)$ and $x = (x_1, \dots, x_J)$. The state of the Markov chain at iteration t is $S^t = (J^t, u^t, x^t)$, corresponding to a function $\theta^t(y) = \sum_{j=1}^{J^t} u_j^t K(x_i^t; y)$. We denote by $\mathcal{L}(S^t \mid y^n)$ the likelihood of S^t (equivalently θ^t) under the observations $y^n = (y_1, \dots, y_n)$. We let $p_b, p_d \in (0,1)$ with $p_b + p_d < 1$, and $p_u = 1 - p_b - p_d$. The main steps involved in the algorithm are the following.

- `BIRTH` step. Set $J^* = J^{t-1} + 1$ and draw uniformly a index j in $\{1, \dots, J^*\}$. Draw $u_j^* \sim P_{\epsilon,u}$ and $x_j^* \sim \alpha(\cdot)/\overline{\alpha}$. It is possible to draw u_j^* directly from $P_{\epsilon,u}$ by numerically inverting the CDF; otherwise draw from other distribution and adapt the Metropolis-Hastings ratio below. Set $u^* = (u_1^{t-1}, \dots, u_{j-1}^{t-1}, u_j^*, u_j^{t-1}, \dots)$ and $x^* = (x_1^{t-1}, \dots, x_{j-1}^{t-1}, x_j^*, x_j^{t-1}, \dots)$. With probability $1 \wedge H$ accept the proposal and set $S^t = S^*$; else reject the proposal and set $S^t = S^{t-1}$. The Metropolis Hastings ratio H for this move is

$$H = \frac{\mathcal{L}(S^* \mid y^n)}{\mathcal{L}(S^{t-1} \mid y^n)} \frac{2\overline{\alpha} E_1(\epsilon)}{J^*} \frac{\left[p_d + p_u \left(\Phi\left(\frac{-u_j^* + \epsilon/\eta}{\sigma_u}\right) - \Phi\left(\frac{-u_j^* - \epsilon/\eta}{\sigma_u}\right)\right)\right] / J^*}{p_b / J^*},$$

  where Φ is the CDF of the $N(0,1)$ distribution.
- `UPDATE` step. When this step is implemented, we have at our disposal a random index j and a candidate u_j^*. We obtain S^t after the following sub-steps.
  - First, we accept the proposal u_j^* with probability $1 \wedge H$, where H is the classical Metropolis Hastings ratio for this move.
  - Second, we update x_j^{t-1} for x_j using random-walk Metropolis-Hastings.
- `DEATH` step. Set $u^* = (u_1^{t-1}, \dots, u_{j-1}^{t-1}, u_{j+1}^{t-1}, \dots)$ and $x^* = (x_1^{t-1}, \dots, x_{j-1}^{t-1}, x_{j+1}^{t-1}, \dots)$. Let $J^* = J^{t-1} - 1$, $S^* = (J^*, u^*, x^*)$. With probability $1 \wedge H$ we accept the proposal

and set $S^t = S^*$, otherwise we reject the proposal and set $S^t = S^{t-1}$. The Metropolis Hastings ratio H for this move is

$$H = \frac{\mathcal{L}(S^* \mid y^n)}{\mathcal{L}(S^{t-1} \mid y^n)} \frac{J^{t-1}}{2\overline{\alpha} E_1(\epsilon)} \frac{p_b / J^*}{\left[p_d + p_u \left(\Phi\left(\frac{-u_j^{t-1} + \epsilon/\eta}{\sigma_u} \right) - \Phi\left(\frac{-u_j^{t-1} - \epsilon/\eta}{\sigma_u} \right) \right) \right] / J^*}.$$

Erhardsson's importance sampling algorithm

Erhardsson (2008) proposed an importance sampling algorithm for simulation of Gamma process mixtures. His algorithm should be rather straightforwardly symmetrized to handle symmetric Gamma process mixtures.

Broadly, Erhardsson (2008) used the Markov Chain from Feigin and Tweedie (1989) — whose invariant distribution is a Dirichlet Process — in combination with the discussion of section 1.2.2 regarding the representation of Gamma random measures in term of the Dirichlet Process. This yields the algorithm 3, where we use essentially the same notations as in previous section. We should not insist more since the algorithm is relatively inefficient, however, the idea is interesting and might be improved in the light of recent progress made around importance sampling algorithms.

Algorithm 3 Erhardsson (2008) Importance sampling algorithm

- Initialize $G_0(y) = K(x_0; y)$ for some arbitrary $x_0 \in \mathcal{G}$.
- Let $\{U_i : i \geq 1\}$ iid. with distribution $\text{Beta}(\alpha(\mathcal{G}), 1)$, $\{x_i : i \geq 1\}$ iid. and independent of $\{U_i : i \geq 1\}$ with distribution $\alpha(\cdot)/\alpha(\mathcal{G})$, and independently of everything else $\{Z_i : i \geq 1\}$ be iid. with distribution $\text{Ga}(\alpha(\mathcal{G}), \eta)$.
- For $1 \leq i \leq N$, let $G_i(y) = (1 - U_i)K(x_i; y) + U_i G_{i-1}(y)$, and $G_i^\dagger(y) = Z_i G_i(y)$.
- Compute the importance weights $w_i = \mathcal{L}(G_i^\dagger \mid y^n)$.
- Define the mixture $\widehat{X}_n$ such that $\mathbb{P}(\widehat{X}_N = G_i^\dagger \mid G_1^\dagger, \ldots, G_N^\dagger) = w_i / \sum_{i=1}^N w_i$ for all $i = 1, \ldots, N$. Then $\widehat{X}_n(y)$ converges in probability to a $\theta(y)$, where θ is a Gamma process mixtures of K with base measure α and scale parameter η.

1.3.3. Contribution to posterior sampling of symmetric Gamma process mixtures

Recall the definition of a Pólya urn sequence from section 1.2.1. We say that a random variable J has a $\text{SGa}(1,1)$ distribution if it is distributed as the difference of two independent $\text{Ga}(1,1)$ random variables. Then, inspired from a paper of Favaro et al. (2012), we prove in chapter 2 the following representation of symmetric Gamma random measures. The convergence of signed measures in the theorem is understood with respect to the weak-* topology on the space of signed Radon measures, identified to the dual space of the space of continuous functions with compact support.

Theorem 1.4 – *Let $\mathcal{X}$ be a Polish space with Borel σ-algebra, $p > 0$ be integer, $T \sim \text{Ga}(\alpha, \eta)$, independently, $J_1, \ldots, J_p \overset{\text{i.i.d}}{\sim} \text{SGa}(1,1)$, and $\{X_i \in \mathcal{X} : 1 \leq i \leq p\}$ a Pólya urn sequence with base distribution $\alpha F(\cdot)$, independent of T and of the J_i's. Define the*

random measure, $Q_p := \sqrt{T/p}\sum_{i=1}^{p} J_i\,\delta_{X_i}$. *Then* $Q_p \xrightarrow{d} Q$, *where* Q *is a symmetric Gamma random measure with base distribution* $\alpha F(\cdot)$ *and scale parameter* $\sqrt{\eta}$.

We use the previous representation of symmetric Gamma random measures to adapt algorithm 1 for sampling posterior distribution of symmetric Gamma process mixtures, as described in chapter 2.

1.4. Quantum Homodyne Tomography

This last section is devoted to the introduction of necessary material for proper understanding of chapter 4. When we studied mixtures as prior distribution in direct regression problems (chapters 2 and 3) we never lost sight of our original belief that mixtures were good priors for solving statistical inverse problems. Chapter 4 constitutes the achievement of our initial goal, we propose a mixture model for estimating the quantum state of a light beam through Quantum Homodyne Tomography measurements. Quantum Homodyne Tomography is a very challenging, nonlinear statistical inverse problem.

1.4.1. Physical background

Quantum mechanics studies the *microworld*, the physical laws taking place at the microscopic level, that cannot be adequately described by classical mechanics. Unlike classical mechanics, the predictions of quantum mechanics are probabilistic so that we cannot (in general) infer the result of a single measurement, but only the distribution of possible outcomes.

Here we formulate the basic principles of quantum mechanics using precise mathematical language. For more insights onto mathematical foundations of quantum mechanics we recommend Takhtadzhian (2008); Hall (2013). For a more physical exposition, see Cohen-Tannoudji et al. (2006). The following axioms constitute the basis of quantum mechanics.

(A1) With every quantum system is associated an infinite-dimensional separable Hilbert space $\mathscr{H}$ over $\mathbb{C}$, with inner product $\langle\cdot,\cdot\rangle$, the *space of states*.

(A2) The set of *observables* $\mathscr{A}$ of a quantum system consists of all self-adjoint operators on $\mathscr{H}$, *i.e.* $\mathscr{A} = \{A : \mathscr{H} \to \mathscr{H} \;:\; \langle Af_1,\, f_2\rangle = \langle f_1,\, Af_2\rangle \quad \forall (f_1, f_2) \in \mathscr{H}\times\mathscr{H}\}$.

(A3) The set of *states* $\mathscr{S}$ of a quantum system with a Hilbert space $\mathscr{H}$ consists of all trace class operators ρ with $\operatorname{Tr}\rho = 1$. *Pure states* are projection operators onto one-dimensional subspaces of $\mathscr{H}$. For $\psi \in \mathscr{H}$ with $\|\psi\| = 1$, the corresponding projection is denoted ρ_ψ. All other states are called *mixed states*.

(A4) Let $\mathscr{P}(\mathbb{R})$ denote the set of probability measures on $\mathbb{R}$. A *measurement* is a mapping $\mathscr{A}\times\mathscr{S} \ni (A,\rho) \mapsto \mu_A \in \mathscr{P}(\mathbb{R})$, which to every observable $A \in \mathscr{A}$ and state $\rho \in \mathscr{S}$ assigns a probability measure μ_A on $\mathbb{R}$. For every Borel subset $E \subseteq \mathbb{R}$, the quantity $\mu_A(E)$ is the probability that for a quantum system in a state ρ the result of a measurement of the observable A lies in E.

To explicit the expression of the measure μ_A in the last axiom, we need the following definition of a *projection-valued measure* (PVM).

Definition 1.12 – *Let $(X, \mathcal{X})$ be a measured space, $\mathscr{H}$ a Hilbert space and $\mathscr{B}(\mathscr{H})$ be the Banach space of bounded operators on $\mathscr{H}$. A map $\mu : \mathcal{X} \to \mathscr{B}(\mathscr{H})$ is called a projector-valued measure if the following properties are satisfied.*

(1) For each $E \in \mathcal{X}$, $\mu(E)$ is an orthogonal projection.

(2) $\mu(\varnothing) = 0$ and $\mu(X) = \mathrm{I_d}$.

(3) If $E_1, E_2, E_3, \ldots$ in $\mathcal{X}$ are disjoint, then for all $f \in \mathscr{H}$ we have $\mu(\cup_{j=1}^{\infty} E_j)f = \sum_{j=1}^{\infty} \mu(E_j)f$, where the convergence of the series is in the norm topology of $\mathscr{H}$.

(4) For all $E_1, E_2 \in \mathcal{X}$, we have $\mu(E_1 \cap E_2) = \mu(E_1)\mu(E_2)$.

Suppose now, that $f : (X, \mathcal{X}) \to \mathbb{C}$ is an unbounded measurable function, and we wish to define the integral of f with respect to the PVM $\mu : \mathcal{X} \to \mathscr{B}(\mathscr{H})$. First, note that every function $\psi \in \mathscr{H}$ describes a positive, real-valued measure $\mu_\psi : \mathcal{X} \to [0, \infty]$ by setting

$$\mu_\psi(E) := \langle \psi,\, \mu(E)\psi \rangle.$$

Then we have the following proposition, whose prove can be found in Hall (2013, section 10.1).

Proposition 1.3 – *Suppose $\mu : \mathcal{X} \to \mathscr{B}(\mathscr{H})$ is a PVM and $f : X \to \mathbb{C}$ is measurable (not necessarily bounded). Define a subspace D_f of $\mathscr{H}$ by*

$$D_f := \left\{ \psi \in \mathscr{H} \;:\; \int_X |f(\lambda)|^2 \, d\mu_\psi(\lambda) < +\infty \right\}.$$

Then there exists a unique unbounded operator $F : \mathscr{H} \to \mathscr{H}$ with domain D_f with the property that

$$\langle \psi,\, F\psi \rangle = \int_X f(\lambda)\, d\mu_\psi(\lambda)$$

for all $\psi \in D_f$. The operator F is called the integral of f with respect to the PVM μ, which we also write as $F = \int_X f \, d\mu$.

With proposition 1.3 in mind, we can state Von Neumann's spectral theorem for unbounded self-adjoint operators, whose proof can be found in Hall (2013, section 10.1).

Theorem 1.5 – *Suppose A is a self-adjoint operator on $\mathscr{H}$, and let $\sigma(A)$ denote the spectrum of A. Then there is a unique projection-valued measure $\mathtt{P}_A$ on $\sigma(A)$ with values in $\mathscr{B}(\mathscr{H})$ such that*

$$\int_{\sigma(A)} \lambda \, d\mathtt{P}_A(\lambda) = A.$$

Going back to axiom (A4), the probability measure μ_A is given by the *Born-von Neumann formula*

$$\mu_A(E) = \mathrm{Tr}\, \mathtt{P}_A(E)\rho, \quad \forall E \in \mathcal{B}(\mathbb{R}),$$

where $\mathtt{P}_A$ is the unique PVM given by theorem 1.5.

We now state the last axiom of quantum mechanics that is of interest for us (as a matter of completeness, other axioms deal with time evolution of quantum systems, which is of no interest for quantum homodyne tomography).

(A5) Observables $A, B \in \mathscr{A}$ can be measured simultaneously if and only if the corresponding PVM $\mathtt{P}_A$ and $\mathtt{P}_B$ satisfy $\mathtt{P}_A(E_1)\mathtt{P}_B(E_2) = \mathtt{P}_B(E_2)\mathtt{P}_A(E_1)$ for all $E_1, E_2 \in \mathcal{B}(\mathbb{R})$; in that case we say that observables A and B *commute*.

We finish this section with probably the most striking result of quantum mechanics, the so-called *Heisenberg uncertainty relation*, stated here for pure states only. Heisenberg's relation bound the precision at which two observables can be measured simultaneously; in particular, if the two observables are noncommuting, measuring one with precision is always to the detriment of the precision for the other.

Theorem 1.6 – *Let $A, B \in \mathscr{A}$ and ρ_ψ be a pure state with $\psi \in \mathrm{dom}(A) \cap \mathrm{dom}(B)$ and $A\psi, B\psi \in \mathrm{dom}(A) \cap \mathrm{dom}(B)$. Let $\sigma^2_\psi(A)$, $\sigma^2_\psi(B)$ denote respectively the variance of the observable A (respectively B) in the state ρ_ψ. Then, $\sigma^2_\psi(A)\sigma^2_\psi(B) \geq \langle i(AB - BA)\psi,\, \psi\rangle^2/4$.*

Finally, notice that the principles stated here cannot be verified directly. However, whenever the quantum mechanics is applicable, it is continuously verified that the experiments agree with the predicitions of the theory.

1.4.2. Statement of the problem

Here we make a brief statement of the problem of state estimation in Quantum Homodyne Tomography. For more details and motivations, we advice reading Artiles et al. (2005); Butucea et al. (2007).

We consider the quantum system of monochromatic light in a cavity, whose state is described by self-adjoint operator acting on the space of complex-valued square integrable functions on the real line $L^2(\mathbb{R})$.

The observables of interest for this quantum system are the electric and magnetic fields, whose corresponding self-adjoint operators on $L^2(\mathbb{R})$ are given by $\mathbf{Q}$ and, respectively $\mathbf{P}$, with domains $D(\mathbf{Q}) := \{\psi \in L^2(\mathbb{R}) : x \mapsto x\psi(x) \in L^2(\mathbb{R})\}$ and $D(\mathbf{P}) := \{\psi \in L^2(\mathbb{R}) : x \mapsto \psi'(x) \in L^2(\mathbb{R})\}$. The operarors $\mathbf{Q}$ and $\mathbf{P}$ act on $D(\mathbf{Q})$, respectively $D(\mathbf{P})$, as

$$\mathbf{Q}\psi(x) = x\psi(x), \text{ and} \qquad \mathbf{P}\psi(x) = -i\psi'(x).$$

The derivative in the definitions of $D(\mathbf{P})$ and $\mathbf{P}$ is understood in the distributional sense.

The observables $\mathbf{P}, \mathbf{Q}$ are noncommuting, hence from theorem 1.6 we cannot expect to have a joint distribution for the simultaneous measurement of $\mathbf{P}$ and $\mathbf{Q}$ on a system in state ρ. But, the *Wigner transform* of ρ, $W_\rho : \mathbb{R}^2 \to \mathbb{R}$, as defined below, is the closest object of a joint probability density function for the simultaneous measurement of $\mathbf{P}$ and $\mathbf{Q}$ (Folland, 1989). In particular, it is real valued and its marginals are probability density functions, with respect to Lebesgue measure on $\mathbb{R}$. However, W_ρ may contains patches of negative values, thus it cannot be a joint probability distribution in general. The Wigner transform of state ρ is defined by the property that the Fourier transform $\widehat{W}_\rho$ of W_ρ with respect to both variables has the expression

$$\widehat{W}_\rho(x, y) := \mathrm{Tr}(\rho \exp(ix\mathbf{Q} + iy\mathbf{P})).$$

Although we cannot measure simultaneously the observables $\mathbf{P}$ and $\mathbf{Q}$, it is possible to measure the *quadrature observables*, defined as $\mathbf{X}_\theta := \mathbf{Q}\cos\theta + \mathbf{P}\sin\theta$ for all $\theta \in [0, \pi]$. We denote by X^ρ_θ the random variable whose distribution is the measurement of $\mathbf{X}_\theta$ on the quantum system in state ρ. Assuming that θ is drawn uniformly from $[0, \pi]$, the joint probability density function (with respect to the Lebesgue measure on $\mathbb{R} \times [0, \pi]$) for

(X_θ^ρ, θ) is given by the *Radon transform* of the Wigner distribution W_ρ, that is

$$p_\rho(x,\theta) := \frac{1}{\pi}\int_{\mathbb{R}} W_\rho(x\cos\theta - \xi\sin\theta, x\sin\theta + \xi\cos\theta)\, d\xi.$$

Given n quantum systems prepared in the same state ρ, the aim of Quantum Homodyne Tomography is to reconstruct ρ from n independent measurements of the *quadrature observables* $\mathbf{X}_\theta := \mathbf{Q}\cos\theta + \mathbf{P}\sin\theta$, for some values of $\theta \in [0,\pi]$.

As mentioned in Butucea et al. (2007), the experimental setup for Quantum Homodyne Tomography is far from perfect, and we have to consider some noise addition in the statistical model to be realistic. In the same reference, authors justify that a good model is to correct the density p_ψ by taking into account the efficiency $\eta \in (0,1)$ of the detector. The efficiency-corrected probability density is then the convolution

$$p_\rho^\eta(x,y) := \sqrt{\frac{2}{1-\eta}}\int_{\mathbb{R}} p_\rho(x,\theta)\exp\left[-\frac{2\pi\eta}{1-\eta}(x-y)^2\right] dx. \tag{1.6}$$

From a statistical perspective, we have n observations $(Y_1,\theta_1),\dots,(Y_n,\theta_n)$ independently and identically distributed with density $p_\rho^\eta(y,\theta)$ (with respect to Lebesgue measure), where $\eta \in (0,1)$ is assumed to be known. Then the aim is to estimate W_ρ, or better, to estimate ρ.

1.4.3. Frequentist results and minimax study

The problem of QHT is a statistical nonparametric *ill-posed* inverse problem that has been relatively well studied from a frequentist point of view in the last few years, and now quite well understood. We mention here only papers with theoretical analysis of the performance of their estimation procedure, which we summarize; many other physical papers references can be found therein.

We should classify frequentist methods in two categories, depending if they are based on estimating the state ρ, or estimating W_ρ (although $\rho \mapsto W_\rho$ is one-to-one, methods based on estimating W_ρ don't permit to do the reverse path from $W_\rho \mapsto \rho$).

State estimation

The estimation of the state ρ from QHT measurements has been considered in the ideal situation ($\eta = 1$, no noise) by Artiles et al. (2005). To summarize their result, we need to introduce the *Fock basis* $(\varphi_j)_{j\in\mathbb{N}}$ for $L^2(\mathbb{R})$,

$$\varphi_j(x) := \frac{1}{\sqrt{\sqrt{\pi}2^j j!}} H_j(x) e^{-x^2/2},$$

where H_j are the *Hermite polynomials*. Artiles et al. (2005) used that the entries of the state ρ, expressed in the Fock basis, can be rewritten as

$$\rho_{j,k} = \frac{1}{\pi}\int_{-\infty}^{+\infty}\int_0^\pi p_\rho(x,\theta) f_{j,k}(x) e^{-2\pi i(k-j)\theta}\, d\theta dx,$$

where $f_{j,k}$ are known functions called *pattern functions*, which exact expressions can be found in various papers of the quantum homodyne litterature. Then they proposed the estimator $\widetilde{\rho}^{(N,n)}$ of ρ, such that the entries of $\widetilde{\rho}^{(N,n)}$ expressed in the Fock basis are

$$\widetilde{\rho}_{j,k}^{(N,n)} = \begin{cases} n^{-1}\sum_{l=1}^{n} f_{j,k}(Y_l)e^{-2\pi i(k-j)\theta_l} & \text{if } 0 \le j,k \le N-1 \\ 0 & \text{if } \max(j,k) \ge N. \end{cases}$$

The main result of Artiles et al. (2005) state that $\mathbb{E}[\mathrm{Tr}\,|\widetilde{\rho}^{(N,n)} - \rho|^2] \to 0$ as $n \to \infty$, provided that $N \to \infty$ with $N = o(n^{3/7})$.

The case where $\eta \neq 1$ is investigated in Aubry et al. (2008) under Frobenius-norm risk. They used a different estimation method depending on whether $0 < \eta \le 1/2$ or $1/2 < \eta \le 1$. We detail only the case $1/2 < \eta \le 1$ here. They considered the estimator,

$$\widetilde{\rho}_{j,k}^{N,\eta} = \begin{cases} n^{-1}\sum_{l=1}^{n} f_{j,k}^{\eta}(Y_l)e^{-2\pi i(k-j)\theta_l} & \text{if } 0 \le j+k \le N-1 \\ 0 & \text{otherwise.} \end{cases} \tag{1.7}$$

where $f_{j,k}^{\eta}$ have Fourier transform $\widehat{f}_{j,k}^{\eta}$ satisfying $\widehat{f}_{j,k}^{\eta}(x) = \widehat{f}_{j,k}(x)e^{-\gamma x^2}$, with $\gamma := \pi(1-\eta)/(2\eta)$ and $f_{j,k}$ the same pattern functions as previously. Then for the class of states

$$\mathcal{R}(B,r,L) := \Big\{\rho \text{ quantum state} : |\rho_{j,k}| \le L\exp\Big(-B(j+k)^{r/2}\Big)\Big\},$$

where the entries $\rho_{j,k}$ are expressed in the Fock basis, they proved that for a constant $C_0 > 0$,

$$\limsup_{n\to\infty} \sup_{\rho\in\mathcal{R}(B,r,L)} \varphi_n^{-2}\mathbb{E}\left(\sum_{j,k}|\widetilde{\rho}_{j,k}^{N,\eta} - \rho_{jk}|^2\right) \le C_0 \tag{1.8}$$

where

- $\varphi_n^2 = (\log n)^{(12\gamma+B)/(12\gamma+3B)}n^{-B/(4\gamma+B)}$, if $r = 2$ and $N = \frac{\log n}{8\gamma+2B}(1 + \frac{2}{3}\frac{\log\log n}{\log n})$.
- $\varphi_n^2 = (\log n)^{2-r/3}e^{-2BN^{r/2}}$, if $0 < r < 2$ and N solution of $8\gamma N + 2BN^{r/2} = \log n$.

Clearly the estimator in equation (1.7) is not rate adaptive over $\mathcal{R}(B,r,L)$. Alquier et al. (2013) tackled the problem of adaptivity over $\mathcal{R}(B,r,L)$ using a *soft-thresholded* version of equation (1.7). For a prescribed tolerance level $\epsilon \in (0,1)$, they estimated the state ρ by $\widetilde{\rho}_{\text{thres}}^{\eta}$, where the entries of $\widetilde{\rho}_{\text{thres}}^{\eta}$ in the Fock basis are

$$\widetilde{\rho}_{\text{thres}}^{\eta}(j,k) = \frac{\widetilde{\rho}_{j,k}^{\eta}}{|\widetilde{\rho}_{j,k}^{\eta}|}(|\widetilde{\rho}_{j,k}^{\eta} - t_{j,k}|)_+, \qquad t_{j,k} = 2\|f_{j,k}^{\eta}\|_\infty\sqrt{\frac{\log\left(\frac{2N(N+1)}{\epsilon}\right)}{n}},$$

and $\widetilde{\rho}_{j,k}^{\eta}$ are computed as in equation (1.7). They found that there estimator is rate-adaptive over all parameters $L \ge 1$, $B \ge B_0$, $r \ge r_0$, where B_0, r_0 can be chosen arbitrary small. They obtained up to additional power of $\log n$ factors, the same upper bound on the rates as in equation (1.8)

Finally, let mention the the problem of goodness-of-fit testing for quantum state estimation is investigated in Méziani (2008).

Wigner density estimation

Regarding frequentist methods for estimating W_ρ, the first result goes back to Guţă and Artiles (2007), where sharp minimax results for the pointwise risk are given over a class of ultra-smooth Wigner functions

$$\mathcal{A}(\beta, r, L) := \left\{ W_\rho \,:\, \int |\widehat{W}_\rho(z)|^2 e^{2\beta\|z\|^r}\, dz \leq L^2 \right\},$$

where $\widehat{W}_\rho$ is the Fourier transform of W_ρ with respect to both variables. They proposed a kernel-type estimator based on the following band-limiter filter

$$K_\delta(x) := \frac{1}{2}\int_{-\delta}^{\delta} r e^{2\pi i r x}\, dr. \tag{1.9}$$

Their kernel estimator of W_ρ based on the observations $(Y_1, \theta_n), \ldots, (Y_n, \theta_n)$ is

$$\widetilde{W}_\rho(x, \omega) := \frac{1}{n}\sum_{l=1}^{n} K_{\delta_n}(x\cos\theta_l + \omega \sin\theta_l - Y_l). \tag{1.10}$$

Then they found that the pointwise risk over $\mathcal{A}(\beta, 1, L)$ when $\delta_n = 2\beta/\log n$ satisfy the bound $\mathbb{E}[(\widetilde{W}_\rho(z) - W_\rho(z))^2] \leq C(z) n^{-1} (\log n)^3 (1 + o(1))$, where the constant $C(z) > 0$ is made precise in the original paper. More precisely, Guţă and Artiles (2007) provided a lower bound for the pointwise risk over a large subset of $\mathcal{A}(\beta, 1, L)$ that matches the previous upper bound, showing that the constant $C(z)$ computed in their paper is sharp. As far as we are not concerned with the case of ideal detection in this manuscript, we will not detail more this result.

The noisy framework has been considered in Butucea et al. (2007); authors proposed to modify the kernel of equation (1.9) to take into account the noise in the observations. They introduced instead[3]

$$K_h^\eta(x) = \frac{1}{2}\int_{-1/h}^{1/h} \frac{\exp(-2\pi i x t)|t|}{\exp(-\gamma t^2)}\, dt,$$

where $\gamma := \pi(1-\eta)/(2\eta)$. Their estimator is very similar to equation (1.10),

$$\widetilde{W}_\rho^\eta(x, \omega) := \frac{1}{n}\sum_{l=1}^{n} K_h^\eta(x\cos\theta_l + \omega\sin\theta_l - Y_l). \tag{1.11}$$

They proved that if h is solution of $2\beta/h^r + 2\gamma/h^2 = \log n$, then

$$\limsup_{n\to\infty} \sup_{z\in\mathbb{R}^2} \sup_{W_\rho \in \mathcal{A}(\beta,r,L)} \mathbb{E}[|\widetilde{W}_h^\eta(z) - W_\rho(z)|^2]\varphi_n^{-2} \leq C_0,$$

where

- $C_0 = 1$ and $\varphi_n^2 = \frac{\sqrt{L}h^{r-1}}{2\beta r}\exp\left(-\frac{2\beta}{h^r}\right)$ if $0 < r < 2$.
- $C_0 > 0$ and $\varphi_n^2 = n^{-\beta/(\beta+\gamma)}$ if $r = 2$.

[3] In the original paper, authors consider that the observations are $(\sqrt{\eta}Y_i, \theta_i)$ instead of (Y_i, θ_i) in this manuscript, leading to a slightly different definition for the kernel.

Furthermore, they also proved that these rates are minimax efficient for $0 < r < 2$ and nearly minimax for $r = 2$, as they found the lower bounds

$$\liminf_{n\to\infty} \inf_{\widetilde{W}_\rho} \sup_{W_\rho \in \mathcal{A}(\beta,r,L)} \mathbb{E}[|\widetilde{W}_\rho(z) - W_\rho(z)|^2]\varphi_n^{-2} \geq 1, \quad \forall z \in \mathbb{R}^2 \text{ if } 0 < r < 2$$

$$\liminf_{n\to\infty} \inf_{\widetilde{W}_\rho} \sup_{W_\rho \in \mathcal{A}(\beta,r,L)} \mathbb{E}[|\widetilde{W}_\rho(z) - W_\rho(z)|^2](n \log n)^{\beta/(\beta+\gamma)} > 0, \quad \forall z \in \mathbb{R}^2 \text{ if } r = 2,$$

where the infinimum is taken over all possible estimators $\widetilde{W}_\rho$ of W_ρ. Finally, for $0 < r < 1$, Butucea et al. (2007) established a bandwith h_{ad} for which the estimator (1.11) is adaptive over all $\beta > 0$, $0 < r < 1$ and $L > 0$.

In, Aubry et al. (2008), authors showed that the class $\mathcal{R}(B, r, L)$ defined in the previous section relates to $\mathcal{A}(\beta, r, L)$ in the following way. If $0 < r < 2$ and $\beta < 2^{-r}B$, then $\mathcal{R}(B, r, L) \subset \mathcal{A}(\beta, r, L')$ for a suitable constant $L' > 0$. For $r = 2$, the result is a little bit different, if $4\beta < B/(1 + \sqrt{B})^2$, then $\mathcal{R}(B, r, L) \subset \mathcal{A}(\beta, r, L'')$ for a suitable $L'' > 0$. However, unlike Wigner functions in $\mathcal{A}(\beta, r, L)$, the Wigner functions for state in $\mathcal{R}(B, r, L)$ have very fast decay, which allowed Aubry et al. (2008) to derive an upper bound for the quadratic risk for estimating W_ρ, $\rho \in \mathcal{R}(B, r, L)$, using a truncated version of the estimator (1.11). For a sequence $s_n \to \infty$, their estimator is

$$\widetilde{W}_h^{\eta,*} := \widetilde{W}_h^{\eta}(z)\mathbb{1}(\|z\| \leq s_n).$$

Then they proved that for a constant $C_0 > 0$,

$$\limsup_{n\to\infty} \sup_{\rho \in \mathcal{R}(B,r,L)} \mathbb{E}\left[\|\widetilde{W}_h^{\eta,*} - W_\rho\|_2^2\right] \varphi_n^{-2} \leq C_0,$$

where

- $\varphi_n^2 = h^{3r-10} \exp(-2\beta(2h)^{-r})$ if $0 < r < 2$, $\beta < B$, $s_n = 1/h$ and h solution of the equation $2\beta(2h)^{-r} + 2\gamma h^{-2} = \log n - (\log\log n)^2$.
- $\varphi_n^2 = (\log n)^{(16\gamma+3\beta)/(8\gamma+2\beta)} n^{-\beta/(4\gamma+\beta)}$ if $r = 2$, $\beta = B/(1 + \sqrt{B})^2$, $s_n = 1/h$ and $h^2 = \frac{2}{4\gamma+\beta} \log n + \frac{1}{4\gamma+\beta} \log\log n$.

More recently, Lounici et al. (2015) established the first sup-norm risk upper bound over $\mathcal{A}(\beta, r, L)$ with the estimator (1.11), as well as the first minimax lower bounds for both sup-norm and L^2-norm risk over $\mathcal{A}(\beta, 2, L)$; showing in particular that (1.11) achieve optimal rates over $\mathcal{A}(\beta, 2, L)$ for both the sup-norm and L^2-risk. Moreover, they also made the estimator (1.11) adaptive over all $\beta, L > 0$, $r \in (0, 2]$, using a Lepski type procedure (Lepski and Spokoiny, 1997). As their paper is still under revision, we should not detail more their results.

1.4.4. Bayesian nonparametric approach

To our knowledge, no Bayesian nonparametric method has been proposed to address the problem of QHT with noisy data, a gap that we try to fill with in this thesis. In particular, we propose two families of prior distributions over pure states that can be useful in practice, namely *mixtures of coherent-states* and *random Wilson series*. Recalling that pure states are projection operator onto one-dimensional subspaces of $L^2(\mathbb{R})$, the problem of estimating a pure state reduce to estimating a function $\psi \in L^2(\mathbb{R})$ with

$\|\psi\|_2 = 1$, called the *wave function*, from noisy QHT measurements $(Y_1, \theta_1), \ldots, (Y_n, \theta_n)$. We will discuss in chapter 4 how the method can be extended to handle mixed-states estimation.

The mixture of coherent states prior, as described in the next section, is motivated by Wolpert et al. (2002), where authors proposed to use mixture priors to address general inverse problems. Although we implement successfully mixtures of coherent states in practice, the posterior asymptotic, however, is still challenging today (this will be discussed more thoroughly at the end of chapter 4). Nevertheless, random Wilson series (section 1.4.4) may also be useful priors in practice, and an asymptotic study of this family of prior is made in chapter 4.

Coherent states mixtures

In quantum optics, a coherent state refers to a state of the quantized electromagnetic field that describes a classical kind of behavior.

Let $T_x f(y) := f(y-x)$, $M_\omega f(y) = e^{2\pi i \omega y} f(y)$, denote the translation and modulation operators, respectively, and g a window function with $\|g\|_2 = 1$; most of time g is chosen as $g(x) = 2^{-1/4} \exp(-\pi x^2)$. Mathematically speaking, coherent states are pure states ρ_ψ, that is projection operators onto $\psi \in L^2(\mathbb{R})$, described by a *wave-function* ψ belonging to

$$\left\{ \psi \in L^2(\mathbb{R}) \,:\, \psi = T_x M_\omega g \quad (x, \omega) \in \mathbb{R}^2 \right\}.$$

We suggest a mixture of coherent states as prior distribution on the wave function ψ. For a random measure Q on $\mathbb{R}^2 \times [0, 2\pi]$, our model may be summarized by the following hierarchical representation. P_ψ^η denote the probability distribution having the density of equation (1.6), with $\rho = \rho_\psi$.

$$\begin{aligned} (Y_1, \theta_1), \ldots, (Y_n, \theta_n) &\overset{\text{i.i.d}}{\sim} P_\psi^\eta, \quad \text{with } \psi = \widetilde{\psi}/\|\widetilde{\psi}\|_2 \\ \widetilde{\psi}(z) &= \int_{\mathbb{R}^2 \times [0,2\pi]} e^{i\phi} T_x M_\omega g(z)\, Q(dx d\omega d\phi) \\ Q &\sim \Pi. \end{aligned}$$

Although this prior is appealing, in particular because all the states prepared in the lab have a very sparse representations in terms of coherent states, we are unable at this time to derive posterior contraction rates. However, simulation results given in chapter 4 show that these priors work well in practice.

Random Wilson series

For $a, b > 0$, $g \in L^2(\mathbb{R})$, the set of time-frequency shifts $\mathcal{G}(g; a, b) = \{M_n T_m g \,:\, (n, m) \in a\mathbb{Z} \times b\mathbb{Z}\}$ is a *Gabor frame* if there exist positive constants $A, B > 0$, called the frame bounds, such that

$$A\|f\|_2^2 \leq \sum_{m,n} |\langle f,\, T_m M_n g \rangle|^2 \leq B\|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}).$$

The frame operator $S_g f := \sum_{m,n} \langle f,\, T_m M_n g \rangle\, T_m M_n g$ is a bounded, positive and invertible mapping of $L^2(\mathbb{R})$ onto itself. Consequently, the *canonical dual window* $h := S^{-1} g$

is also in $L^2(\mathbb{R})$. Moreover, from Gröchenig (2001, proposition 5.2.1) the *canonical dual frame* of $\mathcal{G}(g;a,b)$ is $\mathcal{G}(h;a,b)$ with frame bounds B^{-1}, A^{-1}, and for every $f \in L^2(\mathbb{R})$

$$f = \sum_{m,n} \langle f,\, T_m M_n g \rangle\, T_m M_n h = \sum_{m,n} \langle f,\, T_m M_n h \rangle\, T_m M_n g,$$

with unconditional convergence of the series in $L^2(\mathbb{R})$. Gabor expansions are not unique, which is appealing from a practical perspective, but appeared to be really annoying in establishing theoretical result. However, based on an idea of Wilson (1987), Daubechies et al. (1991) proposed a modification of Gabor systems for which the redundancy is removed, but good time-frequency localization properties are preserved. More precisely, they constructed a real-valued function φ such that for some $a, b > 0$,

$$|\varphi(x)| \lesssim e^{-a|x|}, \qquad |\widehat{\varphi}(\omega)| \lesssim e^{-b|\omega|},$$

and such that the φ_{lm}, $l \in \mathbb{N}$, $m \in \frac{1}{2}\mathbb{Z}$ defined by

$$\varphi_{lm}(x) := \begin{cases} \varphi(x-2m) & \text{if } l = 0, \\ \sqrt{2}\varphi(x-m)\cos(2\pi l x) & \text{if } l \neq 0 \text{ and } 2m+l \text{ is even}, \\ \sqrt{2}\varphi(x-m)\sin(2\pi l x) & \text{if } l \neq 0 \text{ and } 2m+l \text{ is odd}, \end{cases}$$

constitute an orthonormal base for $L^2(\mathbb{R})$, namely the *Wilson bases of exponential decay*. Following Gröchenig (2001, section 8.5), we may rewrite φ_{lm} in a convenient form for the sequel, emphasizing the relationship with coherent states and Gabor frames,

$$\varphi_{lm} = c_l T_m (M_l + (-1)^{2m+l} M_{-l})\varphi, \qquad (l,m) \in \mathbb{N} \times \tfrac{1}{2}\mathbb{Z},$$

where $c_0 := 1/2$ and $c_l := 1/\sqrt{2}$ for $l \geq 1$.

We then consider the following prior distribution Π on the wave function ψ. Let (φ_{lm}) be the previously defined orthonormal Wilson base with exponential decay. For any positive number Z, let Λ_Z be the spherical array

$$\Lambda_Z := \Big\{ (l,m) \in \mathbb{N} \times \tfrac{1}{2}\mathbb{Z} \,:\, l^2 + m^2 < Z^2 \Big\}.$$

Also define the simplex Δ_Z in the ℓ_2 metric as

$$\Delta_Z := \Big\{ \mathbf{p} = (p_{lm})_{(l,m)\in\Lambda_Z} \,:\, \textstyle\sum_{(l,m)\in\Lambda_Z} p_{lm}^2 = 1,\ p_{lm} \geq 0 \Big\}.$$

Let P_Z be a distribution over $\mathbb{R}^+$ and draw $Z \sim P_Z$. Given Z, draw $\mathbf{p}$ from a distribution $G(\cdot \mid Z)$ over the simplex Δ_Z. Independently of $\mathbf{p}$, draw $\zeta = (\zeta_{lm})_{(l,m)\in\Lambda_Z}$ from a distribution $P_\zeta(\cdot \mid Z)$ over $[0, 2\pi]^{|\Lambda_Z|}$ and set

$$\psi := \sum_{(l,m)\in\Lambda_Z} p_{lm} e^{i\zeta_{lm}}\, \varphi_{lm}.$$

Note that (φ_{lm}) is orthonormal, thus $\|\psi\|_2^2 = \sum_{(l,m)\in\Lambda_Z} p_{lm}^2 = 1$ almost-surely.

This family of prior distribution may be seen as a discretized version of coherent states mixtures. However, unlike coherent states, Wilson bases have nice analytic properties that allow to challenge the posterior asymptotic in the model of QHT with noisy data.

1.4.5. Contribution to QHT

In addition to introduce the two Bayesian nonparametric approaches to the QHT problem described in the section 1.4.4, in chapter 4 we focus on evaluating posterior rates of convergence for random Wilson series distribution as prior over the set of pure states — equivalently the unit sphere of $L^2(\mathbb{R})$ — in the statistical model of equation (1.6).

To establish posterior rates of convergence, we need a suitable smoothness class. We propose to use closed balls in *ultra-modulation spaces*; which we describe here. To this aim, we need the following ingredients: the short-time Fourier transform, a class of windows and a class of weights. For a non-zero window function $g \in L^2(\mathbb{R})$, the short-time Fourier transform of a function $f \in L^2(\mathbb{R})$ with respect to the window g is given by

$$V_g f(x,\omega) := \langle f,\, M_\omega T_x g\rangle = \int_{\mathbb{R}} f(t)\overline{g(t-x)}e^{-2\pi i\omega t}\,dt.$$

We also need a class of analyzing windows g with sufficiently good time-frequency localization properties. We use the *Gelfand-Shilov* space $\mathcal{S}_1^1(\mathbb{R})$. A function $f : \mathbb{R}^d \to \mathbb{C}$ belongs to the Gelfand-Shilov space $\mathcal{S}_1^1(\mathbb{R}^d)$ if $f \in \mathcal{C}^\infty(\mathbb{R}^d)$ and there exist real constants $h > 0$ and $k > 0$ such that

$$\sup_{x\in\mathbb{R}^d} |f(x)e^{h\|x\|}| < +\infty, \qquad \sup_{\omega\in\mathbb{R}^d} |\widehat{f}(\omega)e^{k\|\omega\|}| < +\infty.$$

Next, for $\beta > 0$, $g \in \mathcal{S}_1^1(\mathbb{R})$, and $r \in [0,1)$, we consider the exponential weights on $\mathbb{R}^2$ defined by $x \mapsto \exp(\beta\|x\|^r)$, and we introduce the class of wave-functions

$$\mathcal{C}_g(\beta,r,L) := \left\{\psi \in L^2(\mathbb{R}) \,:\, \int_{\mathbb{R}^2} |V_g\psi(z)|\,\exp(\beta\|z\|^r)dz \le L, \quad \|\psi\|_2 = 1\right\}.$$

It is shown in chapter 4 that $\mathcal{C}_g(\beta,r,L)$ is independent of g in the sense that two functions $g_1, g_2 \in \mathcal{S}_1^1(\mathbb{R})$ lead essentially to the same classes of functions.

Let Π denote the random Wilson series distribution over the unit sphere of $L^2(\mathbb{R})$. Assuming $\psi_0 \in \mathcal{C}_g(\beta,r,L)$, our main theorem state sufficient conditions on P_Z, $G(\cdot \mid Z)$ and $P_\zeta(\cdot \mid Z)$ in the definition of Π to have $M > 0$ such that

$$P_{\psi_0}^{\eta,n}\Pi(\|\psi-\psi_0\|_2 \ge M\epsilon_n \mid (Y_1,\theta_1),\dots,(Y_n,\theta_n)) \to 0,$$

where, for a constant $a \ge 0$ depending only on the prior and $\gamma := \pi(1-\eta)/(2\eta)$,

$$\epsilon_n^2 = (\log n)^{2a}\exp\left\{-\beta\left(\tfrac{\log n}{2\gamma}\right)^{r/2} + O(1)\right\}. \tag{1.12}$$

It is not clear whether or not we can chose P_Z, $G(\cdot \mid Z)$ and $P_\zeta(\cdot \mid Z)$ to make these rates adaptive over $(\beta,r,L) \in (0,\infty)\times[0,1)\times(0,\infty)$. However, we construct in chapter 4 non adaptive examples where equation (1.12) is achieved with $a = 2$.

Bibliography

F. Abramovich, T. Sapatinas, and B. Silverman. Stochastic expansions in an overcomplete wavelet dictionary. *Probability Theory and Related Fields*, 117(1):133–144, 2000.

P. Alquier, K. Meziani, and G. Peyré. Adaptive estimation of the density matrix in quantum homodyne tomography with noisy data. *Inverse Problems*, 29(7):075017, 2013.

C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

L. Artiles, R. Gill, et al. An invitation to quantum tomography. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):109–134, 2005.

J.-M. Aubry, C. Butucea, and K. Méziani. State estimation in quantum homodyne tomography with noisy data. *Inverse Problems*, 25(1):015003, 2008.

O. E. Barndorff-Nielsen and J. Schmiegel. Lévy-based spatial-temporal modelling, with applications to turbulence. *Russian Mathematical Surveys*, 59(1):65, 2004.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

J. M. Bernardo and A. F. Smith. Bayesian theory, 2001.

L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques*, volume 42, pages 273–325, 2006.

D. Blackwell and J. B. MacQueen. Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355, 1973.

N. Bochkina and J. Rousseau. Adaptive density estimation based on a mixture of Gammas. *ArXiv e-prints*, May 2016.

C. Butucea, M. Guţă, and L. Artiles. Minimax and adaptive estimation of the wigner function in quantum homodyne tomography with noisy data. *The Annals of Statistics*, 35(2):465–494, 2007.

A. Canale and P. De Blasi. Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404, 02 2017. doi: 10.3150/15-BEJ746. URL http://dx.doi.org/10.3150/15-BEJ746.

O. Cappé, C. P. Robert, and T. Rydén. Reversible jump, birth-and-death and more general continuous time markov chain monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):679–700, 2003.

I. Castillo. Lower bounds for posterior rates with gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

E. Çınlar. *Probability and stochastics*, volume 261. Springer Science & Business Media, 2011.

C. Cohen-Tannoudji, B. Diu, F. Laloe, and B. Dui. *Quantum Mechanics (2 vol. set)*. Wiley-Interscience, Oct. 2006. ISBN 0471569526. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471569526.

I. Daubechies, S. Jaffard, and J.-L. Journé. A simple wilson orthonormal basis with exponential decay. *SIAM Journal on Mathematical Analysis*, 22(2):554–573, 1991.

R. De Jonge and J. Van Zanten. Adaptive nonparametric bayesian inference using location-scale mixture priors. *The Annals of Statistics*, 38(6):3300–3320, 2010.

P. Diaconis and D. Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

J. L. Doob. Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27, 1949.

R. M. Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

T. Erhardsson. Non-parametric bayesian inference for integrals with respect to an unknown finite measure. *Scandinavian Journal of Statistics*, 35(2):369–384, 2008.

M. D. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

M. D. Fall. *Modélisation stochastique de processus pharmaco-cinétiques, application à la reconstruction tomographique par émission de positrons (TEP) spatio-temporelle*. PhD thesis, Paris 11, 2012.

S. Favaro, A. Guglielmi, and S. G. Walker. A class of measure-valued markov chains and bayesian nonparametrics. *Bernoulli*, 18(3):1002–1030, 08 2012. doi: 10.3150/11-BEJ356. URL http://dx.doi.org/10.3150/11-BEJ356.

P. D. Feigin and R. L. Tweedie. Linear functionals and markov chains associated with dirichlet processes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 105, pages 579–585. Cambridge Univ Press, 1989.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

G. B. Folland. *Harmonic analysis in phase space*. Number 122. Princeton university press, 1989.

D. A. Freedman. On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403, 1963.

C. R. Genovese and L. Wasserman. Rates of convergence for the gaussian mixture sieve. *Annals of Statistics*, pages 1105–1127, 2000.

S. Ghosal. Convergence rates for density estimation with bernstein polynomials. *The Annals of Statistics*, 29(5):1264–1280, 2001.

S. Ghosal and A. W. Van Der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, pages 1233–1263, 2001.

S. Ghosal and A. W. Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007a.

S. Ghosal and A. W. Van Der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007b.

S. Ghosal, J. K. Ghosh, and R. Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *Ann. Statist*, 27(1):143–158, 1999.

S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

J. K. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer, 2003.

E. Giné and R. Nickl. Rates of contraction for posterior distributions in l r-metrics, 1 r. *The Annals of Statistics*, pages 2883–2911, 2011.

A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on r^ d. *Probability Theory and Related Fields*, 159(3-4):479–543, 2014.

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

K. Gröchenig. Foundations of time-frequency analysis, 2001.

M. Guţă and L. Artiles. Minimax estimation of the wigner function in quantum homodyne tomography with ideal detectors. *Mathematical Methods of Statistics*, 16(1):1–15, 2007.

B. C. Hall. *Quantum theory for mathematicians*. Springer, 2013.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 2011.

A. Juditsky and S. Lambert-Lacroix. On minimax density estimation on $\mathbb{R}$. *Bernoulli*, 10(2):187–220, 04 2004. doi: 10.3150/bj/1082380217. URL http://dx.doi.org/10.3150/bj/1082380217.

O. Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and computing*, 21(1):93–105, 2011.

J. F. C. Kingman. *Poisson processes*, volume 3. Clarendon Press, 1992.

W. Kruijer and A. van der Vaart. Posterior convergence rates for dirichlet mixtures of beta densities. *Journal of Statistical Planning and Inference*, 138(7):1981 – 1992, 2008. ISSN 0378-3758. doi: http://dx.doi.org/10.1016/j.jspi.2007.07.012. URL http://www.sciencedirect.com/science/article/pii/S0378375807003321.

W. Kruijer, J. Rousseau, and A. Van Der Vaart. Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 1986.

L. Le Cam and G. L. Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, 25(6):2512–2546, 12 1997. doi: 10.1214/aos/1030741083. URL http://dx.doi.org/10.1214/aos/1030741083.

A. Lijoi and I. Prünster. Models beyond the dirichlet process. *Bayesian nonparametrics*, 28:80, 2010.

K. Lounici, K. Meziani, and G. Peyré. Minimax and adaptive estimation of the wigner function in quantum homodyne tomography with noisy data. *arXiv preprint arXiv:1506.06941*, 2015.

S. N. MacEachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741, 1994.

S. N. MacEachern and P. Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

K. Méziani. Nonparametric goodness-of fit testing in quantum homodyne tomography with noisy data. *Electronic journal of statistics*, 2:1195–1223, 2008.

R. M. Neal. Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer, 1992.

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

O. Papaspiliopoulos and G. O. Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

S. Petrone and L. Wasserman. Consistency of bernstein polynomial posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1):79–100, 2002.

B. S. Rajput and J. Rosinski. Spectral representations of infinitely divisible processes. *Probability Theory and Related Fields*, 82(3):451–487, 1989.

C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Adaptive density estimation: a curse of support? *Journal of Statistical Planning and Inference*, 141(1):115–139, 2011.

C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

J. Rosiński. Series representations of lévy processes from the perspective of point processes. In *Lévy processes*, pages 401–415. Springer, 2001.

J. Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics*, 38(1): 146–180, 2010.

L. Schwartz. On bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26, 1965.

C. Scricciolo. Posterior rates of convergence for dirichlet mixtures of exponential power densities. *Electron. J. Statist.*, 5:270–308, 2011. doi: 10.1214/11-EJS604. URL http://dx.doi.org/10.1214/11-EJS604.

C. Scricciolo. Adaptive bayesian density estimation in l^p-metrics with pitman-yor or normalized inverse-gaussian process kernel mixtures. *Bayesian Analysis*, 9(2):475–520, 2014.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

W. Shen, S. T. Tokdar, and S. Ghosal. Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

L. A. Takhtadzhian. *Quantum mechanics for mathematicians*, volume 95. American Mathematical Soc., 2008.

S. T. Tokdar. Posterior consistency of dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pages 90–110, 2006.

N. Tsilevich, A. Vershik, and M. Yor. Distinguished properties of the gamma process and related topics. *arXiv preprint math/0005287*, 2000.

A. B. Tsybakov. Introduction to nonparametric estimation, 2009.

C. Tu. *Nonparametric modelling using Lévy process priors with applications for function estimation, time series modeling and spatio-temporal modeling.* PhD thesis, Ph. D. Thesis, ISDS, Duke University, USA, 2006.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

S. Walker and P. Damien. Sampling methods for bayesian nonparametric inference involving stochastic processes. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 243–254. Springer, 1998.

S. G. Walker. Sampling the dirichlet mixture model with slices. *Communications in Statistics, Simulation and Computation*, 36(1):45–54, 2007.

M. West and M. D. Escobar. *Hierarchical priors and mixture models, with application in regression and density estimation.* Institute of Statistics and Decision Sciences, Duke University, 1993.

K. G. Wilson. Generalized wannier functions. 1987.

R. Wolpert, K. Ickstadt, and M. B. Hansen. A nonparametric bayesian approach to inverse problems. In *Bayesian Statistics 7*. Oxford University Press, 2002.

R. L. Wolpert, M. A. Clyde, and C. Tu. Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *The Annals of Statistics*, pages 1916–1962, 2011.

CHAPTER 2

Some aspects of symmetric Gamma process mixtures

Résumé

Dans cet article, nous présentons quelques aspects spécifiques aux mélanges par processus Gamma symétriques pour un usage en régression. Nous proposons un nouvel échantilloneur de Gibbs pour simuler la distribution a posteriori et nous établissons des vitesses de convergence adaptatives pour le modèle de régression moyenne avec bruit Gaussien.

Abstract

In this article, we present some specific aspects of symmetric Gamma process mixtures for use in regression models. We propose a new Gibbs sampler for simulating the posterior and we establish adaptive posterior rates of convergence related to the Gaussian mean regression problem.

2.1. Introduction

Recently, interest in a Bayesian nonparametric approach to the sparse regression problem based on mixtures emerged from works of Abramovich et al. (2000), de Jonge and van Zanten (2010) and Wolpert et al. (2011). The idea is to model the regression function as

$$f(\cdot) = \int_{\mathcal{X}} K(x;\cdot)Q(dx), \quad Q \sim \Pi_*, \tag{2.1}$$

where $K : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$ is a jointly measurable kernel function, and Π_* a prior distribution on the space of signed measure over the measurable space $\mathcal{X}$. Although the model (2.1) is popular in density estimation Escobar and West (1994); Müller et al. (1996); Ghosal and van der Vaart (2007a); Shen et al. (2013); Canale and De Blasi (2017) and for modeling hazard rates in Bayesian nonparametric survival analysis Lo and Weng (1989); Peccati and Prünster (2008); De Blasi et al. (2009); Ishwaran and James (2012); Lijoi and Nipoti (2014), it seems that much less interest has been shown in regression.

Perhaps the little interest for mixture models in regression is due to the lack of variety in the choice of algorithms available, and in the insufficiency of theoretical posterior contraction results. To our knowledge, the sole algorithm existing for posterior simulations is to be found in Wolpert et al. (2011), when the mixing measure Q is a *Lévy process*.

On the other hand, The only contraction result available is to be found in de Jonge and van Zanten (2010) for a suitable semiparametric mixing measure.

Indeed, both designing an algorithm or establishing posterior contraction results heavily depends on the choice of K and Π_* in equation (2.1); but above all also on the observation model we consider. This last point makes the study of mixtures in regression nasty to handle because of the diversity of observation models possible. In this article, we focus on the situation when Q is a *symmetric Gamma process* to propose both a new algorithm for posterior simulations and posterior contraction rates results.

In the first part of the paper, we propose a Gibbs sampler to get samples from the posterior distribution of symmetric Gamma process mixtures. The algorithm is sufficiently general to be used in all observation models for which the likelihood function is available. We begin with some preliminary theoretical result about approximating symmetric Gamma process mixtures, before stating the general algorithm. Finally, we make an empirical study of the algorithm, with comparison with the RJMCMC algorithm of Wolpert et al. (2011).

The second part of the paper is devoted to posterior contraction rates results. We consider the mean regression model with normal errors of unknown variance, and two types of mixture priors: location-scale and location-modulation. The latter has never been studied previously, mainly because it is irrelevant in density estimation models. However, we show here that it allows to get better rates of convergence than location-scale mixtures, and thus might be interesting to consider in regression.

2.2. Symmetric Gamma process mixtures

Let $(\Omega, \mathcal{E}, \mathbb{P})$ be a probability space and $(\mathcal{X}, \mathcal{A})$ be a measurable space. We call a mapping $Q : \Omega \times \mathcal{A} \to \mathbb{R} \cup \{\pm\infty\}$ a *signed random measure* if $\omega \mapsto Q(\omega, A)$ is a random variable for each $A \in \mathcal{A}$ and if $A \mapsto Q(\omega, A)$ is a signed measure for each $\omega \in \Omega$.

Symmetric Gamma random measures are infinitely divisible and independently scattered random measures (the terminology Lévy base is also used in Barndorff-Nielsen and Schmiegel (2004), and Lévy random measure in Wolpert et al. (2011)), that is, random measures with the property that for each disjoint $A_1, \dots, A_k \in \mathcal{A}$, the random variables $Q(A_1), \dots, Q(A_k)$ are independent with infinitely divisible distribution. More precisely, given $\alpha, \eta > 0$ and F a probability measure on $\mathcal{X}$, a symmetric Gamma random measure assigns to all measurable set $A \in \mathcal{A}$ random variables with distribution $\mathrm{SGa}(\alpha F(A), \eta)$ (see section 2.A). Existence and uniqueness of symmetric Gamma random measures is stated in Rajput and Rosinski (1989).

In the sequel, we shall always denote by Π_* the distribution of a symmetric Gamma random measure with parameters α, η and F, and we refer αF as the base distribution of $Q \sim \Pi_*$, and η as the scale parameter.

2.2.1. Location-scale mixtures

Given a measurable *mother* function $g : \mathbb{R}^d \to \mathbb{R}$, we define the *location-scale* kernel $K_A(x) := g(A^{-1}x)$, for all $x \in \mathbb{R}^d$ and all $A \in \mathcal{E}$, where $\mathcal{E}$ denote the set of all $d \times d$ positive definite real matrices. Then we consider symmetric Gamma location-scale mixtures of

the type

$$f(x;\omega) := \int_{\mathcal{E}\times\mathbb{R}^d} K_A(x-\mu)\, Q(dAd\mu;\omega), \qquad \forall x \in \mathbb{R}^d, \tag{2.2}$$

where $Q : \mathcal{B}(\mathcal{E}\times\mathbb{R}^d)\times\Omega \to [-\infty,\infty]$ is a symmetric Gamma random measure with base measure αF on $\mathcal{E}\times\mathbb{R}^d$, and scale parameter $\eta > 0$. The precise meaning of the integral in equation (2.2) is made clear in Rajput and Rosinski (1989).

2.2.2. Location-modulation mixtures

As in the previous section, given a measurable mother function $g : \mathbb{R}^d \to \mathbb{R}$, we define the *location-modulation* kernel $K_{\xi,\phi}(x) := g(x)\cos(\sum_{i=1}^d \xi_i x_i + \phi)$, for all $x \in \mathbb{R}^d$, all $\xi \in \mathbb{R}^d$ and all $\phi \in [0,\pi/2]$. Then we consider symmetric Gamma location-modulation mixtures of the type

$$f(x;\omega) := \int_{\mathbb{R}^d\times\mathbb{R}^d\times[0,\pi/2]} K_{\xi,\phi}(x-\mu)\, Q(d\xi d\mu d\phi;\omega), \qquad \forall x \in \mathbb{R}^d, \tag{2.3}$$

where $Q : \mathcal{B}(\mathbb{R}^d\times\mathbb{R}^d\times[0,\pi/2])\times\Omega \to [-\infty,\infty]$ is a symmetric Gamma random measure with base measure αF on $\mathbb{R}^d\times\mathbb{R}^d\times[0,\pi/2]$, and scale parameter $\eta > 0$.

2.2.3. Convergence of mixtures

Given a kernel $K : \mathcal{X}\times\mathbb{R}^d \to \mathbb{R}$ and a symmetric Gamma random measure Q, it is not clear a priori whether or not the mixture $y \mapsto \int K(x;y)\, Q(dx)$ converges or not, and in what sense. According to Rajput and Rosinski (1989) (see also Wolpert et al. (2011)), $y \mapsto \int K(x;y)\, Q(dx)$ converges almost-surely at all y for which

$$\int_{\mathbb{R}\times\mathcal{X}} (1 \wedge |uK(x;y)|)|u|^{-1} e^{-|u|\eta} F(dx) < +\infty.$$

Moreover, from the same references (or also in Kingman (1992)), if M is a complete normed space equipped with norm $\|\cdot\|$, then $y \mapsto \int K(x;y)\, Q(dx)$ converges almost-surely in M if

$$\int_{\mathbb{R}\times\mathcal{X}} (1 \wedge |u|\|K(x;\cdot)\|)|u|^{-1} e^{-|u|\eta} F(dx) < +\infty.$$

Since by definition F is a probability measure, we have for instance that the mixtures of equations (2.2) and (2.3) converges almost surely in L^∞ as soon as $\|K_A\|_\infty < +\infty$ for F-almost every $A \in \mathcal{E}$, or $\|K_{\xi,\phi}\|_\infty < +\infty$ for F-almost every $(\xi,\phi) \in \mathbb{R}^d\times[0,\pi/2]$.

2.3. Simulating the posterior

In this section we propose a Gibbs sampler for exploration of the posterior distribution of a mixture of kernels by a symmetric Gamma random measure. The sampler is based on the series representation of the next theorem, inspired from a result about Dirichlet processes from Favaro et al. (2012), adapted to symmetric Gamma processes. In theorem 2.1, we

consider $\mathcal{M}(\mathcal{X})$ the space of signed Radon measures on the measurable space $(\mathcal{X}, \mathcal{A})$. By the Riesz-Markov representation theorem (Rudin, 1974, Chapter 6), $\mathcal{M}(\mathcal{X})$ can be identified as the dual space of $\mathtt{C}_c(\mathcal{X})$, the space continuous functions with compact support. That said, we endow $\mathcal{M}(\mathcal{X})$ with the topology $\mathcal{T}_v$ of weak-* convergence (sometimes referred as the topology of *vague* convergence), that is, a sequence $\{\mu_n \in \mathcal{M}(\mathcal{X}) : n \in \mathbb{N}\}$ converges to $\mu \in \mathcal{M}(\mathcal{X})$ with respect to the topology $\mathcal{T}_v$, if for all $f \in \mathtt{C}_c(\mathcal{X})$,

$$\int_{\mathcal{X}} f(x)\, d\mu_n(x) \to \int_{\mathcal{X}} f(x)\, d\mu(x).$$

Dealing with prior distributions on $\mathcal{M}(\mathcal{X})$, we shall equip $\mathcal{M}(\mathcal{X})$ with a σ-algebra. Here it is always considered the Borel σ-algebra of $\mathcal{M}(\mathcal{X})$ generated by $\mathcal{T}_v$.

Before stating the main theorem of this section, we recall that a sequence of random variables $\{X_i \in \mathcal{X} : 1 \leq i \leq n\}$ is a *Pólya urn sequence* with base distribution $\alpha F(\cdot)$, where F is a probability distribution on $(\mathcal{X}, \mathcal{A})$ and $\alpha > 0$, if for all measurable set $A \in \mathcal{A}$,

$$P(X_1 \in A) = F(A), \quad P(X_{k+1} \in A \,|\, X_1, \ldots, X_k) = F_k(A)/F_k(\mathcal{X}),\ k = 2, \ldots, n-1,$$

where $F_k := \alpha F + \sum_{i=1}^{k} \delta_{X_i}$. We are now in position to state the main theorem of this section, which proof is given in section 2.A.

Theorem 2.1 – *Let $\mathcal{X}$ be a Polish space with Borel σ-algebra, $p > 0$ be integer, $T \sim \mathrm{Ga}(\alpha, \eta)$, independently, $J_1, \ldots, J_p \overset{\text{i.i.d}}{\sim} \mathrm{SGa}(1,1)$, and $\{X_i \in \mathcal{X} : 1 \leq i \leq p\}$ a Pólya urn sequence with base distribution $\alpha F(\cdot)$, independent of T and of the J_i's. Define the random measure, $Q_p := \sqrt{T/p} \sum_{i=1}^{p} J_i\, \delta_{X_i}$. Then $Q_p \overset{d}{\to} Q$, where Q is a symmetric Gamma random measure with base distribution $\alpha F(\cdot)$ and scale parameter $\sqrt{\eta}$.*

2.3.1. Convergence of sequences of mixtures

In theorem 2.1, we proved weak convergence of the sequence of approximating measures $(Q_p)_{p \geq 1}$ to the symmetric Gamma random measure, but it is not clear that mixtures of kernels by Q_p also converge. The next proposition establish convergence in L^q for general kernels, with $1 \leq q < +\infty$, the proof is similar to the proof of Favaro et al. (2012, Theorem 2), thus we defer it into section 2.6.2. For any kernel $K : \mathcal{X} \times \mathbb{R}^d \to \mathbb{C}$, and any (signed) measure Q on $(\mathcal{X}, \mathcal{A})$, we write

$$f^{(Q)}(y) := \int_{\mathcal{X}} K(x; y)\, Q(dx).$$

Proposition 2.1 – *If $x \mapsto K(x; y)$ is continuous for all $x \in \mathcal{X}$, vanishes outside a compact set, and bounded by a Lebesgue integrable function h, then for any $1 \leq q < +\infty$ we have $\lim_{p \to \infty} \| f^{(Q_p)} - f^{(Q)} \|_q = 0$ almost-surely.*

Under supplementary assumptions on K, we can say a little-more about uniform convergence of the approximating sequence of mixtures. Assuming that $y \mapsto K(x; y)$ is in L^1 for all $x \in \mathcal{X}$, we denote by $(x, u) \mapsto \widehat{K}(x; u)$ the L^1 Fourier transform on the second argument of $(x, y) \mapsto K(x; y)$.

Proposition 2.2 – *Let $y \mapsto K(x; y)$ be in L^1 for all $x \in \mathcal{X}$ and $\widehat{K}$ satisfies the assumption of proposition 2.1. Then $\lim_{p \to \infty} \| f^{(Q_p)} - f^{(Q)} \|_\infty = 0$ almost-surely.*

Proof. We can assume without loss of generality that $f^{(Q_p)}$ and $f^{(Q)}$ are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. By duality, it is clear that $\|f^{(Q_p)}(\cdot;\omega) - f^{(Q)}(\cdot;\omega)\|_\infty \leq \int_{\mathbb{R}^d} |\widehat{f}^{(Q_p)}(u,\omega) - \widehat{f}^{(Q)}(u;\omega)|\, du$, where $\widehat{f}$ denote the L^1 Fourier transform of f. Notice that by assumptions on K, $\widehat{f}^{(Q_p)}$ and $\widehat{f}^{(Q)}$ are well-defined for almost all $\omega \in \Omega$ (see section 2.2.3). Then by Fubini's theorem

$$\begin{aligned}\widehat{f}^{(Q_p)}(u;\omega) &= \int_{\mathbb{R}^d}\int_{\mathcal{X}} K(x;y)\, Q(dx;\omega)\ \mathrm{e}^{-iuy}\ dy \\ &= \int_{\mathcal{X}}\int_{\mathbb{R}^d} K(x;y)\,\mathrm{e}^{-iuy}\ dy\, Q(dx;\omega) = \int_{\mathcal{X}} \widehat{K}(x;u)\, Q(dx;\omega),\end{aligned}$$

and the conclusion follows from proposition 2.1. ■

2.3.2. General algorithm

From theorem 2.1, replacing Q by Q_p for sufficiently large p, we propose a Pólya urn Gibbs sampler adapted from algorithm 8 in Neal (2000). In the sequel, we refer to Q_p as the particle approximation of Q with p particles.

Let $Y = (Y_i)_{i=1}^n$ be observations coming from a statistical model with likelihood function $\mathcal{L}(f|Y)$, where $f : \mathbb{R}^d \to \mathbb{R}$ is the regression function on which we put a symmetric Gamma mixture prior distribution. Let $X = (X_i)_{i=1}^p$ be a Pólya urn sequence, $J := (J_1, \ldots, J_p)$ a sequence of i.i.d. SGa$(1,1)$ random variables, and $T \sim \mathrm{Ga}(\alpha, \eta)$ independent of $(X_i)_{i=1}^p$ and J. We introduce the clustering variables $C := (C_1, \ldots, C_p)$ such that $C_i = k$ if and only if $X_i = X_k^\star$ where $X^\star := X_1^\star, \ldots$ stands for unique values of $(X_i)_{i=1}^p$. In the sequel, C_{-i} stands for the vector obtained from removing the coordinate i to C, and the same definition holds for J *mutatis mutandis*. Given J, C, X, T and a measurable kernel $K : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ we construct f as

$$f(x) = \sqrt{\frac{T}{p}} \sum_{i=1}^p J_i\, K(X_i; x).$$

We propose the following algorithm. At each iteration, successively sample from :

(1) $C_i|C_{-i}, Y, X, J, T$, for $1 \leq i \leq p$. Let $n_{k,i} = \#_{\substack{1\leq l\leq n \\ l\neq i}}\{C_l = k\}$, $\kappa^{(p)}$ the number of distinct X_k values and κ_0 a chosen natural,

$$C_i \overset{\text{ind}}{\sim} \sum_{k=1}^{\kappa^{(p)}} n_{k,i}\, \mathcal{L}_{k,i}(X, J, T|Y)\, \delta_k(\cdot) + \frac{\alpha}{\kappa_0} \sum_{k=1}^{\kappa_0} \mathcal{L}_{k+\kappa^{(p)},i}(X, J, T|Y)\, \delta_{k+\kappa^{(p)}}(\cdot),$$

where $\mathcal{L}_{k,i}(X, J, T|Y)$ stands for the likelihood under hypothesis that particle i is allocated to component k (note that the likelihood evaluation requires the knowledge of whole distribution F under any allocation hypothesis).

(2) $X|C, Y, J, T$. Random Walk Metropolis Hastings on parameters.

(3) $J_i|J_{-i}, K, Y, X, T$, for $1 \leq i \leq p$. Independent Metropolis Hastings with prior SGa$(1,1)$ taken as i.i.d. candidate distribution for J_i. Note that for $n \to \infty$, the posterior distribution of $J_i|J_{-i}, C, Y, Z$ should be SGa$(1,1)$, then the number of particles p may be monitored using the acceptance ratio of the J_i's.

(4) $T|C, Y, X, J$. Random Walk Metropolis Hastings on scale parameter.

2.3.3. Assessing the convergence of the Markov Chain

The previous algorithm produces a Markov Chain whose invariant distribution is (an approximation of) the posterior distribution of a symmetric Gamma process mixture. However, if the Markov Chain is initialized in a region of low posterior probability mass, we may over-sample this region. To avoid such over-sampling, we discard the first n_0 samples of the chain using Geweke's convergence diagnostic (Geweke, 1992).

More precisely, we monitor the convergence of the chain using the log-likelihood function. We start the algorithm with Markov Chain initialized at random from prior distribution. Then after $n \gg n_0$ iterations we compute Geweke's Z-statistic for the log-likelihood using the whole chain; if the statistic is outside the 95% confidence interval we continue to apply the diagnostic after discarding 10%, 20%, 30% and 40% of the chain. If the Z-statistic is still outside 95% confidence interval, the chain is reported as failed to converge, and we restart the algorithm from a different initialization point.

Once we have discarded the first n_0 samples using Geweke's test, we run the chain sufficiently longer to get an *Effective Sample Size* (ESS) of at least 1000 samples, where we measure the ESS through the value of the log-likelihood at each iteration of the Markov Chain. A thinning of the chain is not required in general, however, we found in practice that a slight thinning improves the efficiency of the sampling.

In fig. 2.1, we draw some examples of temporal evolution of the log-likelihood on a simple univariate Gaussian mean regression problem. Here and after, we always choose step sizes in RWMH steps to achieve approximately 30% acceptance rates for each class of updates. Each subfigure represent 10 simulations with random starting point of the Markov Chain, distributed according to the prior distribution. We draw each subfigure varying the parameters liable to influence the mixing time of the chain, notably m and the number of particles. We observe that the speed at which the chain reach equilibrium is fast, especially when the number of particles is high. This last remark have to be balanced with the complexity in time of the algorithm which is $O(mnp)$ for a naive implementation, and, depending on the nature of the likelihood, can be reduced to $O(mp)$ or $O(mp^2)$.

2.4. Examples of simulations

We now turn our attention to simulated examples to illustrate the performance of mixture models. First, we use mixtures as a prior distribution on the regression function in the univariate mean regression problem with normal errors. Of course, the interest for mixture comes when the statistical model is more involved. Hence, in a second time we present simulation results for the multivariate inverse problem of CT imaging.

2.4.1. Mean regression with normal errors

We present results of our algorithm on several standard test functions from the wavelet regression litterature (see Marron et al., 1998), following the methodology from Antoniadis et al. (2001) (*i.e.* Gaussian mean regression with fixed design and unknown variance). However, it should be noticed that mixtures are not a Bayesian new implementation of wavelet regression, and are much more general (see for instance the next section). For each test function, the noise variance is chosen so that the root signal-to-noise ratio is

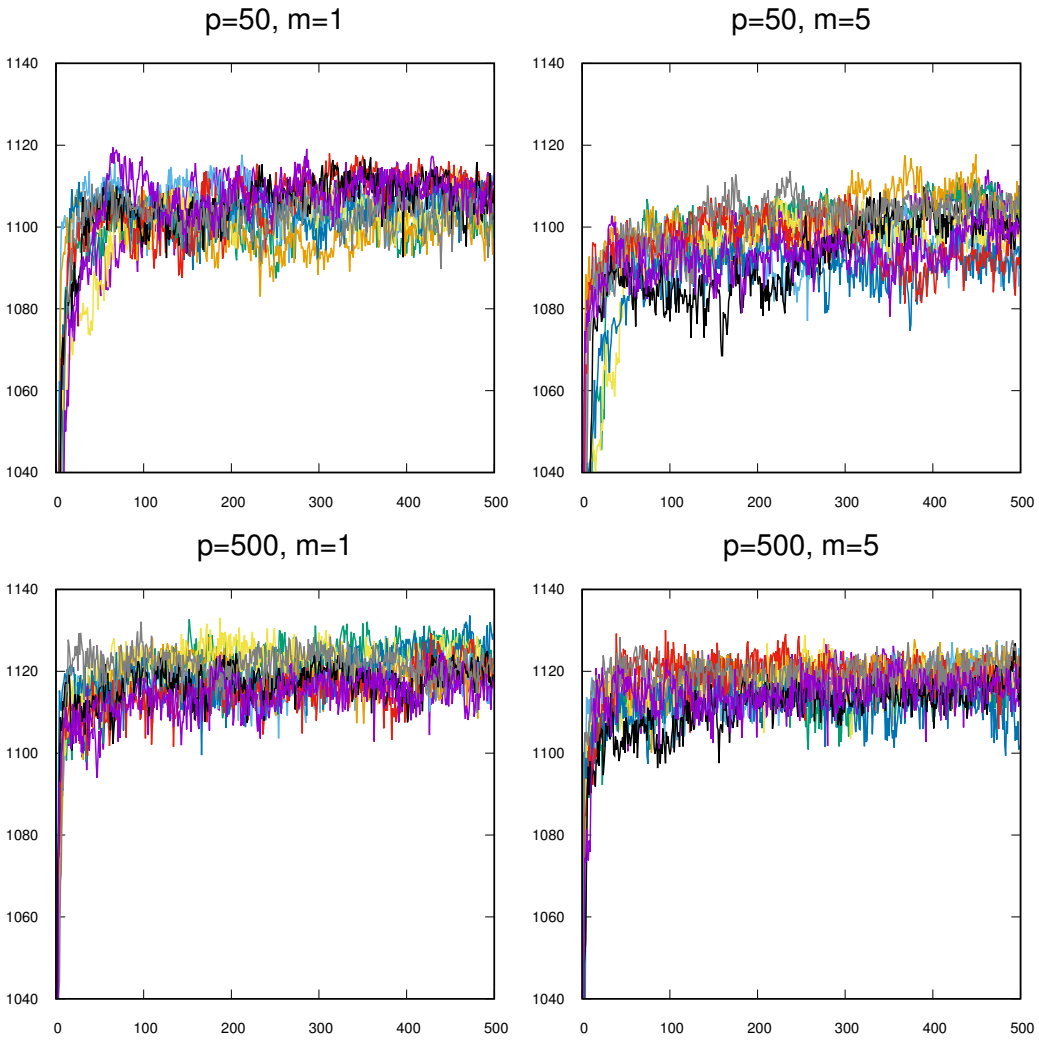

Figure 2.1: Time evolution of the log-likelihood for different starting point of the Markov Chain, chosen according to the prior distribution, and various parameters of the algorithm. The figure are taken from the test function *blip* of the section 2.4.1.

equal to 3 (a high noise level) and a simulation run was repeated 100 times with all simulation parameters constant, excepting the noise which was regenerated. We ran the algorithm for location-scale mixtures of Gaussians and Symmlet8, with normal $\mathcal{N}(0.5, 0.3)$ distribution as prior distribution on translations, and a mixture of Gamma distributions for scales ($\mathrm{Ga}(30, 0.06)$ and $\mathrm{Ga}(2, 0.04)$ with expectation 500 and 50 respectively). In addition of the core algorithm of section 2.3.2, we also added

- a Gibbs step estimation of the noise variance, with Inverse-Gamma prior distributon,
- a $\mathrm{Ga}(2, 0.5)$ (with expectation 4) prior on α, with sampling of α done through a Gibbs update according to the method proposed in West (1992),
- a Dirichlet prior on the weights of the mixture of $\mathrm{Ga}(20, 0.2)$ and $\mathrm{Ga}(2, 0.1)$, with

sampling of the mixture weights done through Gibbs sampling in a standard way,
- a $\mathrm{Ga}(5, 10)$ (with expectation 0.5) prior on T, instead of normally $\mathrm{Ga}(\alpha, \eta)$, which add more flexibility.

The choice of the mixture distribution as prior on scales may appear surprising, but we found in practice that using bimodal distribution on scales substantially improve performance of the algorithm, especially when there are few data available and/or high noise, because in general both large and small scales components are needed to estimate the regression function.

We ran the algorithm for $n = 128$ and $n = 1024$ data, and the performance is measured by its average root mean square error, defined as the average of the square root of the mean squared error $n^{-1} \sum_{i=1}^{n} |\widehat{f}(x_i) - f_0(x_i)|^2$, with $\widehat{f}$ denoting the posterior mean and f_0 the true function. We ran on the same dataset the Translation-Invariant with hard thresholding algorithm (TI-H) and Symmlet8 wavelets (see Antoniadis et al. (2001)), which is one of the best performing algorithm on this collection of test functions. We ran our algorithm with Symmlet8 kernels to make this comparison more relevant, since the choice of the kernel has major impact on the performance of the algorithm (see section 2.4.1 below).

Alternatives

In Wolpert et al. (2011), authors develop a reversible-jump MCMC scheme where the random measure is thresholded, *i.e.* small jumps are removed, yielding to a compound Poisson process approximation of the random measure, with almost-surely a finite number of jumps, allowing numerical computations. We also ran their algorithm with a thresholding level of $\epsilon = 0.05$ (which seems to give the best performance), a $\mathrm{Ga}(15, 1)$ prior on η, and all other parameters being exactly the same as described in the previous section. We use the criteria of section 2.3.3 to stop the running of the chain.

Choosing the number of particles

It is not clear how to choose the number of particles in the algorithm. In theory, the higher is the better. In practice, however, we recommend choosing the number of particles according to the acceptance rate of particles weights move in step 3 of the algorithm. We found in practice that a level of acceptance between 20% and 30% is acceptable, as illustrated in fig. 2.2.

Simulation results

In tables 2.1 and 2.2 we summarize the results for location-scale mixtures of Gaussians and Symmlet8 produced by the algorithm of section 2.3.2 and by the RJMCMC algorithm of Wolpert et al. (2011), with the TI-H method as reference. We used $p = 150$ particles for both the datasets with $n = 128$ covariates and $n = 1024$ covariates, which is a nice compromise in terms of performance and computational cost. Regarding our algorithm and the RJMCMC algorithm, no particular effort was made to determine the value of the fixed parameters.

Obviously the Gibbs algorithm allow for sampling the *full* posterior distribution, pemitting estimation of posterior credible bands, as illustrated in figs. 2.3 and 2.4, where

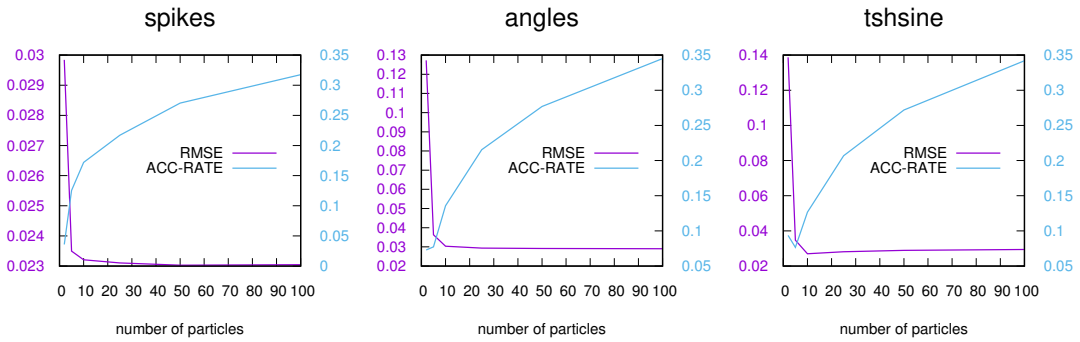

Figure 2.2: Mean over 100 runs of RMSE versus acceptance rate in step 3 of the algorithm for some typical test functions. For each signal the number of covariates is set to 128 and the RNSR is equal to 3.

	TI-H	Gibbs		RJMCMC	
Function	Symm8	Gauss	Symm8	Gauss	Symm8
step	0.0589	0.0517	0.0551	0.0550	0.0565
wave	0.0319	0.0323	0.0306	0.0342	0.0370
blip	0.0307	0.0301	0.0316	0.0323	0.0373
blocks	0.0464	0.0343	0.0374	0.0383	0.0418
bumps	0.0285	0.0162	0.0229	0.0224	0.0345
heavisine	0.0257	0.0267	0.0264	0.0280	0.0289
doppler	0.0443	0.0506	0.0418	0.0526	0.0493
angles	0.0293	0.0266	0.0282	0.0274	0.0305
parabolas	0.0344	0.0301	0.0307	0.0312	0.0396
tshsine	0.0255	0.0285	0.0277	0.0291	0.0339
spikes	0.0237	0.0178	0.0207	0.0199	0.0218
corner	0.0177	0.0171	0.0170	0.0182	0.0255

Table 2.1: Summary of root mean squared errors of different algorithms for $n = 128$ covariates and a root signal to noise ratio of 3.

the credible bands were drawn retaining the 95% samples with the smaller ℓ_2-distance with respect to the posterior mean estimator. Although the algorithm samples an approximated version of the model, it is found that the accuracy of credible bands is quite good since the true regression function almost never comes outside the sampled 95% bands, as it is visible in the example of figs. 2.3 and 2.4. Despite the algorithm efficiency, future work should be done to develop new sampling techniques for regression with mixture models, mainly to improve computation cost.

Discussion

Obviously, the computation cost for our algorithm is high compared to TI-H, or any other classical wavelet thresholding method, even considering that it can intrinsically

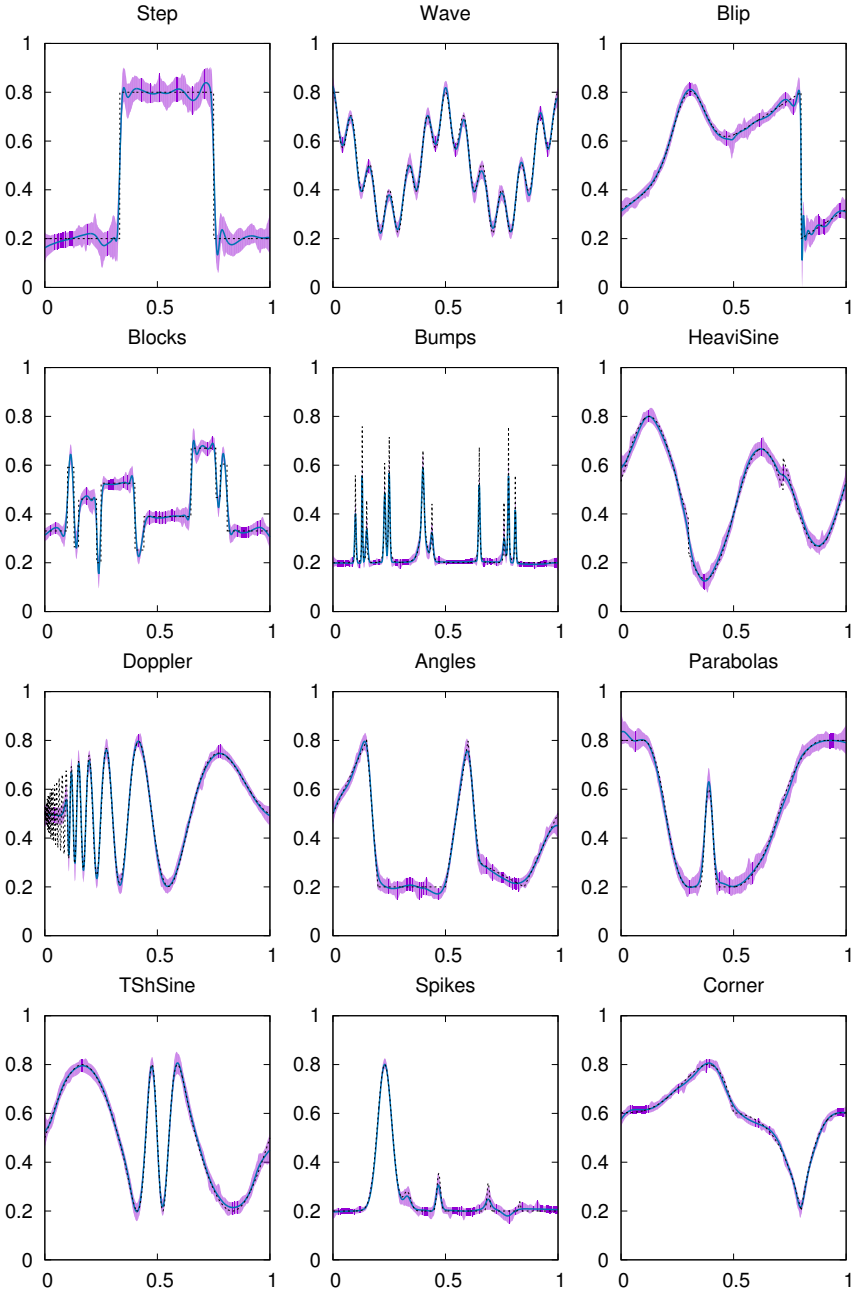

Figure 2.3: Example of simulation results using location-scale mixtures of Gaussians. The root signal-to-noise ratio is equal to 3 for sample size of 1024 design points. The true regression function is represented with dashes, the mean of the sampled posterior distribution in blue and sampled 95% credible bands in pink.

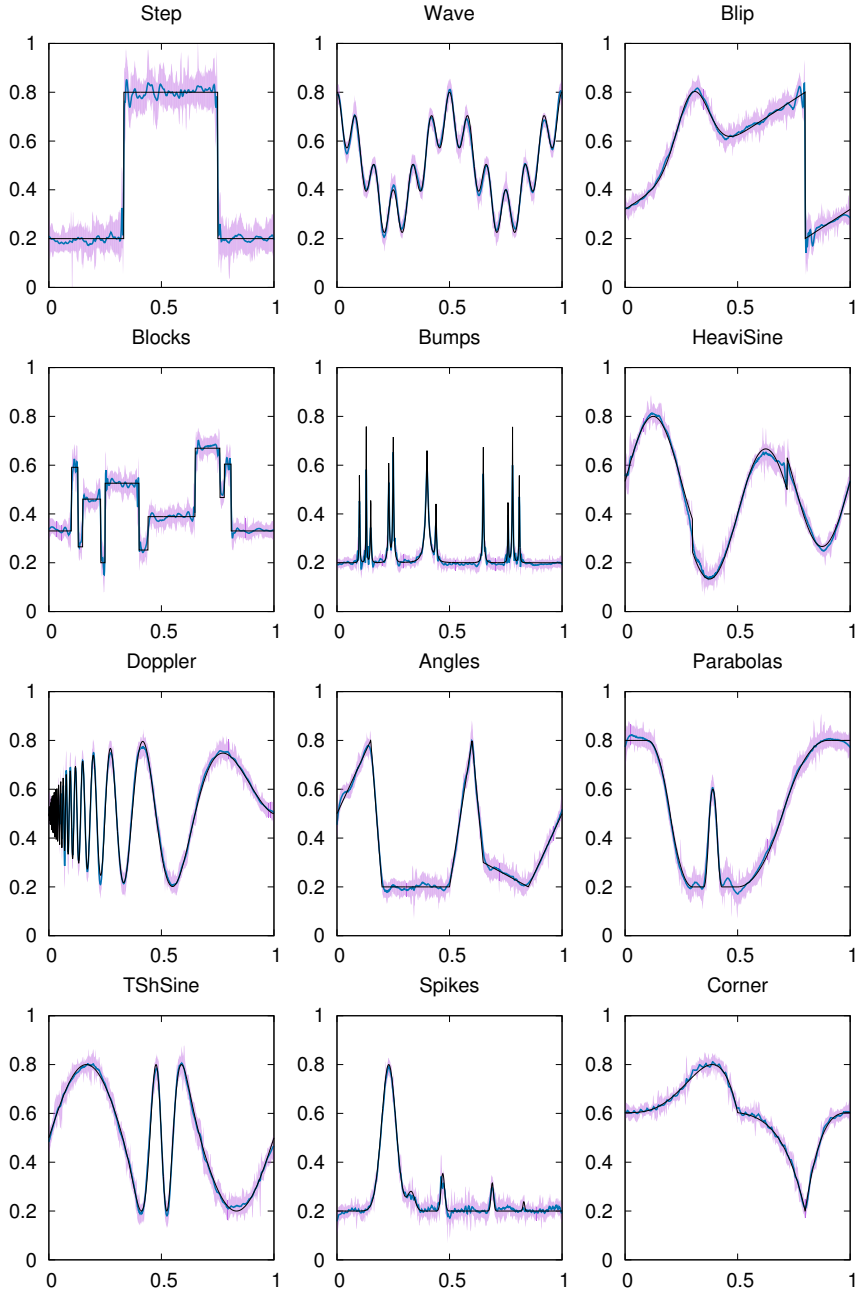

Figure 2.4: Example of simulation results using location-scale mixtures of Symmlet8. The root signal-to-noise ratio is equal to 3 for sample size of 1024 design points. The true regression function is represented with dashes, the mean of the sampled posterior distribution in blue and sampled 95% credible bands in pink.

	TI-H	Gibbs		RJMCMC	
Function	Symm8	Gauss	Symm8	Gauss	Symm8
step	0.0276	0.0268	0.0289	0.0282	0.0300
wave	0.0088	0.0118	0.0108	0.0133	0.0117
blip	0.0148	0.0162	0.0172	0.0180	0.0183
blocks	0.0222	0.0230	0.0241	0.0247	0.0256
bumps	0.0122	0.0132	0.0182	0.0201	0.0232
heavisine	0.0154	0.0134	0.0139	0.0147	0.0147
doppler	0.0180	0.0207	0.0196	0.0261	0.0225
angles	0.0123	0.0120	0.0123	0.0125	0.0128
parabolas	0.0135	0.0124	0.0132	0.0147	0.0145
tshsine	0.0107	0.0109	0.0111	0.0131	0.0120
spikes	0.0110	0.0075	0.0095	0.0095	0.0103
corner	0.0077	0.0075	0.0081	0.0095	0.0085

Table 2.2: Summary of root mean squared errors of different algorithms for $n = 1024$ covariates and a root signal to noise ratio of 3.

compute credible bands. But, as mentioned in Antoniadis et al. (2001), the choice of the kernel is crucial to the performance of estimators. The attractiveness of mixtures then comes because we are not restricted to location-scale or location-modulation kernels, and almost any function is acceptable as a kernel, which is not the case for most regression methods. Moreover, there is no requirements on how the data are spread, which makes the method interesting in inverse problems, such as in the next section.

2.4.2. Multivariate inverse problem example

Many medical imaging modalities, such as X-ray computed tomography imaging (CT), can be described mathematically as collecting data in a Radon transform domain. The process of inverting the Radon transform to form an image can be unstable when the data collected contain noise, so that the inversion needs to be regularized in some way. Here we model the image of interest as a measurable function $f : \mathbb{R}^2 \to \mathbb{R}$, and we propose to use a location-scale mixtures of Gaussians to regularize the inversion of the Radon transform.

More precisely, the Radon transform $R_f : \mathbb{R}^+ \times [0, \pi] \to \mathbb{R}$ of f is such that $R_f(r, \theta) = \int_{-\infty}^{+\infty} f(r\cos\theta - t\sin\theta, r\sin\theta + \cos\theta)\, dt$. Then we consider the following model. Let $n, m \geq 1$. Assuming that the image is supported on $[-1, 1]^2$ we let $r_1, \ldots, r_n$ equidistributed in $[-\sqrt{2}, \sqrt{2}]$ and $\theta_1, \ldots, \theta_m$ equidistributed in $[0, \pi]$. Then,

$$\begin{aligned} Y_{nm} &\sim \mathcal{N}(R_f(r_n, \theta_m), \sigma^2) \quad \forall n, m \\ f &\sim \Pi, \end{aligned}$$

where Π is a symmetric Gamma process location-scale mixture with base measure $\alpha F_A \times F_\mu$ on $\mathcal{E} \times \mathbb{R}^2$, $\alpha > 0$, and scale parameter $\eta > 0$. In the sequel, we use a normal distribution with mean zero and covariance matrix $\mathrm{diag}(\tau, \tau)$ as distribution for F_μ. Regarding F_A,

the choice is more delicate; we choose a prior distribution over the set of *shearlet-type* matrices of the form

$$\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix},$$

where we set a $\mathcal{N}(1, \sigma_a^2)$ distribution over the coefficient a and $\mathcal{N}(0, \sigma_s^2)$ over the coefficient s. This type of prior distribution for F_A is particularly convenient for capturing anisotropic features such as edges in images (Eas, 2009).

We ran our algorithm for $n = 256$ and $m = 128$ (32768 observations, a small amount), using the Shepp and Logan phantom as original image (Shepp and Logan, 1974). The variance of the noise is $\sigma^2 = 0.1$, whereas the image take value between 0 and 2. Both the original image and the reconstruction are visible in fig. 2.5. Finally, we should mention that the choice of the Gaussian kernel for the mixture is convenient since it allows to compute the likelihood analytically. However, from a practical side, a full implementation of the algorithm with the intention of reconstructing CT images may benefit from using a different kernel.

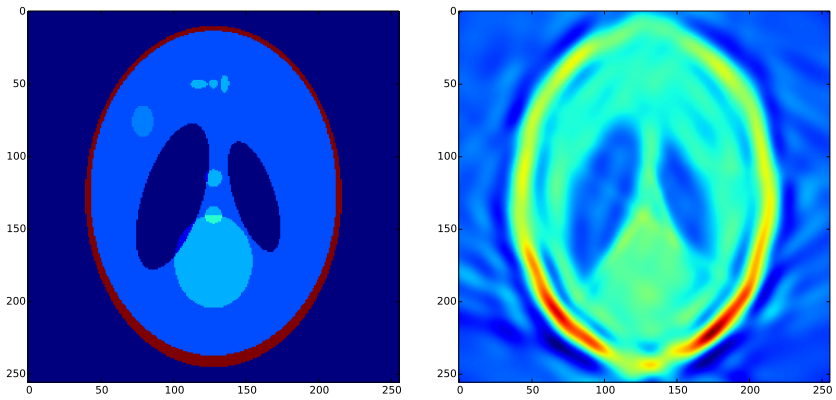

Figure 2.5: Simulation of X-ray computed tomography imaging using symmetric Gamma process location-scale mixture of Gaussians. On the left: the original image. On the right: the reconstructed image from 32768 observations of the Radon transform of the original image in a Gaussian noise.

2.5. Rates of convergence

In this section, we investigate posterior convergence rates in fixed design Gaussian regression for both symmetric Gamma location-scale mixtures and symmetric Gamma location-modulation mixtures.

2.5.1. Notations

In the sequel we use repeatedly the following notations.

- The conventional multi-index notation, for all $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ and all $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$ we write $|\alpha| := \alpha_1 + \cdots + \alpha_d$, $\alpha! := \alpha_1! \ldots \alpha_d!$, and $z^\alpha := z_1^{\alpha_1} \ldots z_d^{\alpha_d}$.

Moreover, for all $f : \mathbb{R}^d \to \mathbb{R}$ with continuous k-th order partial derivatives at $x \in \mathbb{R}^d$ we write

$$D^\alpha f(x) := \frac{\partial^{|\alpha|} f}{\partial z_1^{\alpha_1} \dots \partial z_n^{\alpha_d}}(x), \quad |\alpha| \le k.$$

- Let Ω be an open subset of $\mathbb{R}^d$ and $\overline{\Omega}$ be the closure of Ω. For any $\beta > 0$, we define $\mathcal{C}^\beta(\overline{\Omega})$, the Hölder space on $\overline{\Omega}$, as the set of all functions on $\overline{\Omega}$ such that $\|f\|_{\mathcal{C}^\beta} := \max_{|\alpha|\le k} \sup_{x\in\Omega} |D^\alpha f(x)| + \max_{|\alpha|=k} \sup_{x\neq y\in\Omega} |D^\alpha f(x) - D^\alpha f(y)|/|x-y|^{\beta-k}$ is finite, where k is the largest integer strictly smaller than β.
- We denote by $|\cdot|_d$ the standard euclidean norm on $\mathbb{R}^d$, and, for any $x, y \in \mathbb{R}^d$, xy is the standard inner product. For any $d \times d$ matrix A with real eigenvalues, we denote $\lambda_1(A) \ge \dots \ge \lambda_d(A)$ its eigenvalues in decreasing order, $\|A\| := \sup_{x\neq 0} |Ax|_d/|x|_d$ its spectral norm, and $\|A\|_{\max} := \max_{i,j} |A_{ij}|$, where A_{ij} are the entries of A.
- Given a signed measure μ on a measurable space $(\mathcal{X}, \mathcal{A})$, we let μ^+ and μ^- denote respectively the positive and negative part of the Jordan decomposition of μ. Also, $|\mu| = \mu^+ + \mu^-$ denote the total variation measure of μ.
- Inequalities up to a generic constant are denoted by the symbols $\lesssim$ and $\gtrsim$.

2.5.2. The model

We consider the problem of a random response Y corresponding to a deterministic covariate vector x taking values in $[-S, S]^d$ for some $S > 0$. We aim at estimating the regression function $f : [-S, S]^d \to \mathbb{R}$ such that $f(x_i) = \mathbb{E}\, Y_i$, based on independent observations of Y. More precisely, the nonparametric regression model we consider is the following,

$$\begin{aligned} Y_i|\epsilon_i &= f(x_i) + \epsilon_i, \quad i = 1, \dots, n, \\ \epsilon_1, \dots, \epsilon_n|\sigma^2 &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2), \quad \text{independently of } (f, \sigma), \\ (f, \sigma) &\sim \Pi, \end{aligned}$$

with Π the distribution on an abstract space Θ, given by $\sigma \sim P^\sigma$ independently of f drawn from the distribution of a symmetric Gamma process mixture.

2.5.3. A general result

Let $P_{\theta,i}$ denote the distribution of of Y_i under the parameter $\theta = (f, \sigma)$, P_θ^n the joint distribution of $(Y_1, \dots, Y_n)$, P_θ^∞ the distribution of the infinite sequence $(Y_1, \dots, Y_\infty)$, and $\|f\|_{2,n}^2 := n^{-1} \sum_{i=1}^n |f(x_i)|^2$. Let define the distance $\rho_n(\theta_0, \theta_1) := \|f - f_0\|_{2,n} + |\log \sigma_0 - \log \sigma_1|$. For the regression method based on Π, we say that its posterior convergence rate at θ_0 in the metric ρ_n is ϵ_n if there is $M < +\infty$ such that

$$\lim_{n\to\infty} \Pi\left(\{\theta \in \Theta : \rho_n(\theta, \theta_0) > M\epsilon_n\}|Y_1, \dots, Y_n\right) = 0 \quad P_{\theta_0}^\infty\text{-a.s.} \tag{2.4}$$

Most of the approach to rates of convergence rely on idea coming from density mixtures models (Ghosal et al., 2000; Shen et al., 2013; Canale and De Blasi, 2017). Indeed, we prove that equation (2.4) hold by verifying a set of sufficient conditions established

in theorem 2.2. For $\epsilon > 0$ and any subset A of a metric space equipped with metric ρ, let $N(\epsilon, A, \rho)$ denote the ϵ-covering number of A, *i.e.* the smallest number of balls of radius ϵ needed to cover A. Also, for all $i = 1, \dots, n$, define $K_i(\theta_0, \theta) := \int (\log dP_{\theta_0,i}/dP_{\theta,i})\, dP_{\theta_0,i}$ and $V_{2,i}(\theta_0, \theta) := \int (\log dP_{\theta_0,i}/dP_{\theta,i} - K_i(\theta_0, \theta))^2\, dP_{\theta_0,i}$, and let

$$K_n(\theta_0, \epsilon) := \left\{ \theta \,:\, \frac{1}{n}\sum_{i=1}^{n} K_i(\theta_0, \theta) \le \epsilon^2,\ \frac{1}{n}\sum_{i=1}^{n} V_{2,i}(\theta_0, \theta) \le \epsilon^2 \right\},$$

be the Kullback-Leibler ball of size ϵ around $\theta_0 := (f_0, \sigma_0)$. Theorem 2.2 is the analogue of theorem 5 in Ghosal and van der Vaart (2007b) for the Gaussian mean regression with fixed design ; the major difference reside on constructing suitable test functions, and extra cares have to taken regarding the fact that observations are not i.i.d. The proof of theorem 2.2 is given in section 2.7.

Theorem 2.2 – *Let $K := 3(32 \vee 4\sigma_0^2)^{-1}$, and $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$. Suppose that $\Theta_n \subset \Theta$ is such that $\Pi(\Theta_n^c) \lesssim \mathrm{e}^{-3n\epsilon_n^2}$ for n large enough. Assume that $\Theta_n \subseteq \cup_j \Theta_{n,j}$ is such that for some $M > 0$,*

$$\lim_n \textstyle\sum_j \sqrt{N(M\epsilon_n, \Theta_{n,j}, \rho_n)}\sqrt{\Pi(\Theta_{n,j})}\, \mathrm{e}^{-(KM^2-2)n\epsilon_n^2} = 0,$$
$$\Pi(K_n(\theta_0, \epsilon_n)) \gtrsim \mathrm{e}^{-n\epsilon_n^2}.$$

Then $\Pi(\theta \in \Theta \,:\, \rho_n(\theta_0, \theta) > 12M\epsilon_n | Y_1, \dots, Y_n) \to 0$ in $P_{\theta_0}^n$-probability.

2.5.4. Supplementary assumptions

In order to derive rates of convergence (and only for this) we make supplementary assumptions on the choice of the mother function g and of the base measure αF.

Location-scale mixtures

We restrict our discussion to priors for which the following conditions are verified. We assume that

- $g : \mathbb{R}^d \to \mathbb{R}$ is a non zero Schwartz function such that $|g(x)| \lesssim \exp(-C_0|x|_d^\tau)$ for some $C_0, \tau > 0$. We assume that there is $0 \le \gamma < 1$ such that $\sup_{|\alpha|=k} |D^\alpha g(0)| \lesssim \exp(\gamma k \log k)$ for all k large enough ; this last assumption is not obvious, it is for example met with $\gamma = 1/2$ if g is a multivariate Gaussian (see proposition 2.14 in appendix).
- $\alpha F := \alpha F_A \times F_\mu$, where F_A is a probability measure on $\mathcal{E}_s$, the space of symmetric positive definite $d \times d$ reals matrices, and F_μ a probability measure on $[-2S, 2S]^d$. We also assume that there exist positive constants $\kappa > 0$, $\kappa^* > d(d-1)$, $a_1, \dots, a_5$, $b_1, \dots, b_6$, $C_1, \dots, C_3$ such that for any $0 < s_1 \le \cdots \le s_d$, any $z_0 \in [-2S, 2S]^d$, all

$t \in (0,1)$ and all $x > 0$ sufficiently large

$$F_\mu(z \ : \ |z - z_0| \leq t) \geq b_1 t^{a_1}, \tag{2.5}$$
$$F_A(A \ : \ \lambda_d(A^{-1}) \geq x) \leq b_2 \exp(-C_2 x^{a_2}), \tag{2.6}$$
$$F_A(A \ : \ \lambda_1(A^{-1}) < 1/x) \leq b_3 x^{-a_3}, \tag{2.7}$$
$$F_A\left(A \ : \ s_j < \lambda_j(A^{-1}) < s_j(1+t), \ 1 \leq j \leq d\right) \geq b_4 s_d^{a_4} t^{a_5} \exp(-C_3 s_d^{\kappa/2}), \tag{2.8}$$
$$F_A(A \ : \ \lambda_1(A)/\lambda_d(A) > x) \leq b_6 x^{-\kappa^*}. \tag{2.9}$$

Equations (2.6) to (2.8) are classical and are met for instance with $\kappa = 2$ if F_A is the inverse-Wishart distribution (Shen et al., 2013, lemma 1). For a thorough discussion about equation (2.9) we refer to Canale and De Blasi (2017) and references therein.

- P^σ is a probability distribution on $(0, \infty)$. We also assume that there are positive constants a_7, a_8, a_9, b_7, b_8, C_8, and b_9 eventually depending on $\sigma_0 > 0$, such that for all $t \in (0,1)$

$$P^\sigma(\sigma \ : \ \sigma > x) \leq b_7 x^{-a_7}, \tag{2.10}$$
$$P^\sigma(\sigma \ : \ \sigma \leq 1/x) \leq b_8 \exp(-C_8 x^{a_8}), \tag{2.11}$$
$$P^\sigma(\sigma \ : \ \sigma_0 \leq \sigma \leq \sigma_0(1+t)) \geq b_9 t^{a_9}. \tag{2.12}$$

Location-modulation mixtures

We restrict our discussion to priors for which the following conditions are verified. We assume that

- $g : \mathbb{R}^d \to \mathbb{R}$ is a non zero Schwartz function such that $g(x) \geq 0$ for all $x \in \mathbb{R}^d$ and $|g(x)| \lesssim \exp(-C_0 |x|_d^\tau)$ for some $C_0 > 0$ and $\tau > 1$. We assume that there is a set $E \subseteq [-\pi, \pi]^d$ with strictly positive Lebesgue measure and a constant $C > 0$ such that $g(x) \geq C$ on E. We also assume that there is $0 \leq \gamma < 1$ such that $\sup_{|\alpha|=k} |D^\alpha g(0)| \lesssim \exp(\gamma k \log k)$ for all k large enough. As in the previous section, these assumptions are met for the multivariate Gaussian with $E = [-\pi, \pi]^d$, $\gamma = 1/2$ and $\tau = 2$ (see proposition 2.14 in appendix).
- $\alpha F := \alpha F_\xi \times F_\mu \times F_\phi$, where F_ξ is a probability measure on $\mathbb{R}^d$, F_μ a probability measure on $[-2S, 2S]^d$, and F_ϕ a probability measure on $[0, \pi/2]$. For all $t \in (0,1)$ and all $z_0 \in [-2S, 2S]^d$ we assume that F_μ satisfies equation (2.5). We assume that there are positive constants a_{10}, b_{10} such that for all $t \in (0,1)$ and all $\phi_0 \in [0, \pi/2]$ we have that $F_\phi(\phi \ : \ |\phi - \phi_0| \leq t) \geq b_{10} t^{a_{10}}$. We also assume that there exist positive constants $\eta > (d-1)/2$, $a_{12}, a_{13}, b_{11}, b_{12}$ such that for all $t \in (0,1)$, all $\xi_0 \in \mathbb{R}^d$ and for all $x > 0$

$$F_\xi(\xi \ : \ |\xi|_d \geq x) \leq b_{11}(1+x)^{-2(\eta+1)} \tag{2.13}$$
$$F_\xi(\xi \ : \ |\xi - \xi_0|_d \leq t) \geq b_{12} |\xi_0|_d^{-a_{12}} t^{a_{13}}. \tag{2.14}$$

- P^σ satisfies the same assumptions of equations (2.10) to (2.12).

2.5.5. Results

Theorem 2.2 serves as a starting point for proving rates of contraction for symmetric Gamma process location-scale and location-modulation mixtures in the model of section 2.5.2. The proofs of the next theorems resemble to de Jonge and van Zanten (2010), but, they consider only a location mixture with locations taken on a lattice, allowing for a very specific construction of the sets Θ_n. Here, we do not assume that locations are spread over a lattice, which makes the construction of Θ_n more involved. Our construction is inspired from Shen et al. (2013) for Dirichlet processes mixtures, but adapted to symmetric Gamma processes (indeed, the same construction should work for many Lévy processes). Also, theorem 2.2 allows for partitioning Θ_n onto slices $\Theta_{n,j}$, a step which is unnecessary for location mixtures (de Jonge and van Zanten, 2010; Shen et al., 2013), but yields to better rates and weaker assumptions on the prior when dealing with location-scale (Canale and De Blasi, 2017) and location-modulation mixtures.

Regarding the model of section 2.5.2, with deterministic covariates $x_1, \dots, x_n$ arbitrary spread in $[-S, S]^d$, we have the following theorem for location-scale mixtures. We notice that unlike de Jonge and van Zanten (2010), we do not assume that the covariates are spread on a strictly smaller set than $[-S, S]^d$, *i.e.* the support of the covariates and the domain of the regression function are the same.

Theorem 2.3 – *Let $\zeta = 1 \vee 2/(\tau - \gamma\tau)$. Suppose that $f_0 \in \mathcal{C}^\beta[-S, S]^d$ for some $S > 0$. Under the assumptions of section 2.5.4, the equation (2.4) holds for the location-scale prior with $\epsilon_n^2 = n^{-2\beta/(2\beta+d+\kappa/2)}(\log n)^{2\beta d(\zeta-1)/(2\beta+d+\kappa/2)}$.*

Theorem 2.3 gives a rate of contraction analogous to the rates found in Canale and De Blasi (2017), that is to say, suboptimal with respect to the frequentist minimax rate of convergence. Indeed, if one use an Inverse-Wishart distribution for F_A, then $\kappa = 2$; we can achieve $\kappa = 1$ with a distribution supported on diagonal matrices which assign square of inverse gamma random variables to non-null element of the matrix. Obviously, the choice of F_A matters since it has a direct influence on the rates of contraction of the posterior. Also notice that the rates depends on $\kappa/2$, which is slightly better than the κ dependency found in Canale and De Blasi (2017). The reason is relatively artificial, since this follows from the fact that we put a prior on dilation matrices of the mixture, whereas they set a prior on square of dilation matrices (covariance matrices).

Location-modulation mixtures were never considered before, because they are not satisfactory for estimating a density. In comparison with location-scale mixtures, the major difference in proving contraction rates rely on approximating sufficiently well the true regression function. We use a new approximating scheme, based on standard of Fourier series analysis, yielding the following theorem.

Theorem 2.4 – *Suppose that $f_0 \in \mathcal{C}^\beta[-S, S]^d$ for some $S > 0$. Under the assumptions of section 2.5.4, the equation (2.4) holds for the location-modulation prior with $\epsilon_n^2 = n^{-2\beta/(2\beta+d)}(\log n)^{2\beta(2d+1)/(2\beta+d)}$.*

Although it was not surprising that location-scale mixtures yield suboptimal rates of convergence, we would have expected that location-modulation mixtures could be suboptimal too, which is not the case (up to a power of $\log n$ factor). Moreover, location-modulation mixtures seem less stiff than location mixtures (Shen et al., 2013), hence they might be interesting to consider in regression.

Finally, it should be mentioned that all the rates here are adaptive with respect to $\beta > 0$; that is location-scale and location-modulation mixtures achieve these rates simultaneously for all $\beta > 0$.

2.6. Proofs of section 2.3

2.6.1. Preliminaries on convergence of signed random measures

It is well known for random (non-negative) measures that it is enough to show weak convergence of finite dimensional distributions on a semiring of bounded sets generating $\mathcal{A}$ to prove vague convergence of the distribution, see for instance Kallenberg (1983, Theorem 4.2) or Daley and Vere-Jones (2007, Theorem 11.1.VII). This fact remains true for random signed measures, but not in an obvious way. Indeed, it is well known that the vague topology is not metrizable on $\mathcal{M}(\mathcal{X})$, even if $\mathcal{X}$ is Polish (for example, see Remark 1.2 in Del Barrio et al. (2007)), making the vague topology nasty to handle on $\mathcal{M}(\mathcal{X})$. In particular, it is not as direct as in the case of non-negative measures to prove that the σ-algebra generated by the sets $\{\{\mu \in \mathcal{M}(\mathcal{X}) : \mu(B) \in A\} : A \in \mathcal{B}(\mathbb{R}),\ B \in \mathcal{R}\}$, where $\mathcal{R}$ is a ring of bounded sets generating $\mathcal{A}$, coincides with the Borel σ-algebra of $\mathcal{M}(\mathcal{X})$, given the topology of vague convergence. However, once this last fact is proved, everything in the proof of Kallenberg (1983, Theorem 4.2) remains valid for signed random measures.

Surprisingly, there is not so much literature on vague convergence of signed random measures, and as our knowledge, the only reference available on this subject is Jacob and Oliveira (1995). We state here the result of interest for us, with only a sketch of the proof, as the details can be found in the original article.

Lemma 2.1 – *Let $\mathcal{R} \subset \mathcal{A}$ denote the ring of bounded Borel sets of $\mathcal{X}$. Then the Borel σ-algebra of $\mathcal{M}(\mathcal{X})$ (given the weak-* topology) coincides with the σ-algebra generated by the sets $\{\{\mu \in \mathcal{M} : \mu(B) \in A\} : A \in \mathcal{B}(\mathbb{R}),\ B \in \mathcal{R}\}$ and also $\{\{\mu \in \mathcal{M} : \mu(f) \in A\} : A \in \mathcal{B}(\mathbb{R}),\ f \in \mathtt{C}_c(\mathcal{X})\}$.*

Sketch of proof. First, we shall prove that $\mathcal{S} := \sigma\{\{\mu \in \mathcal{M} : \mu(B) \in A\} : A \in \mathcal{B}(\mathbb{R}),\ B \in \mathcal{R}\} = \sigma\{\{\mu \in \mathcal{M} : \mu(f) \in A\} : A \in \mathcal{B}(\mathbb{R}),\ f \in \mathtt{C}_c(\mathcal{X})\}$. Using the Hahn-Jordan decomposition of signed measures, this is a straightforward adaptation of Kallenberg (1983, Lemma 1.4).

Also, the argument of Kallenberg (1983, Lemma 4.1) for proving $\mathcal{S} \subset \mathcal{B}(\mathcal{M})$ remains valid here, but the converse inclusion is not as direct. Let $\mathcal{M}^+ \subset \mathcal{M}$ denote the cone of non-negative measures, and endow $\mathcal{M}^+$ with the topology $\mathcal{T}_v^+$ of vague convergence (*i.e.* μ_n converges to μ if $\mu_n(f) \to \mu(f)$ for any $f \in \mathtt{C}_c^+$) and corresponding Borel σ-algebra $\mathcal{B}(\mathcal{M}^+)$. We denote $\mathcal{S}^+$ the trace of $\mathcal{S}$ over $\mathcal{M}^+$. Hence, it suffices to prove that

(1) $\mathcal{S}^+ = \mathcal{B}(\mathcal{M}^+)$,
(2) $P : (\mathcal{M}, \mathcal{S}) \to (\mathcal{M}^+ \times \mathcal{M}^+, \mathcal{S}^+ \times \mathcal{S}^+)$, such that $P(\mu) := (\mu^+, \mu^-)$, is measurable,
(3) $R : (\mathcal{M}^+ \times \mathcal{M}^+, \mathcal{B}(\mathcal{M}^+) \times \mathcal{B}(\mathcal{M}^+)) \to (\mathcal{M}, \mathcal{B}(\mathcal{M}))$, such that $R(\mu, \nu) := \mu - \nu$, is measurable.

These 3 conditions imply that $R \circ P : (\mathcal{M}, \mathcal{S}) \to (\mathcal{M}, \mathcal{B}(\mathcal{M}))$ is $\mathcal{S}/\mathcal{B}(\mathcal{M})$-measurable, and since $R \circ P$ is just the identity mapping, this implies $\mathcal{B}(\mathcal{M}) \subset \mathcal{S}$, as required. ■

2.6.2. Proofs

Proof of theorem 2.1. In the whole proof, we use the Pochhammer symbols $x^{(n)}$ and $(x)_n$ for respectively the nth power of the increasing factorial of x, and the nth power of the decreasing factorial of x. Once we took care of subtlety coming with section 2.6.1, the rest of the proof is identical to the proof of Proposition A.1 in Favaro et al. (2012), which we resume here for the sake of completeness. According to section 2.6.1 it is enough to check that

$$(Q_p(A_1), \dots, Q_p(A_k)) \xrightarrow{d} (Q(A_1), \dots, Q(A_k)), \tag{2.15}$$

for any collection of disjoints bounded measurable sets $A_1, \dots, A_k \in \mathcal{A}$, where Q is a symmetric Gamma random measure with parameters $\alpha F(\cdot), \eta$. Oviously, for any vector $(v_1, \dots, v_k) \in \mathbb{R}^k$ the random variable $v_1 Q(A_1) + \cdots + v_k Q(A_k)$ has symmetric Gamma distribution, and hence is determined by its moments (because of proposition 2.10), by Billingsley (2008, Theorem 30.2) the equation (2.15) holds if

$$\mathbb{E}\left[Q_p(A_1)^{r_1} \dots Q_p(A_k)^{r_k}\right] \longrightarrow \mathbb{E}\left[Q(A_1)^{r_1} \dots Q(A_k)^{r_k}\right] \tag{2.16}$$

holds for any disjoints bounded measurable sets $A_1, \dots, A_k \in \mathcal{A}$ and any positive integers $r_1, \dots, r_k$. From now, for all collection of measurable sets $A_1, \dots, A_k \in \mathcal{A}$, we set $A^c := \mathcal{X} \setminus \cup_{i=1}^k A_i$. We recall that if $\{X_i \in \mathcal{X} : i \leq 1 \leq p\}$ is a Pólya urn sequence with base distribution $\alpha F(\cdot)$, and $A_1, \dots, A_k \in \mathcal{A}$ are disjoints, then

$$\begin{aligned} P(\#\{i : X_i \in A_1\} = j_1, \dots, \#\{i : X_i \in A_k\} = j_k) \\ = \binom{p}{j_1 \dots j_k} \frac{(\alpha F(A_1))^{(j_1)} \dots (\alpha F(A_k))^{(j_k)} (\alpha F(A^c))^{(p - \sum_{i=1}^k j_i)}}{(p - \sum_{i=1}^k j_i)!\, \alpha^{(p)}}, \end{aligned}$$

where $(j_1, \dots, j_k) \in \mathcal{E}_{k,p}$, with $\mathcal{E}_{k,p} := \{(j_1, \dots, j_k) \in \{0, \dots, p\}^k : \sum_{i=1}^k j_i \leq p\}$. It is straightforward to show that both the lhs and the rhs of equation (2.16) are null whenever one of the r_i's is odd. Therefore we shall only consider equation (2.16) for even exponents. We deduce from proposition 2.10 that for any disjoints bounded measurable sets $A_1, \dots, A_k \in \mathcal{A}$ and any positive integers $r_1, \dots, r_k$,

$$\begin{aligned} \mathbb{E}\left[Q_p(A_1)^{2r_1} \dots Q_p(A_k)^{2r_k}\right] = \alpha^{(r_1 + \cdots + r_k)} \left(\prod_{i=1}^k \frac{(2r_i)!/r_i!}{(\sqrt{\eta})^{2r_i} p^{r_i}}\right) \\ \times \sum_{(j_1, \dots, j_k) \in \mathcal{E}_{k,p}} \binom{p}{j_1 \dots j_k} \frac{(\alpha F(A_1))^{(j_1)} \dots (\alpha F(A_k))^{(j_k)} (\alpha F(A^c))^{(p - \sum_{i=1}^k j_i)}}{(p - \sum_{i=1}^k j_i)!\, \alpha^{(p)}} \\ \times (j_1)^{(r_1)} \dots (j_k)^{(r_k)}. \end{aligned}$$

Introducing $s(\cdot,\cdot)$ and $S(\cdot,\cdot)$ are the Stirling numbers of the first and second kind, we can mimic Favaro et al. (2012, Appendix A.1) to find that

$$\mathbb{E}\left[Q_p(A_1)^{2r_1}\dots Q_p(A_k)^{2r_k}\right] = \alpha^{(r_1+\dots+r_k)}\left(\prod_{i=1}^{k}\frac{(2r_i)!/r_i!}{(\sqrt{\eta})^{2r_i}p^{r_i}}\right) \times \sum_{m_1=0}^{r_1}|s(r_1,m_1)|\sum_{s_1=0}^{m_1}S(m_1,s_1)\cdots\sum_{m_k=0}^{r_k}|s(r_k,m_k)|\sum_{s_k=0}^{m_k}S(m_k,s_k) \times \frac{(\alpha F(A_1))^{(s_1)}\dots(\alpha F(A_k))^{s_k}}{\alpha^{(s_1+\dots+s_k)}}(p)_{s_1+\dots+s_k}.$$

Therefore, we conclude that,

$$\lim_{p\to\infty}\mathbb{E}\left[Q_p(A_1)^{2r_1}\dots Q_p(A_k)^{2r_k}\right] = \prod_{i=1}^{k}\left(\frac{(2r_i)!}{r_i!}\frac{(\alpha F(A_i))^{(r_i)}}{(\sqrt{\eta})^{2r_i}}\right). \qquad \blacksquare$$

Proof of proposition 2.1. We can assume that all the Q_p and Q are defined on the same probability space $(\Omega,\mathcal{F},\mathbb{P})$. The proof is an adaption of Favaro et al. (2012, Theorem 2). We just have to take care that here, we proved $Q_p \to Q$ vaguely in theorem 2.1, which does not necessarily imply that $f^{(Q_p)}(x) \to f^{(Q)}(x)$ pointwise. But by assumption, $x \mapsto K(x;y)$ is continuous and vanishes outside a compact set, and it is easily seen that the sequence of total mass $|Q|(\cdot;\omega)$ is almost-surely bounded, then by Bauer (2001, Theorem 30.6) (which remains valid for signed measures), we have $f^{(Q_p)}(x) \to f^{(Q)}(x)$ pointwise, almost-surely. The end of the proof is identical to Favaro et al. (2012, Theorem 2) for convergence in L^1, and extension to L^q with $1 \le q < +\infty$ is straightforward. $\blacksquare$

2.7. Proofs of section 2.5.3

Proof of theorem 2.2. The proof is similar to Ghosal and van der Vaart (2007b, theorem 5). The event A_n that $\int\prod_{i=1}^{n}\frac{dP_{\theta,i}}{dP_{\theta_0,i}}(Y_i)\,d\Pi(\theta) \ge \mathrm{e}^{-2n\epsilon_n^2/2}$ satisfies $P_{\theta_0}^n(A_n^c)\to 0$ by Lemma 10 in Ghosal and van der Vaart (2007a) and assumptions on Π. Therefore,

$$\begin{aligned} P_{\theta_0}^n\Pi(\Theta_n^c|Y_1,\dots,Y_n) &\le P_{\theta_0}^n[\Pi(\Theta_n^c|Y_1,\dots,Y_n)\,\mathbb{1}_{A_n}] + P_{\theta_0}^n(A_n^c) \\ &\le e^{2n\epsilon_n^2}P_{\theta_0}^n\int_{\Theta_n^c}\prod_{i=1}^{n}\frac{dP_{\theta,i}}{dP_{\theta_0,i}}(Y_i)\,d\Pi(\theta) + P_{\theta_0}^n(A_n^c) \\ &\le e^{2n\epsilon_n^2}\Pi(\Theta_n^c) + P_{\theta_0}^n(A_n^c) \to 0, \end{aligned}$$

where the last lines follows by Fubini's theorem. For $0<\alpha_j\le 1$, and n large enough, the lemma 2.2 states the existence of tests functions $\psi_{n,j}$ such that

$$P_{\theta_0}^n\psi_{n,j} \le 2\alpha_j N(M\epsilon_n,\Theta_{n,j},\rho_n)\,\mathrm{e}^{-KM^2n\epsilon_n^2}, \qquad P_\theta^n(1-\psi_{n,j}) \le \alpha_j^{-1}\,\mathrm{e}^{-KM^2n\epsilon_n^2},$$

for all $\theta \in \Theta_{n,j}$ with $\rho_n(\theta,\theta_0) > 12M\epsilon_n$. Letting $U_\epsilon := \{\theta \in \Theta : \rho_n(\theta_0,\theta) > 12M\epsilon_n\}$,

$$\begin{aligned}
&P_{\theta_0}^n[\Pi(U_\epsilon \cap \Theta_{n,j}|Y_1,\ldots,Y_n)\,\mathbb{1}_{A_n}] \\
&\qquad \le P_{\theta_0}^n\psi_{n,j} + P_{\theta_0}^n\left((1-\psi_{n,j})\int_{U_\epsilon\cap\Theta_{n,j}}\prod_{i=1}^n \frac{dP_{\theta,i}}{dP_{\theta_0,i}}(Y_i)\,d\Pi(\theta)\right)\mathrm{e}^{2n\epsilon_n^2} \\
&\qquad \le P_{\theta_0}^n\psi_{n,j} + \sup_{U_\epsilon\cap\Theta_{n,j}} P_\theta^n(1-\psi_{n,j})\Pi(\Theta_{n,j})\,\mathrm{e}^{2n\epsilon_n^2} \\
&\qquad \le 2\alpha_j N(M\epsilon_n,\Theta_{n,j},\rho_n)\,\mathrm{e}^{-KM^2n\epsilon_n^2} + \alpha_j^{-1}\Pi(\Theta_{n,j})\,\mathrm{e}^{-(KM^2-2)n\epsilon_n^2},
\end{aligned}$$

where we used Fubini's theorem again. Put $\alpha_j = \sqrt{\Pi(\Theta_{n,j})/N(M\epsilon_n,\Theta_{n,j},\rho_n)}$ (notice that $\alpha_j \le 1$) and sum over j to obtain the result in view of the last equation. ■

2.7.1. Existence of tests

Here we construct the test functions required in the proof of theorem 2.2. We proceed in two steps. First, we construct tests for testing the hypothesis that $\theta = \theta_0$ against θ belongs to a ball of radius $\epsilon/12$ centered at θ_1 with $\rho_n(\theta_0,\theta_1) > \epsilon$; then in lemma 2.2 we construct the tests used in the proof of theorem 2.2.

Let $\theta_0 = (f_0,\sigma_0)$, $\theta_1 = (f_1,\sigma_1)$, $\theta_{10} = (f_1,\sigma_0)$, $\delta = \sqrt{\epsilon^2 + (108/n)\log(1/\alpha)}$, and define,

$$\begin{aligned}
A_n &:= \left\{y \in \mathbb{R}^n : \sum_{i=1}^n \log\frac{dP_{\theta_0,i}}{dP_{\theta_{10},i}}(y_i) < -\frac{n\epsilon^2}{96\sigma_0^2} + 2\log\alpha\right\}, \\
B_n^c &:= \left\{y \in \mathbb{R}^n : n(1-\delta/3) \le \sum_{i=1}^n\left(\frac{y_i - f_0(x_i)}{\sigma_0}\right)^2 \le n(1+\delta/3)\right\}.
\end{aligned}$$

Then we construct the sequence $(\phi_n)_{n\ge0}$ as

$$\begin{aligned}
&\phi_n(Y_1,\ldots,Y_n) \\
&\qquad := \mathbb{1}_{A_n}(Y_1,\ldots,Y_n) + \mathbb{1}_{B_n}(Y_1,\ldots,Y_n) - \mathbb{1}_{A_n}(Y_1,\ldots,Y_n)\,\mathbb{1}_{B_n}(Y_1,\ldots,Y_n).
\end{aligned}$$

Proposition 2.3 – *Let $K = 3(32 \vee 4\sigma_0^2)^{-1}$. The tests ϕ_n defined above satisfy $P_{\theta_0}^n\phi_n \le \mathrm{e}^{-Kn\epsilon^2/144}$ and $\sup_{\theta\in\Theta\,:\,\rho_n(\theta,\theta_1)<\epsilon/12} P_\theta^n(1-\phi_n) \le \mathrm{e}^{-Kn\epsilon^2/144}$ for all $\theta_1 \in \Theta$ such that $\rho_n(\theta_0,\theta_1) > \epsilon$ and all $0 < \epsilon \le 1$.*

Proof. *Type I error of ϕ_n.* It is clear that $P_{\theta_0}^n\phi_n \le P_{\theta_0}^n(A_n) + P_{\theta_0}^n(B_n)$. Moreover, by proposition 4 in Birgé (2006), we have $P_{\theta_0}^n(A_n) \le \alpha\,\mathrm{e}^{-n\epsilon^2/(192\sigma_0^2)}$, and regarding the proof of lemma 7 in Choi and Schervish (2007), the bound $P_{\theta_0}^n(B_n) \le 2\,\mathrm{e}^{-n\delta^2/108} = 2\alpha\,\mathrm{e}^{-n\epsilon^2/108}$ holds for n sufficiently large.

Type II error of ϕ_n. Let $\theta = (f,\sigma)$ be such that $\rho_n(\theta,\theta_1) \le \epsilon/12$. Clearly, $P_\theta^n(1-\phi_n) = P_\theta^n(1-\mathbb{1}_{A_n})(1-\mathbb{1}_{B_n}) \le P_\theta^n(A_n^c) \wedge P_\theta^n(B_n^c)$. We should consider two situations, either $|\log\sigma_0 - \log\sigma_1| \le \epsilon/2$, or $|\log\sigma_0 - \log\sigma_1| > \epsilon/2$.

- If $|\log\sigma_0 - \log\sigma_1| \le \epsilon/2$, then $\rho_n(\theta_0,\theta_1) > \epsilon$ implies $\|f_0 - f_1\|_{2,n} > \epsilon/2$, and for all θ with $\rho_n(\theta,\theta_1) \le \epsilon/12$, it is clear that $\|f - f_1\|_{2,n} \le \epsilon/12$. It follows from

proposition 4 in Birgé (2006) that

$$P_\theta^n(A_n^c) \le \exp\left[-\frac{n\|f_0 - f_1\|_{2,n}^2 - n\epsilon^2/8 + 24\sigma_0^2 \log\alpha}{24\sigma_0^2}\right] \le \frac{1}{\alpha}\exp\left[-\frac{n\epsilon^2}{64\sigma_0^2}\right].$$

- If $|\log\sigma_0 - \log\sigma_1| > \epsilon/2$, then $\rho_n(\theta,\theta_1) \le \epsilon/12$ implies $|\log\sigma - \log\sigma_0| > 5\epsilon/12$. We should again subdivise this case, considering either $\sigma/\sigma_0 \ge 1$ or not. For both cases we mimick and adapt the proof of lemma 7 in Choi and Schervish (2007).
  - If $\sigma/\sigma_0 \ge 1$, because $|\log\sigma - \log\sigma_0| > \epsilon/3$ we have $\sigma > \sigma_0\,\mathrm{e}^{\epsilon/3}$, and thus $\sigma > (1+\epsilon/3)\sigma_0$ for any $\epsilon > 0$. Let $W \sim \chi_n^2$ and let W' have a noncentral χ^2 distribution with n degrees of freedom and noncentrality parameter $\sum_{i=1}^n (f(x_i) - f_0(x_i))^2$. Then,

$$\begin{aligned} P_\theta^n(B_n^c) &\le P_\theta^n\left(\sum_{i=1}^n \left(\frac{Y_i - f_0(x_i)}{\sigma_0}\right)^2 \le n\left(1+\frac{\delta}{3}\right)\right) \\ &= \mathbb{P}\left(W' \le n\frac{\sigma_0^2}{\sigma^2}\left(1+\frac{\delta}{3}\right)\right) \le \mathbb{P}\left(W \le n\frac{\sigma_0^2}{\sigma^2}\left(1+\frac{\delta}{3}\right)\right). \end{aligned}$$

    But whenever $0 < \alpha \le 2$, we have

$$\frac{\sigma_0^2}{\sigma^2}\left(1+\frac{\delta}{3}\right) \le \frac{1+\delta/3}{(1+\epsilon/3)^2} \le \frac{1}{1+\epsilon/3} + \frac{(108/n)\log(1/\alpha)}{6\epsilon(1+\epsilon/3)^2}.$$

    Therefore, by Markov's inequality we get for all $t < 1/2$

$$P_\theta^n(B_n^c) \le \exp\left\{-t\frac{108\log(1/\alpha)}{6\epsilon(1+\epsilon/3)^2}\right\}\exp\left\{-\frac{nt}{1+\epsilon/3}\right\}(1-2t)^{-n/2}.$$

    Choosing $t = -\epsilon/18$ leads to

$$\begin{aligned} P_\theta^n(B_n^c) &\le \exp\left\{\frac{\log(1/\alpha)}{(1+\epsilon/3)^2}\right\}\exp\left\{\frac{n}{2}\left(\frac{\epsilon/9}{1+\epsilon/3} - \log(1+\epsilon/9)\right)\right\} \\ &\le \frac{1}{\alpha}\exp\left\{-\frac{7n\epsilon^2}{648}\right\} \le \frac{1}{\alpha}\exp\left\{-\frac{n\epsilon^2}{93}\right\}, \end{aligned}$$

    because we have $0 < \epsilon \le 1$. This concludes the proof when $\sigma/\sigma_0 \ge 1$.
  - On the other direction, $\sigma/\sigma_0 < 1$ and $|\log\sigma - \log\sigma_0| > 5\epsilon/12$ imply that $\sigma < (1-\epsilon/3)\sigma_0$ for any $0 < \epsilon \le 1$. Using the same strategy as in the previous item it is possible to show that the bound $P_\theta^n(B_n^c) \le (1/\alpha)\,\mathrm{e}^{-n\epsilon^2/1536}$ holds. ■

Lemma 2.2 – *Let $\Theta_n \subset \Theta$ and $K := 3(32 \vee 4\sigma_0^2)^{-1}$. Then for any $0 < \alpha \le 1$ there exists a collection of tests functions $(\psi_n)_{n\ge1}$ such that for any $0 < \epsilon \le 1/12$ and any $n \ge 1$*

$$P_{\theta_0}^n\psi_n \le 2\alpha N(\epsilon,\Theta_n,\rho_n)\,\mathrm{e}^{-Kn\epsilon^2}, \qquad \sup_{\theta\in\Theta_n\,:\,\rho_n(\theta,\theta_0)>12\epsilon} P_\theta^n(1-\psi_n) \le \alpha^{-1}e^{-Kn\epsilon^2}.$$

Proof. Let $N \equiv N(\epsilon/12, \Theta_n, \rho_n)$ denote the number of balls of radius $\epsilon/12$ needed to cover Θ_n. Let $(B_1, \ldots, B_N)$ denote the corresponding covering and $(\zeta_1, \ldots, \zeta_N)$ denote the centers of $(B_1, \ldots, B_N)$. Now let J be the index set of balls B_j with $\rho_n(\theta_0, \zeta_j) > \epsilon$. Using proposition 2.3 for $0 < \epsilon \leq 1$ and for any ball B_j with $j \in J$, we can build a test function $\phi_{n,j}$ satisfying

$$P_{\theta_0}^n \phi_{n,j} \leq 2\alpha \, \mathrm{e}^{-Kn\epsilon^2/144}, \qquad \sup_{\theta \in B_j} P_\theta^n (1 - \phi_{n,j}) \leq \alpha^{-1} \, \mathrm{e}^{-Kn\epsilon^2/144},$$

Let $\psi_n := \max_{j \in J} \phi_{n,j}$. Then $P_{\theta_0}^n \psi_n \leq \sum_{j \in J} P_{\theta_0}^n \phi_{n,j} \leq 2\alpha N(\epsilon/12, \Theta_n, \rho_n) \, \mathrm{e}^{-Kn\epsilon^2/144}$ and also $P_\theta^n (1 - \psi_n) \leq \min_{j \in J} \sup_{\theta' \in B_j} P_{\theta'}^n (1 - \phi_{n,j}) \leq \alpha^{-1} \, \mathrm{e}^{-Kn\epsilon^2/144}$ for any $\theta \in \Theta_n$ with $\rho_n(\theta, \theta_0) > \epsilon$. ■

2.8. Proof of theorem 2.3

We prove theorem 2.3 by verifying the set of sufficient conditions established in theorem 2.2.

2.8.1. Sieve construction

For constants $H, M > 0$ to be determined later, we define the sets

$$\mathcal{D}_n := \Big\{ A \in \mathcal{E}_s \, : \, n^{-1/a_2} \leq \lambda_i(A) \leq n^{-1/a_2} (1 + M\epsilon_n/n)^{n^2}, \quad i = 1, \ldots, d \Big\},$$

$$\Theta_n := \left\{ (f, \sigma) \, : \, \begin{array}{l} n^{-2/a_8} < \sigma^2 \leq n^{-2/a_8} (1 + M\epsilon_n)^n, \; f(x) = \int K_A(x - \mu) Q(dA d\mu), \\ Q = \sum_{i=1}^\infty u_i \delta_{A_i, \mu_i}, \; \operatorname{supp} Q = \mathcal{E}_s \times [-2S, 2S]^d, \; \sum_{i=1}^\infty |u_i| \leq n, \\ \#\{ i \, : \, |u_i| > n^{-1}, \; A_i \in \mathcal{D}_n \} \leq Hn\epsilon_n^2 / \log n, \\ \sum_{i=1}^\infty |u_i| \, \mathbb{1}\{A_i \notin \mathcal{D}_n\} \leq M\epsilon_n, \quad \sum_{i=1}^\infty |u_i| \, \mathbb{1}\{|u_i| \leq n^{-1}\} \leq M\epsilon_n \end{array} \right\}.$$

In the sequel, we assume without loss of generality that the jumps of Q in the definition of Θ_n are ordered so that there is no jump with $|u_i| > n^{-1}$ and $A_i \in \mathcal{D}_n$ when $i > Hn\epsilon_n^2 / \log n$. Moreover, we consider the following partition of Θ_n. Let H_n the largest integer smaller than $Hn\epsilon_n^2 / \log n$. Then for any $j = (j_1, \ldots, j_{H_n}) \in \{1, 2, \ldots\}^{H_n}$, inspired by Canale and De Blasi (2017, theorem 2) we define the slices

$$\Theta_{n,j} := \Big\{ (f, \sigma) \in \Theta_n \, : \, n^{2^{j_i - 1}} < \lambda_1(A_i)/\lambda_d(A_i) \leq n^{2^{j_i}} \quad \forall i \leq H_n \Big\}.$$

Lemma 2.3 – *Assume that there is* $0 < \gamma_1 < 1$ *such that* $\epsilon_n^2 \geq n^{-\gamma_1}$ *for all* n *large enough. Then for* $H = 6(1 - \gamma_1)^{-1}$ *it holds* $\Pi(\Theta_n^c) \lesssim \exp(-3n\epsilon_n^2)$ *as* $n \to \infty$.

Proof. From the definition of Θ_n, it is clear that

$$\begin{aligned} \Pi(\Theta \backslash \Theta_n) \leq \Pi\Big(\#\{ i \, : \, |u_i| > n^{-1} \} > Hn\epsilon_n^2 / \log n \Big) + \Pi\left(\textstyle\sum_{i=1}^\infty |u_i| > n \right) \\ + \Pi\left(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{|u_i| \leq n^{-1}\} > M\epsilon_n \right) + \Pi\left(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{(A_i, \mu_i) \notin \mathcal{D}_n\} > M\epsilon_n \right) \\ + P^\sigma(\sigma^2 \leq n^{-2/a_8}) + P^\sigma(\sigma^2 > n^{-2/a_8}(1 + M\epsilon_n)^n). \end{aligned} \tag{2.17}$$

The bounds on the two last terms are obvious in view of equations (2.10) and (2.11).

By the superposition theorem (Kingman, 1992, section 2), for any measurable set $A \subseteq \mathcal{E} \times \mathbb{R}^d$ we have $Q(A) := Q_1(A) + Q_2(A)$ where Q_1 and Q_2 are independent signed random measures with total variation having Laplace transforms (for all measurable $A \subseteq \mathcal{E} \times \mathbb{R}^d$ and all $t \in \mathbb{R}$ for which the integrals in the expression converge)

$$\mathbb{E}\, \mathrm{e}^{t|Q_1|(A)} = \exp\left\{2\alpha F(A) \int_{n^{-1}}^{\infty} (\mathrm{e}^{tx} - 1) x^{-1}\, \mathrm{e}^{-\eta x}\, dx\right\}, \tag{2.18}$$

$$\mathbb{E}\, \mathrm{e}^{t|Q_2|(A)} = \exp\left\{2\alpha F(A) \int_{0}^{n^{-1}} (\mathrm{e}^{tx} - 1) x^{-1}\, \mathrm{e}^{-\eta x}\, dx\right\}. \tag{2.19}$$

The random measures Q_1 and Q_2 are almost-surely purely atomic, the magnitudes of the jumps of Q_1 are all $\geq n^{-1}$, whereas Q_2 has jumps magnitudes all $< n^{-1}$ (almost-surely). Also, the number of jumps of Q_1 is distributed according to a Poisson law with intensity $\alpha E_1(n^{-1}/\eta)$, where E_1 is the exponential integral E_1 function. Recalling that $E_1(x) \asymp \gamma + \log(1/x)$ for x small, it follows $\alpha(\gamma + \log \eta) \leq \alpha E_1(n^{-1}/\eta) \leq 2\alpha \log n \ll H n \epsilon_n^2 / \log n$ when n is large. Then using Chernoff's bound on Poisson law, we get

$$\begin{aligned}
&\Pi\left(\#\{i : |u_i| > n^{-1}\} > H n \epsilon_n^2 / \log n\right) \\
&\qquad \leq \mathrm{e}^{-\alpha E_1(n^{-1}/\eta)} \frac{(\mathrm{e}\, \alpha E_1(n^{-1}/\eta)^{H n \epsilon_n^2 / \log n}}{(H n \epsilon_n^2 / \log n)^{H n \epsilon_n^2 / \log n}} \\
&\qquad \leq (\eta e^{\gamma})^{\alpha} \exp\left\{-\frac{H n \epsilon_n^2}{\log n}\left(\log \frac{H n \epsilon_n^2}{\log n} - \log(2e\alpha \log n)\right)\right\}.
\end{aligned}$$

But,

$$\log \frac{H n \epsilon_n^2}{\log n} - \log(2e\alpha \log n) \geq (1 - \gamma_1) \log n - 2 \log\log n + \log \frac{H}{2e\alpha},$$

which is in turn greater than $(1/2)(1 - \gamma_1) \log n$ when n becomes large. This gives the proof for the first term of the rhs of equation (2.17).

Regarding the second term of the rhs of equation (2.17), it suffices to remark that the random variable $\sum_{i=1}^{n} |u_i|$ has Gamma distribution with parameters $(2\alpha, \eta)$. Then the upper bound on $\Pi(\sum_{i=1}^{n} |u_i| > n)$ follows from Markov's inequality. With the same argument, we have that the random variable $\sum_{i=1}^{\infty} |u_i|\, \mathbb{1}\{|u_i| \leq n^{-1}\}$ is equal in distribution to $|Q_2|(\mathcal{E} \times \mathbb{R}^d)$, thus the bound for the fourth term of the rhs of equation (2.17) follows from Markov's inequality and equation (2.19), because

$$\Pi\left(e^{3n\epsilon_n |Q_2|} > e^{3n\epsilon_n^2}\right) \leq \mathrm{e}^{-3n\epsilon_n^2} \exp\left\{2\alpha \int_{0}^{n^{-1}} (e^{n\epsilon_n x} - 1) x^{-1}\, \mathrm{e}^{-\eta x}\, dx\right\} \lesssim e^{-3n\epsilon_n^2}.$$

The fifth term of the rhs of equation (2.17) is bounded using Chebychev's inequality. Indeed, with the same argument as before, the random variable $X := \sum_{i=1}^{n} |u_i|\, \mathbb{1}\{A_i \notin \mathcal{D}_n\}$ has Gamma distribution with parameters $(2\alpha F_A(\mathcal{D}_n^c), \eta)$. Hence for n sufficiently large we have $\mathbb{E}X = 2\alpha F_A(\mathcal{D}_n^c)/\eta \leq \epsilon_n/2$, and

$$\Pi(X > \epsilon_n) \leq \Pi(X - \mathbb{E}X > \epsilon_n/2) \leq \frac{8\alpha F_A(\mathcal{D}_n^c)}{\eta^2 \epsilon_n^2}.$$

Then the result follows from equations (2.6) and (2.7). ■

Lemma 2.4 – *Let $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$ and $K = 3(32 \vee 4\sigma_0^2)$. Then there exists $M > 0$ such that it holds $\sum_j \sqrt{N(M\epsilon_n, \Theta_{n,j}, \rho_n)}\sqrt{\Pi(\Theta_{n,j})}\, \mathrm{e}^{-(KM^2-2)n\epsilon_n^2} \to 0$.*

Proof. Define the random measures Q_1 and Q_2 as in the proof of lemma 2.3. Then using the Poisson construction of Q_1 (see for instance Wolpert et al. (2011, section 2.3.1)), it follows from equation (2.9) that for any $j \in \{1, 2, \ldots\}^{H_n}$

$$\Pi(\Theta_{n,j}) \leq \prod_{i \leq H_n} F_A(A \ : \ \lambda_1(A)/\lambda_d(A) \geq n^{2^{j_i - 1}}) \leq b_6^{H_n} \prod_{i \leq H_n} n^{-\kappa^* 2^{j_i - 1}}.$$

Moreover, using proposition 2.4 we can find a constant $C > 0$ independent of M such that $N(M\epsilon_n, \Theta_{n,j}, \rho_n) \leq \mathrm{e}^{-2CHn\epsilon_n^2}\, n^{d(d-1)/2 \sum_{i \leq H_n} 2^{j_i}}$ when n is large. Therefore,

$$\sqrt{N(M\epsilon_n, \Theta_{n,j}, \rho_n)}\sqrt{\Pi(\Theta_{n,j})} \leq \exp\left\{Hn\epsilon_n^2 \left(C + \frac{\log b_6}{2 \log n}\right)\right\} \prod_{i \leq H_n} n^{\frac{1}{2}[d(d-1) - \kappa^*] 2^{j_i - 1}}.$$

For n large enough we have $\log b_6 \leq 2C \log n$; then provided $\kappa^* > d(d-1)$, we can sum over $j \in \{1, 2, \ldots\}^{H_n}$ the last expression to get

$$\begin{aligned} \sum_j \sqrt{N(M\epsilon_n, \Theta_{n,j}, \rho_n)}\sqrt{\Pi(\Theta_{n,j})} &\leq \exp\left\{2CHn\epsilon_n^2\right\} \left(\sum_{k \geq 1} n^{\frac{1}{2}[d(d-1) - \kappa^*] 2^{k-1}}\right)^{H_n} \\ &\leq \exp\left\{H(2C + \kappa^*/2) n\epsilon_n^2\right\}. \end{aligned}$$

Now choose $M > 0$ satisfying $KM^2 > 2 + H(2C + \kappa^*/2)$ to obtain the conclusion of the lemma. ■

Proposition 2.4 – *For n large enough there is a constant $C > 0$ independent of M such that for any sequence $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$, the following holds for any $j \in \{1, 2, \ldots\}^{H_n}$.*

$$\log N(M\epsilon_n, \Theta_{n,j}, \rho_n) \leq CHn\epsilon_n^2 + \frac{d(d-1)}{2} \log n \sum_{i \leq H_n} 2^{j_i}.$$

Proof. The proof is based on arguments from Shen et al. (2013), it uses the fact that the covering number $N(M\epsilon_n, \Theta_n, \rho_n)$ is the minimal cardinality of an $M\epsilon_n$-net over Θ_n in the distance ρ_n. Let $\delta_n := M\epsilon_n n^{-(1+1/a_2)}$, $\widehat{R}_n$ be a δ_n-net of $[-2S, 2S]^d$, $\widehat{\Delta}_n$ be a $M\epsilon_n$-net of $\{(u_1, \ldots, u_{H_n}) \in \mathbb{R}^{H_n} \ : \ \sum_{i=1}^{H_n} |u_i| \leq n\}$ in the ℓ_1-distance, and $\widehat{S}_n := \{\sigma > 0 \ : \ \sigma^2 = n^{-2/a_8}(1 + M\epsilon_n)^k, \ k \in \mathbb{N}, \ k \leq n\}$. Also, for any $k \geq 1$ let $\widehat{\mathcal{O}}_k$ be a $n^{-(2^k+1)}M\epsilon_n$-net of the group of $d \times d$ orthogonal matrices equipped with spectral norm $\|\cdot\|$, and define

$$\widehat{\mathcal{D}}_{n,k} := \left\{ A \in \mathcal{D}_n \ : \ \begin{array}{l} A = P\Lambda P^\top, \ P \in \widehat{\mathcal{O}}_k, \ \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \\ \lambda_j = n^{-1/a_2}(1 + M\epsilon_n/n)^k, \ k \in \mathbb{N}, \ k \leq n^2, \ j = 1, \ldots, d \end{array} \right\}.$$

Pick $(f, \sigma) \in \Theta_{n,j}$ with $f(x) = \sum_{i=1}^{\infty} u_i K_{A_i}(x - \mu_i)$. Clearly we can find $\widehat{u} \in \widehat{\Delta}$ such that $\sum_{i \leq H_n} |u_i - \widehat{u}_i| \leq M\epsilon_n$, $\widehat{\mu} \in \widehat{R}_n^{H_n}$ such that $|\mu_i - \widehat{\mu}_i|_d \leq \delta_n$ for all $i = 1, \ldots, H_n$, and $\widehat{\sigma} \in \widehat{S}_n$ such that $|\log \sigma - \log \widehat{\sigma}| \leq M\epsilon_n$. We also claim that we can find $\widehat{A}_i \in \widehat{\mathcal{D}}_{n,j_i}$ such that $\|I - A_i^{-1}\widehat{A}_i^{-1}\| \leq 3dM\epsilon_n/n$ for all $i \leq H_n$. We defer the proof of the claim to later. Let $\widehat{f}(x) = \sum_{i \leq H_n} \widehat{u}_i K_{\widehat{A}_i}(x - \widehat{\mu}_i)$ denote the function built from the parameters chosen

as above ; it follows

$$\begin{aligned}\|f-\widehat{f}\|_{2,n} &\le \sum_{i>H_n}|u_i| + \sum_{i\le H_n}|u_i-\widehat{u}_i| + \sum_{i\le H_n}|u_i|\|K_{A_i}(\cdot-\mu_i)-K_{\widehat{A}_i}(\cdot-\widehat{\mu}_i)\|_{2,n}\\ &\le 2M\epsilon_n + C'\sum_{i\le H_n}|u_i|\|I-A_i^{-1}\widehat{A}_i\| + C'\sum_{i\le H_n}|u_i|\|A_i^{-1}\|\|\mu_i-\widehat{\mu}_i|_d\\ &\le M(2+C'+3C'd)\epsilon_n,\end{aligned}$$

where the two last inequalities hold by proposition 2.12 for a constant $C'>0$ depending only on g, and because $\|A_i^{-1}\| \le n^{1/a_2}$ for all $i \le H_n$. Thus a $(2+C'+3C'd)M\epsilon_n$-net of $\Theta_{n,j}$ in the distance ρ_n can be constructed with $(\widehat{f},\widehat{\sigma})$ as above. Recall that $\#\widehat{R}_n \le (4S/\delta_n)^d$, $\#\widehat{\Delta}_n \le (n/(M\epsilon_n))^{H_n}$, $\#\widehat{S}_n = n$ and $\#\widehat{\mathcal{O}}_k \le (n^{-(2^k+1)}M\epsilon_n)^{-d(d-1)/2}$. It turns out that $\#\widehat{\mathcal{D}}_{n,k} \le n^{2d}\times\#\widehat{\mathcal{O}}_k$. Then the total number of $(\widehat{f},\widehat{\sigma})$ is bounded by a multiple constant of

$$n\times\left(\frac{4S}{\delta_n}\right)^{H_n}\times\left(\frac{n}{M\epsilon_n}\right)^{H_n}\prod_{i\le H_n}\left(n^{2d}\times\left(\frac{n^{2^{j_i}+1}}{M\epsilon_n}\right)^{d(d-1)/2}\right).$$

Finally, $H_n|\log M| \ll H_n\log n$ when n is large proving that the constant $C>0$ can be chosen independent of M, and the constant factor $2+C'+3C'd$ can be absorbed into the bound.

It remains to prove that for any $A\in\mathcal{D}_n$ with $\lambda_1(A)/\lambda_d(A)\le n^{2^k}$ we can find $\widehat{A}\in\widehat{\mathcal{D}}_{n,k}$ such that $\|I-A^{-1}\widehat{A}\|\le 3dM\epsilon_n/n$. Let $A =: P\Lambda P^\top$ denote the spectral decomposition of A (recall that A is symmetric). Clearly, we can find a matrix $\widehat{A} := \widehat{P}\widehat{\Lambda}\widehat{P}^\top$ in $\widehat{\mathcal{D}}_{n,k}$ with $\|P-\widehat{P}\|\le n^{-(2^k+1)}M\epsilon_n$ and $1\le\lambda_j(\Lambda)/\lambda_j(\widehat{\Lambda})\le 1+M\epsilon_n/n$ for all $j=1,\dots,d$. Let $\widetilde{A} := \widehat{P}\Lambda\widehat{P}^\top$ and remark that

$$\begin{aligned}\|I-A^{-1}\widehat{A}\| &\le \|I-A^{-1}\widetilde{A}\| + \|A^{-1}\widetilde{A}\|\|I-\widetilde{A}^{-1}\widehat{A}\|\\ &\le \|I-A^{-1}\widetilde{A}\| + \|I-\widetilde{A}^{-1}\widehat{A}\|(1+\|I-A^{-1}\widetilde{A}\|).\end{aligned}\tag{2.20}$$

Let $B := P^\top\widehat{P}-I$, so that $\|B\|_{\max}\le\|B\|\le\|P^\top\widehat{P}-I\|\le\|P-\widehat{P}\|\le n^{-(2^k+1)}M\epsilon_n$, and $I-A^{-1}\widetilde{A} = P(B-\Lambda^{-1}B\Lambda)\widehat{P}^\top$. It follows,

$$\|I-A^{-1}\widetilde{A}\|\le\|B-\Lambda^{-1}B\Lambda\|\le d\|B\|_{\max}\frac{\lambda_1(\Lambda)}{\lambda_d(\Lambda)}\le dM\epsilon_n/n,$$

because the entries of $B-\Lambda^{-1}B\Lambda$ are equal to $B_{ij}(1-\Lambda_j/\Lambda_i)$ and $\|\cdot\|\le d\|\cdot\|_{\max}$. Moreover, $I-\widetilde{A}^{-1}\widehat{A} = \widehat{P}(I-\Lambda^{-1}\widehat{\Lambda})\widehat{P}^\top$ implies $\|I-\Lambda^{-1}\widehat{\Lambda}\|\le dM\epsilon_n/n$. Then the conclusion follows from equation (2.20). ■

2.8.2. Approximation of functions

In order to prove the prior positivity of Kullback-Leibler balls around θ_0, we need to approximate $f_0\in\mathcal{C}^\beta[-S,S]^d$ by finite location-scale mixtures of kernels. We mostly follow the approach of de Jonge and van Zanten (2010, lemma 3.4).

Nevertheless, as mentioned in de Jonge and van Zanten (2010), we shall extend f_0 defined on $[-S,S]^d$ onto a (smooth) function defined on $\mathbb{R}^d$ to be able to approximate

properly f_0; otherwise we could have troubles at the boundaries of $[-S,S]^d$. Clearly, without any precaution, $h^{-d}K_{hI} * f_0(x) \to f_0(x)/2$ as $h \to 0$ when x belongs to the boundary of $[-S,S]^d$. De Jonge and van Zanten (2010) assume that the covariates are spread onto $[a,b]^d$ with $a > -S$ and $b < S$ and extend f_0 by multiplying it by a smooth function that equal 1 on $[a,b]^d$ and 0 outside $[-S,S]^d$. Here we assume that the covariates are spread onto $[-S,S]^d$ and we use Whitney's extension theorem (Whitney, 1934) to find a function $\widetilde{f}_0 : \mathbb{R}^d \to \mathbb{R}$ such that $\widetilde{f}_0 \in \mathcal{C}^\beta(\mathbb{R}^d)$ and $D^\alpha \widetilde{f}_0(x) = D^\alpha f_0(x)$ for all $x \in [-S,S]^d$ and all $|\alpha| \leq \beta$. Then we apply the method of de Jonge and van Zanten (2010, lemma 3.4) to $\widetilde{f}_0$. We find this approach more elegant since we do not have to assume that f_0 is defined on a larger set than the support of the covariates.

For each $\alpha \in \mathbb{N}^d$, let $\mathfrak{m}_\alpha^h := h^{-d} \int x^\alpha K_{hI}(x)\, dx$. For $\alpha \in \mathbb{N}^d$ with $|\alpha| \geq 1$, define two sequences of numbers by the following recursion. If $|\alpha| = 1$ set $c_\alpha = 0$ and $d_\alpha = -1/\alpha!$, and for $|\alpha| \geq 2$ define

$$c_\alpha := \sum_{\substack{l+k=\alpha \\ |l|\geq 1,\, |k|\geq 1}} \frac{(-1)^{|\alpha|}}{\alpha!} \left(\frac{\mathfrak{m}_l^h \mathfrak{m}_k^h}{\mathfrak{m}_\alpha^h} \right) d_k, \quad d_\alpha := \frac{(-1)^{|\alpha|}}{\alpha!} - c_\alpha. \tag{2.21}$$

Given $\beta > 0$, $h > 0$ and p the largest integer strictly smaller than β, define

$$f_\beta := \widetilde{f}_0 - \sum_{1 \leq |\alpha| \leq p} d_\alpha \mathfrak{m}_\alpha^h D^\alpha \widetilde{f}_0.$$

Proposition 2.5 – *Let $h > 0$. For any $\beta > 0$ and any function $f_0 \in \mathcal{C}^\beta[-S,S]^d$ there is a positive constant M_β such that $|h^{-d}K_{hI} * f_\beta(x) - f_0(x)| \leq M_\beta h^\beta$ for all $x \in [-S,S]^d$.*

Proof. Noticing that $\mathfrak{m}_\alpha^h \lesssim h^{|\alpha|}$, the proof follows from the same argument as in (Shen et al., 2013, lemma 2), because $\widetilde{f}_0(x) = f_0(x)$ for all $x \in [-S,S]^d$. ■

The proposition 2.5 shows that any sufficiently regular function can be approximated by continuous location mixtures of K_{hI}, provided h is chosen small enough and g has enough finite moments. In the sequel, we will need slightly more, that is approximating any β-Hölder continuous function by *discrete* mixtures of K_{hI} ; this is done by discretizing the convolution operator in the next proposition. Compared to Ghosal and van der Vaart (2001, lemma 3.1), we need to take extra cares regarding the fact that f_0 can take negative values, and also to control the "total mass" of the mixing measure.

Proposition 2.6 – *Let $h > 0$ be small enough and $\zeta = 1 \vee 2/(\tau - \gamma\tau)$. There exists a discrete mixture $f(x) = \sum_{i=1}^N \alpha_i K_{hI}(x - \mu_i)$ with $N \lesssim h^{-d}(\log h^{-1})^{d(\zeta-1)}$, $\mu_i \in [-2S, 2S]^d$ for all $i = 1, \ldots, N$; such that $|f(x) - f_0(x)| \lesssim h^\beta$ for all $x \in [-S,S]^d$. Moreover $\sum_{i=1}^N |\alpha_i| \lesssim h^{-d}$, and $|\mu_i - \mu_j|_d \geq h^{\beta+1}$ for any $i \neq j$.*

Proof. Let Q be the signed measure defined by $A \mapsto \int_A f_\beta(y)dy$ for any measurable set $A \subseteq \mathbb{R}^d$. Let $M_h := (C_0^{-1}(\beta + d) \log h^{-1})^{1/\tau}$. To any $j \in \mathbb{Z}^d$ we associate the cube $B_j := hM_h(j + [0,1]^d)$ and the signed measure Q_j such that $Q_j(A) := Q(A \cap B_j)$ for all measurable $A \subseteq \mathbb{R}^d$. Let Q_j^+, Q_j^- denote respectively the positive and negative part of the Jordan decomposition of Q_j. It is a classical result from Tchakaloff (1957) that we can construct discrete (positive) measures $P_{j,k}^+$, $P_{j,k}^-$ each having at most $(k+d)!/(k!d!)$ atoms and satisfying $\int R(x)Q_j^\pm(dx) = \int R(x)P_j^\pm(dx)$ for any polynomial $R(x)$ of degree

$|\alpha| \leq k$. Let $\Lambda_h := \{j \in \mathbb{Z}^d : |j| \leq 1+S/(hM_h)\}$ and for any $x \in \mathbb{R}^d$ let $N_x := \{j \in \Lambda_h : \inf\{|x-y|_d : y \in B_j\} \leq hM_h\}$. For the signed measure $P_k := \sum_{j\in\Lambda_h}(P_{j,k}^+ - P_{j,k}^-)$ the total variation of P_k satisfy the bound

$$|P_k| \leq \sum_{j\in\Lambda_h} P_{j,k}^+ + \sum_{j\in\Lambda_h} P_{j,k}^- \leq \sum_{j\in\mathbb{Z}^d}(P_{j,k}^+ + P_{j,k}^-) = |Q|.$$

Notice that $|Q| < +\infty$ since we have $f_\beta \in L^1(\mathbb{R}^d)$. Moreover, letting $P_{j,k} = P_{j,k}^+ - P_{j,k}^-$

$$\begin{aligned}\int K_{hI}(x-y)(Q-P_k)(dy) = &\sum_{j\notin\Lambda_h}\int_{B_j} g\left(\frac{x-y}{h}\right)Q_j(dy)\\ &+\sum_{j\in\Lambda_h\setminus N_x}\int_{B_j} g\left(\frac{x-y}{h}\right)(Q_j-P_{j,k})(dy)\\ &+\sum_{j\in N_x}\int_{B_j} g\left(\frac{x-y}{h}\right)(Q_j-P_{j,k})(dy). \quad (2.22)\end{aligned}$$

By assumptions on g, for any $x \in [-S,S]^d$ the first term of the rhs of equation (2.22) is bounded by $|Q|h^{\beta+d}$. With the same argument, using the definition of N_x, the second term of the rhs of equation (2.22) is bounded by $2|Q|h^{\beta+d}$. Regarding the last term, using multivariate Taylor's formula we write

$$\begin{aligned}\int_{B_j} g\left(\frac{x-y}{h}\right)(Q_j-P_{j,k})(dy) = &\sum_{|\alpha|\leq k}\frac{D^\alpha g(0)}{\alpha!}\int_{B_j}\left(\frac{x-y}{h}\right)^\alpha(Q_j-P_{j,k})(dy)\\ &+\int_{B_j} R_k\left(\frac{x-y}{h}\right)(Q_j-P_{j,k})(dy), \quad (2.23)\end{aligned}$$

where $|R_k(x)| \leq \sup_{|\alpha|=k}|D^\alpha g(0)||x|_d^k/k!$. The first term of the rhs of equation (2.23) vanishes by construction of $P_{j,k}$. For any $j \in N_x$ and any $y \in B_j$ it holds $|x-y|_d \leq 2hM_h$; then using Stirling's formula and assumptions on $D^\alpha g$ the second term of the rhs of equation (2.23) is bounded by

$$\sup_{|\alpha|=k}|D^\alpha g(0)|\frac{(2eM_h)^k}{\sqrt{2\pi k}k^k}\int_{B_j}|Q_j-P_{j,k}|(dy) \leq K_1\exp\left\{-k(1-\gamma)\log k + k\log(2eM_h)\right\},$$

whenever $j \in N_x$, for a constant K_1 depending only on f_0, β and g. Therefore, choosing $k \geq (2eM_h)^{2/(1-\gamma)}$, we deduce from equations (2.22) and (2.23) that

$$\left|\int K_{hI}(x-y)(Q-P_k)(dy)\right| \leq 3|Q|h^{\beta+d} + K_1\exp\left\{-\frac{1-\gamma}{2}k\log k\right\}. \quad (2.24)$$

Now if $(2eM_h)^{2/(1-\gamma)} \geq 2(\beta+d)/(1-\gamma)\log h^{-1}$ set k to be the smaller integer larger than $(2eM_h)^{2/(1-\gamma)}$; otherwise set k to be the larger integer greater than $2(\beta+d)/(1-\gamma)\log h^{-1}$. This yields the first part of the proposition with $f(x) = h^{-d}\int K_{hI}(x-y)P_k(dy)$ because of equation (2.24), of proposition 2.5 and because each of the $P_{j,k}$ has a number of atoms proportional to $(\log h^{-1})^{d\zeta}$ by Tchakaloff's theorem, all in $[-2S,2S]^d$ if h is small enough.

It remains to prove the separation between the atoms of Q'_k. But the cost to the supremum norm of moving one μ_i of $h^{\beta+1}$ is proportional to h^β by proposition 2.12. Hence we can assume that the support point of Q'_k are chosen on a regular grid with $h^{\beta+1}$ separation within nodes (see also Shen et al. (2013, corollary B1)). ■

2.8.3. Kullback-Leibler property

A simple computation shows that (see for instance Choi and Schervish (2007)) for $\theta_0 = (f_0, \sigma_0)$ and $\theta = (f, \sigma)$,

$$K_i(\theta_0, \theta) = \log \frac{\sigma}{\sigma_0} - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{2}\frac{|f_0(x_i) - f(x_i)|^2}{\sigma^2},$$
$$V_{2;i}(\theta_0, \theta) = \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right)^2 + \frac{\sigma_0^4}{\sigma^4}|f_0(x_i) - f(x_i)|^2.$$

Therefore, for all $0 < \epsilon \leq 1/2$, there exists a constant $C_0 > 0$ (depending only on θ_0) such that one has the inclusions

$$K_n(\theta_0, \epsilon) \supseteq \Big\{(f, \sigma) \,:\, \|f - f_0\|_\infty^2 \leq C_0\epsilon^2,\ \sigma_0 \leq \sigma \leq \sigma_0(1 + C_0\epsilon^2)\Big\}, \tag{2.25}$$

hence probabilities of Kullback-Leibler balls around θ_0 are lower bounded by the probability of the sets defined in the rhs of equation (2.25). Now we state and prove the main result of this section.

Lemma 2.5 – *Let $f_0 \in \mathcal{C}^\beta[-S, S]^d$, and $\zeta > 1$ as in proposition 2.6. Then there exists a constant $C > 0$, not depending on n, such that $\Pi(K_n(\theta_0, \epsilon_n)) \gtrsim \exp(-n\epsilon_n^2)$ for $\epsilon_n^2 = Cn^{-2\beta/(2\beta+d+\kappa/2)}(\log n)^{2\beta d(\zeta-1)/(2\beta+d+\kappa/2)}$.*

Proof. By proposition 2.6 for any $h > 0$ sufficiently small, there is $N \lesssim h^{-d}(\log h^{-1})^{d(\zeta-1)}$ and a function $f_h(x) = \sum_{j=1}^N \alpha_j K_{hI}(x - \mu_j)$ such that $|f_h(x) - f_0(x)| \lesssim h^\beta$ for all $x \in [-S, S]^d$, with $\alpha_j \in \mathbb{R}$ for all $j = 1, \dots N$, $\mu_i \in [-2S, 2S]^d$ for all $i = 1, \dots, N$, and $|\mu_i - \mu_j|_d \geq h^{\beta+1}$ whenever $i \neq j$. Let define

$$\mathcal{E}_{s,h} := \Big\{A \in \mathcal{E}_s \,:\, h^{-1} \leq \lambda_i(A^{-1}) \leq h^{-1}(1 + h^{\beta+d}), \quad i = 1, \dots, d\Big\}.$$

We construct a partition of $\mathcal{E}_s \times [-2S, 2S]^d$ in the following way : for all $j = 1, \dots, N$, let U_j be the closed ball of radius $h^{\beta+d+1}$ centered at μ_j (observe that these balls are disjoint), and set $V_j := \mathcal{E}_{s,h} \times U_j$, $V^c := \mathcal{E}_s \times [-2S, 2S]^d \backslash U_{j=1}^N V_j$. Let $\mathcal{Q}$ denote the set of signed measures on $\mathcal{E}_s \times [-2S, 2S]^d$ satisfying $Q \in \mathcal{Q} \Rightarrow |Q(V_j) - \alpha_{h,j}| \leq h^\beta N^{-1}$ for all $j = 1, \dots, N$, and $|Q|(V^c) \leq h^\beta$. Notice that for any $Q \in \mathcal{Q}$ we have $|Q| \leq \sum_{j=1}^N |Q(V_j) - \alpha_{h,j}| + \sum_{j=1}^N |\alpha_{h,j}| \lesssim h^\beta + h^{-d} \lesssim h^{-d}$ because of proposition 2.6. Then for any $Q \in \mathcal{Q}$ and all $x \in [-S, S]^d$, using proposition 2.12,

$$\left|\int_{\mathcal{E}_s \times [-2S,2S]^d} K_A(x - \mu)\, Q(dAd\mu) - f_h(x)\right| \lesssim \sum_{j=1}^N |Q(V_j) - \alpha_{h,j}| + |Q|(V^c)$$
$$+ \sum_{j=1}^N \int_{V_j} |K_A(x - \mu) - K_{hI}(x - \mu_j)|\ |Q|(dAd\mu) \lesssim h^\beta.$$

Thus for all $Q \in \mathcal{Q}$ and all $x \in [-S, S]^d$, we have $|\int K_A(x - \mu)\, Q(dAd\mu) - f_0(x)| \leq |\int K_A(x - \mu)\, Q(dAd\mu) - f_h(x)| + |f_h(x) - f_0(x)| \leq K_1 h^\beta$ for a constant $K_1 > 0$ not

depending on h. By the assumptions of equations (2.5) and (2.8) we have for any $j = 1, \dots, N$

$$\begin{aligned}\alpha F_A(\mathcal{E}_{s,h}) F_\mu(U_j) &\geq \alpha b_1 b_4 h^{a_1(\beta+d+1)-a_4+a_5(\beta+d)} \exp(-C_3 h^{-\kappa/2}) \\ &=: K_2 h^q \exp(-C_3 h^{-\kappa/2}),\end{aligned}$$

where $q := a_1(\beta + d + 1) - a_4 + a_5(\beta + d)$ and the constant $K_2 > 0$ not depending on h.

For $h > 0$ sufficiently small, it is clear that $K_2 h^q \exp(-C_3 h^{-\kappa/2}) < F(V_j) \leq 1$ for all $j = 1, \dots, N$. We also assume without loss of generality that $K_2 h^q \exp(-C_3 h^{-\kappa/2}) \leq F(V^c) \leq 1$ and we set $V_{N+1} := V^c$, $\alpha_{h,N+1} := 0$; otherwise we subdivide V^c onto smaller subsets for which the relation is verified. Because F is a probability measure, this can be done with a finite number of subsets not depending on h. Now let $W := \{\sigma > 0 : \sigma_0 \leq \sigma \leq \sigma_0(1 + C_0\epsilon_n^2)\}$ and $\epsilon_n = C_0^{-1} K_1 h^\beta$. Notice that $P^\sigma(W) \geq K_3 \epsilon_n^{2a_9}$ with a constant $K_3 > 0$ eventually depending on θ_0. The sets $\mathcal{E}_{s,h} \times U_j$ are disjoint, hence by equation (2.12) and proposition 2.11 we deduce that there is a constant $K_4 > 0$ such that

$$\begin{aligned}\Pi(K_n(\theta_0, \epsilon_n)) \geq P^\sigma(W) \Pi_*(\mathcal{Q}) &\gtrsim \epsilon_n^{2a_9} \prod_{i=1}^{N+1} \left(\frac{h^\beta N^{-1} \, \mathrm{e}^{-(3+\eta)|\alpha_j|}}{\Gamma(\alpha(F_A(\mathcal{E}_h) F_\mu(U_j))} \right) \\ &\geq \exp\left\{ -K_4 h^{-(d+\kappa/2)} (\log h^{-1})^{d(\zeta-1)} \right\},\end{aligned}$$

where we used that $N \lesssim h^{-d} (\log h^{-1})^{d(\eta-1)}$, $\sum_{j=1}^N |\alpha_{h,j}| \lesssim h^{-d}$ and $\Gamma(x) \lesssim x^{-1}$ for $x > 0$ sufficiently small. This concludes the proof. ■

2.9. Proof of theorem 2.4

As in section 2.5, the proof of theorem 2.4 consists on verifying the condition established in theorem 2.3.

2.9.1. Sieve construction

For constants $H, M > 0$ to be determined later, we define

$$\Theta_n := \left\{ (f, \sigma) \; : \; \begin{array}{l} f(x) = \int K_{\xi,\phi}(x - \mu) Q(d\xi d\mu d\phi), \; \operatorname{supp} Q = \mathbb{R}^d \times [-2S, 2S]^d \times [0, \pi/2], \\ Q = \sum_{i=1}^\infty u_i \delta_{\xi_i, \mu_i, \phi_i}, \; n^{-2/a_8} < \sigma^2 \leq n^{-2/a_8} (1 + M\epsilon_n)^n \\ \sum_{i=1}^\infty |u_i| \leq n, \; \#\{i : |u_i| > n^{-1}, \; |\xi_i|_d \leq \mathrm{e}^{2n\epsilon_n^2}\} \leq H n \epsilon_n^2 / \log n, \\ \sum_{i=1}^\infty |u_i| \, \mathbb{1}\{|\xi_i|_d > \mathrm{e}^{2n\epsilon_n^2}\} \leq M\epsilon_n, \quad \sum_{i=1}^\infty |u_i| \, \mathbb{1}\{|u_i| \leq n^{-1}\} \leq M\epsilon_n \end{array} \right\}.$$

In the sequel, we assume without loss of generality that the jumps of Q in the definition of Θ_n are ordered so that there is no jump with $|u_i| > n^{-1}$ and $|\xi_i|_d \leq \mathrm{e}^{2n\epsilon_n^2}$ when $i > H n \epsilon_n^2 / \log n$. Moreover, we consider the following partition of Θ_n. Let H_n be the largest integer smaller than $H n \epsilon_n^2 / \log n$. Then for any $j = (j_1, \dots, j_{H_n}) \in \{1, 2, \dots\}^{H_n}$ we define the slices

$$\Theta_{n,j} := \{(f, \sigma) \in \Theta_n \; : \; \sqrt{n}(j_i - 1) \leq |\xi|_d < \sqrt{n} j_i, \quad \forall i \leq H_n\}.$$

Lemma 2.6 – *Assume that there is $0 < \gamma_1 < 1$ such that $\epsilon_n^2 \geq n^{-\gamma_1}$ for all n large enough. Then for $H = 6(1 - \gamma_1)^{-1}$ it holds $\Pi(\Theta_n^c) \lesssim \exp(-3n\epsilon_n^2)$ as $n \to \infty$.*

Proof. According to the proof of lemma 2.3, the result holds if $F_\xi(\xi \,:\, |\xi_d| \geq \mathrm{e}^{2n\epsilon_n^2}) \lesssim \epsilon_n^2 \exp(-3n\epsilon_n^2)$ for n sufficiently large. Then the conclusion follows from equation (2.13) because $\eta > 0$. ■

Lemma 2.7 – *Let $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$ and $K = 3(32 \vee 4\sigma_0^2)$. Then there exists $M > 0$ such that it holds $\sum_j \sqrt{N(M\epsilon_n, \Theta_{n,j}, \rho_n)}\sqrt{\Pi(\Theta_{n,j})}\,\mathrm{e}^{-(KM^2-2)n\epsilon_n^2} \to 0$.*

Proof. With the same argument as in lemma 2.4, it follows from equation (2.13) that for any $j \in \{1, 2, \ldots\}^{H_n}$

$$\Pi(\Theta_{n,j}) \leq \textstyle\prod_{i\leq H_n} F_\xi(\xi \,:\, |\xi|_d \geq \sqrt{n}(j_i - 1)) \leq b_{11}^{H_n} \prod_{i\leq H_n}(1 + \sqrt{n}(j_i - 1))^{-2(\eta+1)}.$$

Moreover, using proposition 2.7 we can find a constant $C > 0$ independent of M such that $N(M\epsilon_n, \Theta_{n,j}, \rho_n) \leq \exp(2CHn\epsilon_n^2) \prod_{i\leq H_n} j_i^{d-1}$ when n is large. Therefore, for those n

$$\begin{aligned}\sqrt{N(M\epsilon_n, \Theta_{n,j}, \rho_n)}&\sqrt{\Pi(\Theta_{n,j})} \\ &\leq \exp\left\{Hn\epsilon_n^2\left(C + \frac{\log b_{11}}{2\log n}\right)\right\} \prod_{i\leq H_n} j_i^{(d-1)/2}[1 + \sqrt{n}(j_i - 1)]^{-(\eta+1)}.\end{aligned}$$

For n large enough we have $\log b_{11} \leq 2C \log n$; then provided $\eta > (d-1)/2$, we can sum over $j \in \{1, 2, \ldots\}^{H_n}$ the last expression to get

$$\begin{aligned}\textstyle\sum_j \sqrt{N(M\epsilon_n, \Theta_{n,j}, \rho_n)}&\sqrt{\Pi(\Theta_{n,j})} \\ &\leq \exp\left\{2CHn\epsilon_n^2\right\}\left(\textstyle\sum_{k\geq 1} k^{(d-1)/2}[1 + \sqrt{n}(k-1)]^{-(\eta+1)}\right)^{H_n} \\ &\leq \exp\left\{2CHn\epsilon_n^2\right\}\left(1 + n^{-(\eta+1)/2}\textstyle\sum_{k\geq 1} k^{(d-1)/2-(\eta+1)}\right)^{H_n} \lesssim \exp\{3CHn\epsilon_n^2\},\end{aligned}$$

where the last inequality holds for n sufficiently large. Now choose $M > 0$ satisfying $KM^2 > 2 + 3CH$ to obtain the conclusion of the lemma. ■

Proposition 2.7 – *For n large enough there is a constant $C > 0$ independent of M such that for any sequence $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$, the following holds for any $j \in \{1, 2, \ldots\}^{H_n}$.*

$$\log N(M\epsilon_n, \Theta_{n,j}, \rho_n) \leq CHn\epsilon_n^2 + (d-1)\sum_{i\leq H_n} \log j_i.$$

Proof. The proof is similar to proposition 2.4. Let $\widehat{R}_n$ be a $(M\epsilon_n/n)$-net of $[-2S, 2S]^d$, $\widehat{\Delta}_n$ be a $M\epsilon_n$-net of $\{(u_1, \ldots, u_{H_n}) \in \mathbb{R}^{H_n} \,:\, \sum_{i=1}^{H_n} |u_i| \leq n\}$ in the ℓ_1-distance, $\widehat{S}_n := \{\sigma > 0 \,:\, \sigma^2 = n^{-2/a_8}(1 + M\epsilon_n)^k,\ k \in \mathbb{N},\ k \leq n\}$, $\widehat{U}_n$ be a $(M\epsilon_n/n)$-net of $[0, \pi/2]$, and for all $k = 1, \ldots, H_n$, let $\widehat{V}_{n,k}$ a $(M\epsilon_n/n)$-net of $\{\xi \in \mathbb{R}^d \,:\, \sqrt{n}(k-1) \leq |\xi|_d < \sqrt{n}k\}$. Pick $(f, \sigma) \in \Theta_{n,j}$ with $f(x) = \sum_{i=1}^\infty u_i K_{\xi_i,\phi_i}(x - \mu_i)$. Clearly we can find $\widehat{u} \in \widehat{\Delta}$ such that $\sum_{i\leq H_n} |u_i - \widehat{u}_i| \leq M\epsilon_n$, $\widehat{\mu} \in \widehat{R}_n^{H_n}$ such that $|\mu_i - \widehat{\mu}_i|_d \leq M\epsilon_n/n$ for all $i = 1, \ldots, H_n$, $\widehat{\phi} \in \widehat{U}_n^{H_n}$ such that $|\phi_i - \widehat{\phi}_i| \leq M\epsilon_n/n$ for all $i = 1, \ldots, H_n$, $\widehat{\xi}_i \in \widehat{V}_{n,j_i}$ such that $|\xi_i - \widehat{\xi}_i|_d \leq M\epsilon_n/n$ for all $i = 1, \ldots, H_n$, and $\widehat{\sigma} \in \widehat{S}_n$ such that $|\log\sigma - \log\widehat{\sigma}| \leq M\epsilon_n$. Let

$\widehat{f}(x) = \sum_{i \le H_n} \widehat{u}_i K_{\widehat{\xi}_i, \widehat{\phi}_i}(x - \widehat{\mu}_i)$ denote the function built from the parameters chosen as above ; it follows

$$\begin{aligned}\|f - \widehat{f}\|_{2,n} &\le \sum_{i > H_n} |u_i| + \sum_{i \le H_n} |u_i - \widehat{u}_i| + \sum_{i \le H_n} |u_i| \|K_{\xi_i,\phi_i}(\cdot - \mu_i) - K_{\widehat{\xi}_i,\widehat{\phi}_i}(\cdot - \widehat{\mu}_i)\|_{2,n} \\ &\le 2M\epsilon_n + C' \sum_{i \le H_n} |u_i||\xi_i - \widehat{\xi}_i|_d + C' \sum_{i \le H_n} |u_i||\mu_i - \widehat{\mu}_i|_d + C' \sum_{i \le H_n} |u_i||\phi_i - \widehat{\phi}_i| \\ &\le 2M(1 + 3C)\epsilon_n,\end{aligned}$$

for a constant $C' > 0$ depending only on g, because of proposition 2.12. Thus a $2(1 + 3C')M\epsilon_n$-net of $\Theta_{n,j}$ in the distance ρ_n can be constructed with $(\widehat{f}, \widehat{\sigma})$ as above. Recall that $\#\widehat{R}_n \le (4Sn/(M\epsilon_n)^d$, $\#\widehat{\Delta}_n \le (n/(M\epsilon_n))^{H_n}$, $\#\widehat{S}_n = n$, $\#\widehat{U}_n \le \pi n/(2M\epsilon_n)$ and $\#\widehat{V}_k \le (n^{3/2}k/(M\epsilon_n) + 1)^d - (n^{3/2}(k-1)/(M\epsilon_n) - 1)^d \lesssim (n^{3/2}/(M\epsilon_n))^d k^{d-1}$, where we used $u^d - v^d \le d(u - v)u^{-1}$ for $v > u$. Then the end of the proof is identical to proposition 2.4. ■

2.9.2. Approximation of functions

Let $\xi > 0$ and $m, r \ge 1$ be two positive integers. Let define the approximating kernel $L_{m,\xi} : \mathbb{R}^d \to \mathbb{R}$ by the expression $L^{\xi}_{m,r}(x) := \lambda^{\xi}_{m,r} g(x) \prod_{i=1}^d \sin^{2r}(m\xi x_i)/\sin^{2r}(\xi x_i)$, where $\lambda^{\xi}_{m,r} > 0$ is chosen so that $\int_{\mathbb{R}^d} L^{\xi}_{m,r}(x)\,dx = 1$. Also let $\widetilde{f}_0$ denote a suitable Whitney extension of f_0 from $[-S, S]^d$ to $\mathbb{R}^d$ (see the proof of proposition 2.5). We may assume that $\widetilde{f}_0$ and all its derivatives (up to order β) are zero outside $[-2S, 2S]^d$. If it is not the case, it suffices to multiply $\widetilde{f}_0$ by a smooth function that equal 1 on $[-S, S]^d$ and 0 outside $[-2S, 2S]^d$ (for instance, think about the convolution of a bump function with a proper indicator set function).

In order to achieve good order of approximation of f_0 when β is large, we construct a transformation of $\widetilde{f}_0$ as follows. In the sequel we let p be the largest integer strictly smaller than β. For all multi-index $\alpha \in \mathbb{N}^d$, we define $\mathfrak{m}^{m,r,\xi}_{\alpha} := \int_{\mathbb{R}^d} x^{\alpha} L^{\xi}_{m,r}(x)\,dx$. By definition of $L^{\xi}_{m,r}$, the $\mathfrak{m}^{m,r,\xi}_{\alpha}$'s are always finite. Then we define

$$f_{\beta} \equiv f^{m,r,\xi}_{\beta} := \widetilde{f}_0 - \sum_{1 \le |\alpha| \le p} d_{\alpha} \mathfrak{m}^{m,r,\xi}_{\alpha} D^{\alpha} \widetilde{f}_0,$$

where the coefficients (d_{α}) are defined in the same fashion as equation (2.21), with obvious modifications.

Proposition 2.8 – *Let $m, r \ge 1$ be integers. For any $\beta > 0$ and any function $f_0 \in \mathcal{C}^{\beta}[-S, S]^d$ there is a constant $M_{\beta} > 0$ such that $|L^r_{m,\xi} * f_{\beta}(x) - f_0(x)| \le M_{\beta}(\log m/m)^{\beta}$ for all $x \in [-S, S]^d$ if $2r \ge p + 1$ and $\xi = K_0(\log m)^{-1}$ for a constant K_0 depending only on g, β and r.*

Proof. First assume $0 < \beta \le 1$. By assumptions on f_0, there is $M > 0$ such that for all $x, y \in \mathbb{R}^d$ we have $|\widetilde{f}_0(x) - \widetilde{f}_0(y)| \le M|x - y|^{\beta}_d$. Then,

$$\begin{aligned}\left|\widetilde{f}_0(x) - L^{\xi}_{m,r} * \widetilde{f}_0(x)\right| &\le \int_{\mathbb{R}^d} \left|\widetilde{f}_0(x) - \widetilde{f}_0(y)\right| |L^{\xi}_{m,r}(x - y)|\,dy \\ &\le M \int_{\mathbb{R}^d} |x - y|^{\beta}_d |L^{\xi}_{m,r}(x - y)|\,dy.\end{aligned}$$

Remark that for any $\tau > 0$ and all $u \in \mathbb{R}^d$ we have $\sum_{i=1}^d |u_i|^\tau \le d \max_{i=1,\dots,d} |u_i|^\tau \le d(\sum_{i=1}^d |u_i|^2)^{\tau/2}$ and $|x-y|_d^\beta = (\sum_{i=1}^d |x_i - y_i|^2)^{\beta/2} \le d^{\beta/2} \max_{i=1,\dots,d} |x_i - y_i|^\beta \le d^{\beta/2} \sum_{i=1}^d |x_i - y_i|^\beta$. Then, because $|g(x)| \lesssim \exp(-C_0 |x|_d^\tau)$,

$$\begin{aligned}
&\left| \widetilde{f}_0(x) - L^\xi_{m,r} * \widetilde{f}_0(x) \right| \\
&\lesssim \lambda^\xi_{m,r} \sum_{i=1}^d \int_{\mathbb{R}^d} |u_i|^\beta \exp\left(-C_0 d^{-1} \textstyle\sum_{j=1}^d |u_j|^\tau\right) \prod_{j=1}^d \frac{\sin^{2r}(m\xi u_j)}{\sin^{2r}(\xi u_j)}\, du \\
&\lesssim \lambda^\xi_{m,r} \left(\int_{\mathbb{R}} |u|^\beta \, \mathrm{e}^{-C_0 |u|^\tau / d} \frac{\sin^{2r}(m\xi u)}{\sin^{2r}(\xi u)}\, du \right) \left(\int_{\mathbb{R}} \mathrm{e}^{-C_0 |u|^\tau / d} \frac{\sin^{2r}(m\xi u)}{\sin^{2r}(\xi u)}\, du \right)^{d-1}. \qquad (2.26)
\end{aligned}$$

We now bound the first integral of the rhs of equation (2.26). Let split the domain into three parts : $D_1 := (-1/(\xi m), 1/(\xi m))$, $D_2 := [-1/(\xi m), -\pi/\xi] \cup [1/(\xi m), \pi/\xi]$ and $D_3 := \mathbb{R} \backslash (D_1 \cup D_2)$. On D_1 and D_3 we always have $\sin^2(m\xi u)/\sin^2(\xi u) \lesssim m^2$, whereas on D_2 it holds $\sin^2(m\xi u)/\sin^2(\xi u) \lesssim 1/(\xi x)^2$. Therefore,

$$\begin{aligned}
&\int_{\mathbb{R}} |u|^\beta \, \mathrm{e}^{-|u|^\tau/d} \frac{\sin^2(m\xi u)}{\sin^2(\xi u)}\, du \\
&\lesssim m^{2r} \int_{D_1} |u|^\beta \, du + \xi^{-2r} \int_{D_2} |u|^{\beta - 2r} \, du + m^{2r} \int_{D_3} |u|^\beta \, \mathrm{e}^{-C_0 |u|^\tau / d} \, du =: I_1 + I_2 + I_3.
\end{aligned}$$

The bounds $I_1 \lesssim m^{-\beta + (2r-1)} \xi^{-(\beta+1)}$ and $I_2 \lesssim \xi^{-(\beta+1)}(1 + m^{-\beta+(2r-1)})$ are obvious. Now we bound I_3. By Markov's inequality, for any $t < C_0/d$, we have

$$\int_{\pi/\xi}^\infty u^\beta \exp(-C_0 u^\tau / d) \le \mathrm{e}^{-\pi/\xi} \int_0^\infty u^\beta \exp(-C_0 u^\tau / d + ut)\, du.$$

Now it is clear that $I_3 \lesssim m^{2r} \exp(-\pi/\xi)$ since by assumption $\tau > 1$ and we can choose $t < C_0/d$. It follows $I_3 \lesssim \xi^{-(\beta+1)}$ if $\xi = K_0 (\log m)^{-1}$ for a suitable constant $K_0 > 0$ depending only on g, β and r. The same reasoning applies to the second integral of the rhs of equation (2.26), yielding the bound

$$\left| \widetilde{f}_0(x) - L_{m,\xi} * \widetilde{f}_0(x) \right| \lesssim \lambda^\xi_{m,r} m^{-\beta + d(2r-1)} (\log m)^{\beta + d}, \qquad (2.27)$$

whenever $\xi = K_0 (\log m)^{-1}$. Hence, it remains to bound $\lambda^\xi_{m,r}$. By assumption, we have $g(x) \ge 0$ for all $x \in \mathbb{R}^d$ and a constant $C > 0$ such that $g(x) > C$ on a set $E \subseteq [-\pi, \pi]^d$; thus

$$\begin{aligned}
\frac{1}{\lambda^\xi_{m,r}} &\ge \int_E g(x) \prod_{i=1}^d \frac{\sin^{2r}(m\xi x_i)}{\sin^{2r}(\xi x_i)}\, dx \gtrsim C m^{2dr} \int_E \prod_{i=1}^d \frac{\sin^{2r}(m\xi x_i)}{(m\xi x_i)^{2r}}\, dx \\
&\gtrsim \frac{m^{d(2r-1)}}{\xi^d} \int_{E'} \prod_{i=1}^d \frac{\sin^2(u_i)}{u_i^2}\, du \gtrsim \frac{m^{d(2r-1)}}{\xi^d},
\end{aligned}$$

where $E' := \{m\xi x \,:\, x \in E\}$ has non-null Lebesgue measure by assumption. Combining the last result with equation (2.27), we get the estimate $|\widetilde{f}_0(x) - L_{m,\xi} * \widetilde{f}_0(x)| \lesssim m^{-\beta} (\log m)^\beta$ for all $x \in \mathbb{R}^d$ provided $\xi \le K_0 (\log m)^{-1}$.

Now assume that $\beta > 1$. Acting as in the previous paragraph, we can have $\mathfrak{m}_\alpha^{m,r,\xi} \le m^{-|\alpha|} (\log m)^{|\alpha|}$ for all $|\alpha| \le p$, provided $2r > p + 1$ and $\xi = K_0' (\log m)^{-1}$ for a suitable constant $K_0' > 0$. Then the proof is identical to Shen et al. (2013, lemma 2). ■

Proposition 2.9 – *Let $m \geq 1, r \geq (\beta+1)/2$ be integers and $\xi = K_0(\log m)^{-1}$, with K_0 as in proposition 2.8. There exists a discrete mixture $f(x) = \sum_{i=1}^N \alpha_i K_{\xi_i,\phi_i}(x-\mu_i)$ with $N \lesssim (m\log m)^d$ and for all $i = 1,\dots,N$: $\mu_i \in [-2S,2S]^d$, $\xi_i \in [0, 2rK_0 m/\log m]^d$, $\phi_i \in [0,\pi/2]$; such that $|f(x)-f_0(x)| \lesssim (\log m/m)^\beta$ for all $x \in [-S,S]^d$. Moreover $\sum_{i=1}^N |\alpha_i| \lesssim 1$, and for any $i \neq j$ it holds $|\xi_i - \xi_j|_d \geq 2(\log m/m)^\beta$, $|\mu_i - \mu_j|_d \geq 2(\log m/m)^\beta$ and $|\phi_i - \phi_j| \geq 2(\log m/m)^\beta$.*

Proof. We rewrite $L^\xi_{m,r}$ in a more convenient form for the sequel. Let $a_0 := 1$ and $a_k = 2(1-k/m)$ for all $k = 1,\dots,m-1$. Then first step is to notice that

$$L^\xi_{m,r}(x) = m^{dr}\lambda^\xi_{m,r}\, g(x) \prod_{i=1}^d \left[\sum_{k=0}^{m-1} a_k \cos(2\xi k x_i)\right]^r.$$

From here, letting $\mathcal{I}_r := \{0,\dots,m-1\}^r$ and $\mathcal{S} = \{-1,1\}$,

$$L^\xi_{m,r}(x) = m^{dr}\lambda^\xi_{m,r}\, g(x) \prod_{i=1}^d \left[\sum_{k\in\mathcal{I}_r} a'_k 2^{-r} \sum_{e\in\mathcal{S}^r} \cos\left(2\xi x_i \textstyle\sum_{j=1}^r e_j k_j\right)\right],$$

where $a'_k := a_{k_1}\dots a_{k_r}$, and because $\prod_{j=1}^r \cos(2\xi k_j x_i) = 2^{-r}\sum_{e\in\mathcal{S}^r}\cos(2\xi\sum_{j=1}^r e_j k_j x_i)$. Notice that $|a'_k|2^{-r} \leq 1$ for all $k \in \mathcal{I}_r$, and that $2|\sum_{j=1}^r e_j k_j|$ can take at most $1+r(m-1)$ values ; we denote these unique values ω_j with $j \in \mathcal{J} := \{0,\dots,r(m-1)\}$. Then we can rewrite,

$$L^\xi_{m,r}(x) = m^{dr}\lambda^\xi_{m,r}\, g(x) \prod_{i=1}^d \left[\sum_{k\in\mathcal{J}} a''_k \cos(\xi\omega_j x_i)\right],$$

where the coefficients a''_k satisfy $|a''_k| \leq 2\#(\mathcal{I}_r\times\mathcal{S}^r) \leq 2(2m)^r$. Finally, for all $k \in \mathcal{J}^d$ letting $b_k := 2^{-d}a''_{k_1}\dots a''_{k_d}$ and $\omega_{k,i} := \omega_{k_i}$, with the same arguments as previously,

$$L^\xi_{m,r}(x) = m^{dr}\lambda^\xi_{m,r}\, g(x) \sum_{k\in\mathcal{J}^d}\sum_{e\in\mathcal{S}^d} b_k \cos\left(\xi\textstyle\sum_{i=1}^d \omega_{k,i}e_i x_i\right),$$

where $|b_k| \leq (2m)^{dr}$ for all $k \in \mathcal{J}^d$. Therefore,

$$\begin{aligned}
&(m^{dr}\lambda^\xi_{m,r})^{-1} L^\xi_{m,r} * f_\beta(x) \\
&\quad= \sum_{k\in\mathcal{J}^d}\sum_{e\in\mathcal{S}^d} b_k \int_{\mathbb{R}^d} f_\beta(y) g(x-y) \cos\left(\xi\textstyle\sum_{i=1}^d \omega_{k,i}e_i(x_i-y_i)\right) dy \\
&\quad= \sum_{k\in\mathcal{J}^d}\sum_{e\in\mathcal{S}^d} b_k \cos\left(\xi\textstyle\sum_{i=1}^d \omega_{k,i}e_i x_i\right) \int_{\mathbb{R}^d} f_\beta(y) g(x-y) \cos\left(\xi\textstyle\sum_{i=1}^d \omega_{k,i}e_i y_i\right) dy \\
&\qquad+ \sum_{k\in\mathcal{J}^d}\sum_{e\in\mathcal{S}^d} b_k \sin\left(\xi\textstyle\sum_{i=1}^d \omega_{k,i}e_i x_i\right) \int_{\mathbb{R}^d} f_\beta(y) g(x-y) \sin\left(\xi\textstyle\sum_{i=1}^d \omega_{k,i}e_i y_i\right) dy.
\end{aligned}$$

We finish the proof by discretizing the integrals in the last equation. Obviously the proof are identical for both integrals, hence we only consider the first one. To ease notations, we set $h_k(x) := f_\beta(x)\prod_{i=1}^d \cos(\xi\sum_{i=1}^d \omega_{k,i}e_i x_i)$. For any integer $q \geq 1$, proceed as in the proof of proposition 2.6 to find a signed measure $P_{k,q} =: \sum_{l\in\mathcal{L}} p_{k,l}\delta_{x_{k,l}}$ such that

$\int_{[-2S,2S]^d} R(x)\, dP_{k,q}(x) = \int_{[-2S,2S]^d} R(x)\, h_k(x)dx$ for all polynomials $R(x)$ of degree $\leq q$, with $\#\mathcal{L} \leq (q+d)!/(q!d!)$ and $\sum_{l\in\mathcal{L}} |p_{k,l}| = \int_{[-2S,2S]^d} |h_k(x)|\, dx \leq M$ for a positive constant M (recall that by construction of f_β, we have $\|f_\beta\|_\infty < +\infty$, and $\operatorname{supp} f_\beta \subseteq [-2S, 2S]^d$). Then for any $x \in \mathbb{R}^d$,

$$\begin{aligned}\left| \int_{\mathbb{R}^d} h_k(y) g(x-y)\, dy - \int_{[-2S,2S]^d} g(x-y)\, dP_{k,q}(y) \right| \\ \leq \sum_{|\alpha|\leq r} \frac{|D^\alpha g(0)|}{\alpha!} \left| \int_{[-2S,2S]^d} (x-y)^\alpha\, h_k(y)dy - \int_{[-2S,2S]^d} (x-y)^\alpha\, dP_{k,q}(y) \right| \\ + \int_{[-2S,2S]^d} |R_q(y)|\, |h_k(y)|dy + \int_{[-2S,2S]^d} |R_q(y)|\, d|P_{k,q}|(y), \quad (2.28)\end{aligned}$$

where $|R_q(y)| \leq \sup_{|\alpha|=q} |D^\alpha g(0)||y|_d^q/q!$. The first term of the rhs of equation (2.28) is null by construction of $P_{k,q}$. As in the proof of proposition 2.6, the two last terms of equation (2.28) are bounded by a constant multiple of

$$\exp\left\{-(1-\gamma)q\log q + q(1+\log(2\sqrt{d}S))\right\}.$$

Then the error of approximating the integrals is $o(m^{-\beta})$ if $q = K_1 \log m$ for a suitable constant $K_1 > 0$ depending only on β and γ. Since for $\xi = K_0'(\log m)^{-1}$ we have,

$$m^{dr}\lambda_{m,r}^\xi \sum_{k\in\mathcal{J}^d} \sum_{e\in\mathcal{S}^d} |b_k| \lesssim m^{dr} \times m^{-d(2r-1)} \times (\log m)^{-d} \times \#\mathcal{J}^d \times m^{dr} \lesssim (\log m)^{-d},$$

the error of approximating $L_{m,r}^\xi * f_\beta$ by the discretized version does not exceed $o(m^{-\beta})$ when $q = K_1 \log m$. The conclusion of the proposition follows from elementary manipulation of trigonometric functions and because $\#\mathcal{L} \lesssim q^d \lesssim (\log m)^d$.

It remains to prove the separation between the atoms of the mixing measure, but this follows from proposition 2.13 with the same argument as in proposition 2.6. ■

2.9.3. Kullback-Leibler property

Lemma 2.8 – *Let $f_0 \in \mathcal{C}^\beta[-S,S]^d$. Then there exists a constant $C > 0$, not depending on n, such that $\Pi(K_n(\theta_0, \epsilon_n)) \gtrsim \exp(-n\epsilon_n^2)$ for $\epsilon_n^2 = Cn^{-2\beta/(2\beta+d)}(\log n)^{2\beta(2d+1)/(2\beta+d)}$.*

Proof. Let $f_m(x) = \sum_{i=1}^N \alpha_i\, K_{\xi_i,\phi_i}(x-\mu_i)$ be as in proposition 2.9. For any $i = 1,\ldots,N$ define the sets $U_i := \{\xi \in \mathbb{R}^d : |\xi - \xi_i|_d \leq (\log m/m)^\beta\}$, $V_i := \{\mu \in [-2S,2S]^d : |\mu - \mu_i|_d \leq (\log m/m)^\beta\}$ and $W_i := \{\phi \in [0,\pi/2] : |\phi - \phi_i| \leq (\log m/m)^\beta\}$. Notice that these sets are disjoint, and for any $i = 1,\ldots,N$ we have

$$\alpha F(U_i \times V_i \times W_i) \gtrsim |\xi_i|_d^{-a_{12}} (\log m/m)^{\beta(a_1+a_{10}+a_{13})} \gtrsim (\log m/m)^q,$$

where $q := da_{12} + \beta(a_1 + a_{10} + a_{13})$. Then proceed as in lemma 2.5, to find constants $K_1, K_4 > 0$ such that with $\epsilon_n = C_0^{-1} K_1 (\log m/m)^\beta$,

$$\Pi(K_n(\theta_0, \epsilon_n)) \geq \exp\left\{-K_4 m^d (\log m)^{d+1}\right\}. \qquad ■$$

Appendix

2.A. Symmetric Gamma distribution

The symmetric Gamma distribution $\mathrm{SGa}(a,b)$, with $a,b>0$ is the distribution having Fourier transform $t \mapsto (1+t^2/b^2)^{-a}$. It is easily seen that if $X \sim \mathrm{Ga}(a,b)$ and $Y \sim \mathrm{Ga}(a,b)$, with X and Y independent, then $X-Y$ has $\mathrm{SGa}(a,b)$ distribution.

Proposition 2.10 – *Let $Z \sim \mathrm{SGa}(a,b)$. Then for any positive integer n,*

$$\mathbb{E}\, Z^{2n} = \frac{(2n)!}{n!}\frac{(a)^{(n)}}{b^{2n}}, \qquad \mathbb{E}\, Z^{2n+1} = 0.$$

Moreover, the distribution $\mathrm{SGa}(a,b)$ is determined by its moments (in the sense that $\mathrm{SGa}(a,b)$ is the only distribution with this sequence of moments).

Proof. From definition of $\mathrm{SGa}(a,b)$, the random variable Z is distributed as $X-Y$, where $X,Y \sim \mathrm{Ga}(a,b)$ and X,Y are independent. Then it is obvious that all odd moments must vanish. For the even moments, we write,

$$\begin{aligned}\mathbb{E}(X-Y)^{2n} &= \sum_{k=0}^{2n}\binom{2n}{k}(-1)^k\,\mathbb{E}\,X^{2n-k}\,\mathbb{E}\,Y^k \\ &= \frac{1}{b^{2n}}\sum_{k=0}^{2n}\binom{2n}{k}(-1)^k(a)^{(2n-k)}(a)^{(k)} = \frac{(2n)!}{n!}\frac{(a)^{(n)}}{b^{2n}},\end{aligned}$$

where the last equality can be obtained after some algebra. To see that $\mathrm{SGa}(a,b)$ is determined by its moments, we check that Carleman's criteria applies (Gut, 2006), which is straightforward. ■

Proposition 2.11 – *Let $X \sim \mathrm{SGa}(\alpha,\eta)$, with $0<\alpha\leq 1$ and $\eta>0$. Then there is a constant $C>0$ such that for any $x\in\mathbb{R}$ and any $0<\delta\leq(3+\eta)^{-1}$ we have $\mathbb{P}(|X-x|\leq\delta) \geq C\delta\,\mathrm{e}^{-(3+\eta)|x|}\,\Gamma(\alpha)^{-1}$.*

Proof. Assume for instance that $x\geq 0$. Recalling that X is distributed as the difference of two independent $\mathrm{Ga}(\alpha,\eta)$ distributed random variables, it follows

$$\mathbb{P}(|X-x|\leq\delta) \geq \frac{1}{\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}\,\mathrm{e}^{-\eta y}\,\frac{1}{\Gamma(\alpha)}\int_{x+y}^{x+y+\delta} z^{\alpha-1}\,\mathrm{e}^{-\eta z}\,dz\,dy.$$

Because $\alpha\leq 1$, the mapping $z\mapsto z^{\alpha-1}e^{-\eta z}$ is monotonically decreasing on $\mathbb{R}^+$, then the last integral in the rhs of the previous equation is lower bounded by $\delta(x+y+$

$\delta)^{\alpha-1}\,\mathrm{e}^{-\eta(x+y+\delta)} \geq \delta\,\mathrm{e}^{-(3+\eta)(x+y+\delta)}$. Then

$$\begin{aligned}\mathbb{P}(|X-x|\leq\delta) &\geq \frac{\delta\,\mathrm{e}^{-(3+\eta)(x+\delta)}}{\Gamma(\alpha)^2}\int_0^\infty y^{\alpha-1}\,\mathrm{e}^{-(3+2\eta)y}\,dy\\ &= \frac{\delta\,\mathrm{e}^{-(3+\eta)(x+\delta)}}{(3+2\eta)^\alpha\Gamma(\alpha)} \geq \frac{\delta\,\mathrm{e}^{-(3+\eta)|x|}}{\mathrm{e}(3+2\eta)^\alpha\Gamma(\alpha)}.\end{aligned}$$

The proof when $x<0$ is obvious. ■

2.B. Auxiliary results

Proposition 2.12 – *Let $K_A(x)=g(A^{-1}x)$, and assume that for all multi-index $k\in\mathbb{N}^d$ with $|k|=0,1,2$ the mapping $x\mapsto x^k g(x)$ belongs to $L^1(\mathbb{R}^d)$. Let $\|\cdot\|$ be the spectral norm on $\mathcal{E}$. Then there is a constant $C>0$ such that for all $x,\mu_1,\mu_2\in\mathbb{R}^d$ and all $A_1,A_2\in\mathcal{E}$ arbitrary with $\|I-A_1^{-1}A_2\|\wedge\|I-A_2^{-1}A_1\|$ small enough,*

$$\begin{aligned}|K_{A_1}(x-\mu_1)-K_{A_2}(x-\mu_2)| \leq C\|I-A_1^{-1}A_2\|\wedge C\|I-A_2^{-1}A_1\|\\ + C\left(\|A_1^{-1}\|\wedge\|A_2^{-1}\|\right)|\mu_1-\mu_2|_d.\end{aligned}$$

Proof. Starting from the triangle inequality, we have

$$\begin{aligned}|K_{A_1}(x-\mu_1)-K_{A_2}(x-\mu_2)| \leq |K_{A_1}(x-\mu_2)-K_{A_2}(x-\mu_2)|\\ + |K_{A_1}(x-\mu_1)-K_{A_1}(x-\mu_2)| \quad (2.29)\end{aligned}$$

We recall that $K_A(x):=g(A^{-1}x)$. To bound the first term, it is enough to bound $g(x)-g(A_1^{-1}A_2x)$ for all $x\in\mathbb{R}^d$. Let $(B_n)_{n\in\mathbb{N}}$ and $(C_n)_{n\in\mathbb{N}}$ be two arbitrary sequences in $\mathcal{E}$ such that $\|I-B_n^{-1}C_n\|\leq 1/n$, and let $\widehat{g}$ denote the Fourier transform of g. Then,

$$\sup_{x\in\mathbb{R}^d}\left|g(x)-g(B_n^{-1}C_nx)\right| \leq \int_{\mathbb{R}^d}\left|\widehat{g}(\xi)-|\det(B_n^{-1}C_n)|\,\widehat{g}(B_n^{-1}C_n\xi)\right|\,d\xi.$$

Remark that $|\det B_n^{-1}C_n|\leq 1+|\det(I-B_n^{-1}C_n)|$, and $\|I-B_n^{-1}C_n\|\leq 1/n$ implies that $|\det(I-B_n^{-1}C_n)|\leq\sqrt{d}/n^d$. Also, $|B_n^{-1}C_n\xi|_d\leq\|I-B_n^{-1}C_n\||\xi|_d+|\xi|_d\leq(1+1/n)|\xi|_d$. It turns out that,

$$\lim_{n\to\infty}|\det B_n^{-1}C_n|\,\widehat{g}(B_n^{-1}C_n\xi)=\widehat{g}(\xi).$$

We now prove that $\{|\det B_n^{-1}C_n|\,\widehat{g}(B_n^{-1}C_n\xi) : n\geq 2\}$ is dominated. By assumption, $g\in L^1(\mathbb{R}^d)$, as well as $x\mapsto x^k g(x)$ with $|k|=1,2$. This implies that $|\widehat{g}(\xi)|\leq C(1+|\xi|_d)^{-2}$ for some $C>0$. We already saw that $|\det B_n^{-1}C_n|\leq 1+1/n^d$, and $|\xi|_d\leq|B_n^{-1}C_n\xi|_d+|(I-B_n^{-1}C_n)\xi|_d$ implies $|B_n^{-1}C_n\xi|_d\geq(1-1/n)|\xi|_d$. Therefore, for any $n\geq 2$,

$$|\det B_n^{-1}C_n|\,\widehat{g}(B_n^{-1}C_n\xi)\leq\frac{C|\det B_n^{-1}C_n|}{(1+|B_n^{-1}C_n\xi|_d)^2}\leq\frac{C(1+2^{-d})}{(1+|\xi|_d/2)^2}.$$

Then the dominated convergence applies, and

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}^d}|g(x)-g(B_n^{-1}C_nx)|=0.$$

The second term of the rhs of equation (2.29) is bounded above by $|A_1^{-1}(\mu_1 - \mu_2)|_d \leq \|A_1^{-1}\|\,|\mu_1 - \mu_2|_d$, using Lipshitz continuity of g. Using a symmetry argument, the conclusion of the proposition follows. ■

Proposition 2.13 – *Let $K_{\xi,\phi}(x) = g(x)\cos(\sum_{i=1}^d \xi_i x_i + \phi)$, and assume that for all multi-index $k \in \mathbb{N}^d$ with $|k| \leq 1$ we have $\sup_{x\in\mathbb{R}^d} |x^k g(x)| < +\infty$ and $\sup_{x\in\mathbb{R}^d} |D^k g(x)|$. Then there is a constant $C > 0$ such that for all $x, \mu_1, \mu_2, \xi_1, \xi_2 \in \mathbb{R}^d$ and all $\phi_1, \phi_2 \in [0, \pi/2]$*

$$|K_{\xi_1}(x - \mu_1) - K_{\xi_2}(x - \mu_2)| \leq C|\xi_1 - \xi_2|_d + C|\mu_1 - \mu_2|_d + C|\phi_1 - \phi_2|.$$

Proof. We write,

$$\begin{aligned}|K_{\xi_1,\phi_1}(x - \mu_1) - K_{\xi_2,\phi_2}(x - \mu_2)| &\leq |K_{\xi_1,\phi_1}(x - \mu_1) - K_{\xi_1,\phi_1}(x - \mu_2)| \\ &\quad + |K_{\xi_1,\phi_1}(x - \mu_2) - K_{\xi_1,\phi_2}(x - \mu_2)| + |K_{\xi_1,\phi_2}(x - \mu_2) - K_{\xi_2,\phi_2}(x - \mu_2)|.\end{aligned}$$

Because g has bounded first derivatives, it is Lipschitz continuous for some Lipschitz contant $K > 0$, then the first term of the rhs is bounded above by $K|\mu_1 - \mu_2|_d$. With the same argument, the second term is bounded by a constant multiple of $\|g\|_\infty|\phi_1 - \phi_2|$. The last term of the rhs is easily bounded, because for all $x \in \mathbb{R}^d$:

$$\begin{aligned}|K_{\xi_1,\phi_2}(x) - K_{\xi_2,\phi_2}(x)| &\leq |\cos(\textstyle\sum_{i=1}^d \xi_{1,i}x_i + \phi_2) - \cos(\sum_{i=1}^d \xi_{2,i}x_i + \phi_2)||g(x)| \\ &\leq \sum_{i=1}^d |\xi_{1,i}x_i - \xi_{2,i}x_i||g(x)| \\ &\leq \left(\sum_{i=1}^d |\xi_{1i} - \xi_{2i}|^2\right)^{1/2} \left(\sum_{i=1}^d |x_i g(x)|^2\right)^{1/2},\end{aligned}$$

where the last line holds by Hölder's inequality. Then the conclusion follows $x \mapsto x^k g(x)$ is bounded for all $|k| = 1$. ■

Proposition 2.14 – *Let $g(x) = \exp(-|x|_d^2/2)$. Then $\sup_{x\in\mathbb{R}^d} |D^\alpha g(x)| \lesssim \exp(\frac{1}{2}|\alpha|\log|\alpha|)$ for all $\alpha \in \mathbb{N}^d$.*

Proof. For any $\alpha \in \mathbb{N}^d$, let $k = |\alpha| = \sum_{i=1}^d \alpha_i$. When $k < 2$, the result is obvious. Now assume that $k \geq 2$. By Fourier duality, we have for all $x \in \mathbb{R}^d$

$$|D^\alpha g(x)| \leq \int |u^\alpha g(u)|\,du \lesssim 2^{k/2} \prod_{i=1}^d \Gamma\left(\frac{\alpha_i + 1}{2}\right) \lesssim 2^{k/2} \prod_{i=1}^d \left(\frac{2}{\alpha_i + 1}\right)\left(\frac{\alpha_i + 1}{2e}\right)^{\frac{\alpha_i+1}{2}},$$

where the last inequality follows from Stirling formula. Then it is clear that,

$$|D^\alpha g(0)| \lesssim \exp\left\{-\frac{k}{2} - \frac{1}{2}\sum_{i=1}^d \log(1 + \alpha_i) + \frac{1}{2}\sum_{i=1}^d \alpha_i \log(1 + \alpha_i)\right\}.$$

The result follows because for all $k \geq 2$ we have $\sum_{i=1}^d \alpha_i \log(1+\alpha_i) \leq \sum_{i=1}^d \alpha_i \log(1+k) \leq (1/2 + \log k)\sum_{i=1}^d \alpha_i \leq k/2 + k\log k$. ■

Acknowledgements

The authors are grateful to Judith Rousseau and Trong Tuong Truong for their helpful support and valuable advice throughout the writing of this article, and also to Pr. Robert L. Wolpert for discussions and RJMCMC source code.

Bibliography

Improved radon based imaging using the shearlet transform, volume 7343, 2009. doi: 10.1117/12.820066. URL http://dx.doi.org/10.1117/12.820066.

F. Abramovich, T. Sapatinas, and B. Silverman. Stochastic expansions in an overcomplete wavelet dictionary. *Probability Theory and Related Fields*, 117(1):133–144, 2000. ISSN 1432-2064. doi: 10.1007/s004400050268. URL http://dx.doi.org/10.1007/s004400050268.

A. Antoniadis, J. Bigot, and T. Sapatinas. Wavelet estimators in nonparametric regression: a comparative simulation study. *Journal of Statistical Software*, 6:1–83, 2001. URL http://hal.archives-ouvertes.fr/hal-00823485/.

O. E. Barndorff-Nielsen and J. Schmiegel. Lévy-based spatial-temporal modelling, with applications to turbulence. *Russian Mathematical Surveys*, 59(1):65, 2004. URL http://stacks.iop.org/0036-0279/59/i=1/a=R06.

H. Bauer. *Measure and integration theory*, volume 26. Walter de Gruyter, 2001.

P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 42(3):273–325, 2006. ISSN 0246-0203. doi: http://dx.doi.org/10.1016/j.anihpb.2005.04.004. URL http://www.sciencedirect.com/science/article/pii/S0246020305000841.

A. Canale and P. De Blasi. Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404, 02 2017. doi: 10.3150/15-BEJ746. URL http://dx.doi.org/10.3150/15-BEJ746.

T. Choi and M. J. Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969 – 1987, 2007. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2007.01.004. URL http://www.sciencedirect.com/science/article/pii/S0047259X07000048.

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*, volume 2. Springer Science & Business Media, 2007.

P. De Blasi, G. Peccati, and I. Prünster. Asymptotics for posterior hazards. *Ann. Statist.*, 37(4):1906–1945, 08 2009. doi: 10.1214/08-AOS631. URL http://dx.doi.org/10.1214/08-AOS631.

R. de Jonge and J. H. van Zanten. Adaptive nonparametric bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320, 12 2010. doi: 10.1214/10-AOS811. URL http://dx.doi.org/10.1214/10-AOS811.

E. Del Barrio, P. Deheuvels, and S. Van De Geer. *Lectures on empirical processes*. European Mathematical Society, 2007.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1994.

S. Favaro, A. Guglielmi, and S. G. Walker. A class of measure-valued markov chains and bayesian nonparametrics. *Bernoulli*, 18(3):1002–1030, 08 2012. doi: 10.3150/11-BEJ356. URL http://dx.doi.org/10.3150/11-BEJ356.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *IN BAYESIAN STATISTICS*, pages 169–193. University Press, 1992.

S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 02 2007a. doi: 10.1214/009053606000001172. URL http://dx.doi.org/10.1214/009053606000001172.

S. Ghosal and A. van der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 04 2007b. doi: 10.1214/009053606000001271. URL http://dx.doi.org/10.1214/009053606000001271.

S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29 (5):1233–1263, 10 2001. doi: 10.1214/aos/1013203452. URL http://dx.doi.org/10.1214/aos/1013203452.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 04 2000. doi: 10.1214/aos/1016218228. URL http://dx.doi.org/10.1214/aos/1016218228.

A. Gut. *Probability: A Graduate Course*. Springer Science & Business Media, 2006.

H. Ishwaran and L. F. James. Computational methods for multiplicative intensity models using weighted gamma processes. *Journal of the American Statistical Association*, 2012.

P. Jacob and P. E. Oliveira. A representation of infinitely divisible signed random measures. *Portugaliae Mathematica*, 52(2):211–220, 1995.

O. Kallenberg. *Random measures*. Academic Pr, 1983.

J. F. C. Kingman. *Poisson processes*, volume 3. Oxford university press, 1992.

A. Lijoi and B. Nipoti. A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association*, 109(506): 802–814, 2014. doi: 10.1080/01621459.2013.869499. URL http://dx.doi.org/10.1080/01621459.2013.869499.

A. Y. Lo and C.-S. Weng. On a class of bayesian nonparametric estimates: Ii. hazard rate estimates. *Annals of the Institute of Statistical Mathematics*, 41(2):227–245, 1989. ISSN 1572-9052. doi: 10.1007/BF00049393. URL http://dx.doi.org/10.1007/BF00049393.

J. S. Marron, S. Adak, I. M. Johnstone, M. H. Neumann, and P. Patil. Exact risk analysis of wavelet regression. *Journal of Computational and Graphical Statistics*, 7(3):278–309, 1998. doi: 10.1080/10618600.1998.10474777. URL http://amstat.tandfonline.com/doi/abs/10.1080/10618600.1998.10474777.

P. Müller, A. Erkanli, and M. West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79, 1996. doi: 10.1093/biomet/83.1.67. URL http://biomet.oxfordjournals.org/content/83/1/67.abstract.

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. doi: 10.1080/10618600.2000.10474879. URL http://amstat.tandfonline.com/doi/abs/10.1080/10618600.2000.10474879.

G. Peccati and I. Prünster. Linear and quadratic functionals of random hazard rates: An asymptotic analysis. *Ann. Appl. Probab.*, 18(5):1910–1943, 10 2008. doi: 10.1214/07-AAP509. URL http://dx.doi.org/10.1214/07-AAP509.

B. S. Rajput and J. Rosinski. Spectral representations of infinitely divisible processes. *Probability Theory and Related Fields*, 82(3):451–487, 1989. ISSN 1432-2064. doi: 10.1007/BF00339998. URL http://dx.doi.org/10.1007/BF00339998.

W. Rudin. *Real and Complex Analysis, 1966*. McGraw-Hill, New York, 1974.

W. Shen, S. T. Tokdar, and S. Ghosal. Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013. doi: 10.1093/biomet/ast015. URL http://biomet.oxfordjournals.org/content/100/3/623.abstract.

L. A. Shepp and B. F. Logan. The fourier reconstruction of a head section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, June 1974. ISSN 0018-9499. doi: 10.1109/TNS.1974.6499235.

V. Tchakaloff. Formules de cubatures mécaniques à coefficients non négatifs. *Bull. Sci. Math*, 81(2):123–134, 1957.

M. West. *Hyperparameter estimation in Dirichlet process mixture models*. Citeseer, 1992.

H. Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934.

R. L. Wolpert, M. A. Clyde, and C. Tu. Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *Ann. Statist.*, 39(4):1916–1962, 08 2011. doi: 10.1214/11-AOS889. URL http://dx.doi.org/10.1214/11-AOS889.

CHAPTER 3

Tails assumptions and posterior concentration rates for mixtures of Gaussians

Résumé

De nos jours en estimation de densité, les vitesses de contraction de la distribution a posteriori pour les mélanges en translation et en moyenne et variance de Gaussiennes ne sont connus que sous des hypothèses de queues légères; avec de meilleures vitesses obtenues pour les mélanges en translation. Il a été conjecturé par Canale et DeBlasi, mais non prouvé, que la situation devrait être inversée sous des hypothèses de queues lourdes. Cette conjecture est basée sur l'intuition qu'il n'est pas nécessaire d'obtenir un bon ordre d'approximation dans les régions avec peu d'observations (disons les queues), favorisant l'usage des mélanges en moyenne et variance qui peuvent adapter l'ordre d'approximation spatialement. Ici, nous testons l'argument précédent sur le problème de régression gaussienne avec covariables aléatoires. Bien que nous ne puissions pas encore invalider la conjecture, nous trouvons que même avec une hypothèse de queues lourdes, les mélanges en moyenne et variance semblent être moins performants que les mélanges en translation. Cependant, les preuves suggèrent l'introduction d'un mélange *hybride*, qui semble être toujours plus performant que les deux autres, qu'importe la nature des queues. Finalement, nous montrons que toutes les hypothèses de queues peuvent être supprimées, au prix de rendre la distribution a priori dépendente des covariables.

Abstract

Nowadays in density estimation, posterior rates of convergence for location and location-scale mixtures of Gaussians are only known under light-tail assumptions; with better rates achieved by location mixtures. It has been conjectured by Canale and DeBlasi, but not proved, that the situation should be reversed under heavy tails assumptions. The conjecture is based on the feeling that there is no need to achieve a good order of approximation in regions with few data (say, in the tails), favoring location-scale mixtures which allow for spatially varying order of approximation. Here we test the previous argument on the Gaussian errors mean regression model with random design, for which the light tail assumption is not required for proofs. Although we cannot invalidate the conjecture due to the lack of lower bound, we find that even with heavy tails assumptions, location-scale mixtures apparently perform in general worst than location mixtures. However, the proofs suggest to introduce *hybrid* location-scale mixtures that are find to outperform both location and location-scale mixtures, whatever the nature of the tails. Finally, we show that all tails assumptions can be released at the price of making the prior distribution covariate dependent.

3.1. Introduction

Nonparametric mixture models are highly popular in the Bayesian nonparametric literature, due to both their reknown flexibility and relative easiness of implementation, see Hjort et al. (2010) for a review. They have been used in particular for density estimation, clustering and classification and recently nonparametric mixture models have also been proposed in nonlinear regression models, see for instance de Jonge and van Zanten (2010); Wolpert et al. (2011); Naulet and Barat (2015).

There is now a large literature on posterior concentration rates for nonparametric mixture models, initiated by Ghosal and Van Der Vaart (2001); Ghosal et al. (2007a) and improved by Kruijer et al. (2010); Shen et al. (2013); Scricciolo (2014) in the context of location mixtures of Gaussian distributions and studied by Canale and De Blasi (2017) in the context of location-scale Gaussian distributions and de Jonge and van Zanten (2010) in the case of location mixture models for nonlinear regression.

Location mixture of Gaussian densities can be writen as

$$f_{\sigma,G}(x) = \int_{\mathbb{R}} \varphi_\sigma(x-\mu) dG(\mu), \tag{3.1}$$

while location-scale mixtures have the form

$$f_G(x) = \int_{\mathbb{R}\times\mathbb{R}^+} \varphi_\sigma(x-\mu) dG(\mu,\sigma). \tag{3.2}$$

These models are used in the Bayesian nonparametric literature to model smooth curves, typically probability densities, by putting a prior on the mixing distribution G (and on σ for location mixtures (3.1)). The most popular prior distributions on G are either finite with unknown number of components, as in Kruijer et al. (2010) and the reknown Dirichlet Process (Ferguson (1973)) or some of its extensions. In both cases G is discrete almost surely.

In Kruijer et al. (2010) and later on in Shen et al. (2013); Scricciolo (2014) it was proved that location mixture of Gaussian distributions lead to adaptive (nearly) optimal posterior concentration rates (for L_1 metrics) over collections of Hölder types functional classes, in the context of density estimation for independently and identically distributed random variables. Contrarywise, in Canale and De Blasi (2017), suboptimal posterior concentration rates are derived and the authors obtain rates that are at best $n^{-\beta/(2\beta+2)}$ up to a $\log n$ term in place of $n^{-\beta/(2\beta+1)}$. These results are obtained under strong assumptions on the tail of the true density f_0, since it is assumed that $f_0(x) \lesssim e^{-c|x|^\tau}$ when x goes to infinity, for some positive c, τ.

In Canale and De Blasi (2017), the authors suggest that location-scale mixtures might lead to suboptimal posterior concentration rates, for light tail distributions but might be more robust to tails, since the rate $n^{-\beta/(2\beta+2)}$ is the minimax estimation rate for density estimation with regularity β, under the L_2 loss, see Reynaud-Bouret et al. (2011); Goldenshluger and Lepski (2014).

The question thus remains open as to how robust to tails mixtures of Gaussian distributions (either location or location-scale) are.

Interestingly in Bochkina and Rousseau (2016), much weaker tail constraints are necessary to achieve the minimax rate $n^{-\beta/(2\beta+1)}$, for estimating densities on $\mathbb{R}^+$ using

mixtures of Gamma distributions. The authors merely require that F_0 allows for a moment of order striclty greater than 2. However in Bochkina and Rousseau (2016) as well as in Kruijer et al. (2010); Shen et al. (2013); Scricciolo (2014), the smoothness functional classes are non standard and roughly correspond to requiring that the log-density is locally Hölder, which blurs the understanding of the robustness of Gaussian mixtures to tails. These smoothness conditions are required to ensure that the density f_0 can be approximated by a mixture $f_{\sigma,G}$ where G is a probability measure in terms of Kullback-divergence. Hence to better understand the ability of mixture models to capture heavy tails we study their use in nonparametric regression models:

$$\begin{aligned} Y_i &= f(X_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} N(0, s^2), \quad i = 1, \dots, n, \\ X_1, \dots, X_n &\overset{\text{i.i.d}}{\sim} Q_0, \quad f \in L^2(Q_0). \end{aligned} \tag{3.3}$$

The parameter is f with prior distribution denoted by Π. We assume that s is known, which is just a matter of convenience for proofs. All the results of the paper can be translated to the case s unknown using the same methodology as Salomond (2013) or Naulet and Barat (2015). Our aim is to study posterior concentration rates around the true regression function f_0 defined by sequences ϵ_n converging to zero with n and such that

$$\Pi\left(n^{-1}\sum_{i=1}^n |f(x_i) - f_0(x_i)|^2 \leq \epsilon_n^2 \mid \mathbf{y}^n, \mathbf{x}^n\right) = 1 + o_p(1), \tag{3.4}$$

under the model f_0. By analogy to the case of density estimation of Reynaud-Bouret et al. (2011) and Goldenshluger and Lepski (2014) we assume that $f_0 \in L^1$ and belongs to a Hölder ball with smoothness β. The tail condition are then on the design distribution and written as $\int_{\mathbb{R}} |x|^p dQ_0(x) < +\infty$, $p \geq 0$, and our aim is to study the posterior concentration rate (3.4) for both location and location-scale mixtures.

We show in section 3.2, that in most cases location mixtures have a better posterior concentration rate than location-scale mixtures and unless p goes to infinity the posterior concentration rates is not as good as the usual $n^{-\beta/(2\beta+1)}$. This rate is suboptimal for light tail design points, since in this case the minimax posterior concentration rate is given by $n^{-\beta/(2\beta+1)}$. To improve on this rate we propose a new version of location-scale mixture models, which we call the hybrid location-scale mixture and we show that this nonparametric mixture model leads to better posterior concentration rates than the location mixture (and thus than the location-scale mixture). All these results are up to $\log n$ terms. The results are summarized in table 3.1 which displays the value q defined by $\epsilon_n^2 = n^{-q}$.

Although the results are presented in the regression model, we believe that similar phenomena should take place in the density estimation problem.

The main results with the description of the three types of prior models and the associated posterior concentration rates are presented in section 3.2. Proofs are presented in section 3.3 and some technical lemmas are proved in the appendix.

3.1.1. Notations

We call $P_f(\cdot \mid X)$ the distribution of the random variable $Y \mid X$ under the model (3.3), associated with the regression function f. Given $(X_1, \dots, X_n)$, $P_f^n(\cdot \mid X_1, \dots, X_n)$ stands for the distribution of the random vector $(Y_1, \dots, Y_n)$ of independent random variables

Table 3.1: Summary of posterior rates of convergence for different types of mixtures. The rates are understood to be in the form $\epsilon_n^2 = n^{-q}$, up to powers of $\log n$ factors, where q is given below.

	$0 < p < 2$			$p \geq 2$		
	$0 < p \leq \frac{2\beta}{\beta+1}$	$\frac{2\beta}{\beta+1} < p \leq 2\beta$	$p > 2\beta$	$0 < p \leq \frac{2\beta}{\beta+1}$	$\frac{2\beta}{\beta+1} < p \leq 2\beta$	$p > 2\beta$
Location	$\frac{2\beta}{3\beta+1}$	$\frac{2\beta}{3\beta+1}$	$\frac{2\beta}{3\beta+1}$	$\frac{2\beta}{2\beta+1+2\beta/p}$	$\frac{2\beta}{2\beta+1+2\beta/p}$	$\frac{2\beta}{2\beta+1+2\beta/p}$
Location-scale	$\frac{2\beta}{3\beta+2}$	$\frac{2\beta}{2\beta+1+2\beta/p}$	$\frac{2\beta}{2\beta+2}$	$\frac{2\beta}{3\beta+2}$	$\frac{2\beta}{2\beta+1+2\beta/p}$	$\frac{2\beta}{2\beta+2}$
Hybrid	$\frac{2\beta}{3\beta+1}$	$\frac{p}{p+1}$	$\frac{2\beta}{2\beta+1}$	$\frac{2\beta}{3\beta+1}$	$\frac{p}{p+1}$	$\frac{2\beta}{2\beta+1}$

$Y_j \sim P_f(\cdot \mid X_j)$. Also, for any random variable X with distribution P, and any function g, $Pg(X)$ denote the expectation of $g(X)$.

For any $\alpha > 0$, we let $\mathrm{SGa}(\alpha)$ denote the symmetric Gamma distribution with parameter α; that is $X \sim \mathrm{SGa}(\alpha)$ has the distribution of the difference of two independent Gamma random variables with parameters $(\alpha, 1)$.

For any finite positive measure α on the measurable space $(X, \mathcal{X})$, let Π_α denote the symmetric Gamma process distribution with parameter α (Wolpert et al., 2011; Naulet and Barat, 2015); that is, an $M \sim \Pi_\alpha$ is a random signed measure on $(X, \mathcal{X})$ such that far any disjoints $B_1, \ldots, B_k \in \mathcal{X}$ the random variables $M(B_1), \ldots, M(B_k)$ are independent with distributions $\mathrm{SGa}(\alpha(B_i))$, $i = 1, \ldots, k$.

For any $\beta > 0$, we let $\mathtt{C}^\beta$ denote the Hölder space of order β; that is the set of all functions $f : \mathbb{R} \to \mathbb{R}$ that have bounded derivatives up to order m, the largest integer smaller than β, and such that the norm $\|f\|_{\mathtt{C}^\beta} := \sup_{k \leq m} \sup_{x \in \mathbb{R}} |f^{(k)}(x)| + \sup_{x \neq y} |f^{(m)}(x) - f^{(m)}(y)|/|x - y|^{\beta - m}$ is finite.

For $1 \leq p < \infty$ we let L^p be the space of function for which the norm $\|f\|_p^p := \int |f(x)|^p \, dx$ is finite; and by L^∞ we mean the space of functions for which $\|f\|_\infty := \sup_{x \in \mathbb{R}} |f(x)|$ is finite. For $0 \leq p, q \leq \infty$ and functions $f \in L^p$, $g \in L^q$, we write $f * g$ the convolution of f and g, that is $f * g(x) := \int f(x - y) g(y) \, dy$ for all $x \in \mathbb{R}$. Moreover, we'll use repeatedly Young's inequality which state that $\|f * g\|_r \leq \|f\|_p \|g\|_q$, with $1/p + 1/q = 1/r + 1$.

If $f \in L^1$, then we define $\widehat{f}$ as the (L^1) Fourier transform of f; that is $\widehat{f}(\xi) := \int f(x) e^{-i\xi x} \, dx$ for all $\xi \in \mathbb{R}$. Moreover, if $\widehat{f} \in L^1$, then the inverse Fourier transform is well-defined and $f(x) = (2\pi)^{-1} \int \widehat{f}(\xi) e^{ix\xi} \, d\xi$. Also, we denote by $\mathcal{S}$ the Schwartz space; that is the space of infinitely differentiable functions $f : \mathbb{R} \to \mathbb{R}$ for which $|x^r f^{(k)}(x)| < +\infty$ for all $r > 0$ and all $k \in \mathbb{N}$. Then $\mathcal{S} \subset L^1$, and it is well known that the Fourier transform maps $\mathcal{S}$ onto itself, thus the Fourier transform is always invertible on $\mathcal{S}$. We note $\|f\|_{r,k} = \sup\{|x|^r |f^{(k)}(x)|, x \in \mathbb{R}\}$ for any $f \in \mathcal{S}$.

For two real numbers a, b, the notation $a \wedge b$ stand for the minimum of a and b whereas $a \vee b$ stand for the maximum. Similarly, given two real valued functions f, g the function $f \wedge g$ is the function which at x assigns the minimum of $f(x)$ and $g(x)$ and $f \vee g$ has obvious definition. Throughout the paper C denotes a generic constant.

Inequalities up to a generic constant are denoted by $\lesssim$ and $\gtrsim$.

3.2. Posterior convergence rates for Symmetric Gamma mixtures

In this section we present the main results of the paper. We first present the three types of priors that are studied; i.e. location mixtures, location-scale mixtures and hybrid location-scale mixtures and for each of these families of priors we provide the associated posterior concentration rates.

Recall that we consider observations $(Y_i, X_i)_{i=1}^n$ independent and identically distributed according to model (3.3) and we note $\mathbf{y}^n = (Y_1, \cdots, Y_n)$ and $\mathbf{x}^n = (X_1, \cdots, X_n)$. We denote the prior and the posterior distribution on f by $\Pi(\cdot)$ and $\Pi(\cdot \mid \mathbf{y}^n, \mathbf{x}^n)$ respectively.

3.2.1. Family of priors

Location mixtures of Gaussians

A symmetric Gamma process location mixture of Gaussians prior Π is the distribution of the random function $f(x) := \int \varphi((x-\mu)/\sigma)\, dM(\mu)$ where $\sigma \sim G_\sigma$ and $M \sim \Pi_\alpha$, with α a finite positive measure on $\mathbb{R}$, G_σ a probability measure on $(0, \infty)$ and $\varphi(x) := e^{-x^2/2}$ for all $x \in \mathbb{R}$.

We restrict our discussion to priors for which the following conditions are verified. We assume that there are positive constants a_1, a_2, a_3 and b_1, b_2, b_3, b_4 such that G_σ satisfies for $x \geq 1$

$$G_\sigma(\sigma > x) \lesssim \exp(-a_1 x^{b_1}) \tag{3.5}$$

$$G_\sigma(\sigma \leq 1/x) \lesssim \exp(-a_2 x^{b_2}) \tag{3.6}$$

$$G_\sigma\left(x^{-1} \leq \sigma \leq x^{-1}(1+t)\right) \gtrsim x^{b_3} t^{b_4} \exp(-a_3 x), \quad \forall t \in (0,1). \tag{3.7}$$

We let $\alpha := \overline{\alpha} G_\mu$ for a positive constant $\overline{\alpha} > 0$ and G_μ a probability distribution on $\mathbb{R}$. We assume that there are positive constants b_5, b_6 such that G_μ satisfies for all $x \in \mathbb{R}$

$$G_\mu(|\mu - x| \leq t) \gtrsim t^{b_5}(1 + |x|)^{-b_6}, \quad \forall t \in (0,1). \tag{3.8}$$

The heavy tail condition on G_μ is required to not deteriorate the rate of convergence when Q_0 is heavy tailed.

Notice that equation (3.5) forbids the use of the classical inverse-Gamma distribution as prior distribution on σ because of its heavy tail. In fact, it is always possible to weaken equation (3.5) to allow for Inverse-Gamma distribution (see Canale and De Blasi (2017); Naulet and Barat (2015)) but it complicates the proofs with no contribution to the subject of the paper. We found that among the usual distributions the inverse-Gaussian is more suitable for our purpose since it fulfills all the equations (3.5) to (3.7), as shown in proposition 3.1. We recall that the inverse-Gaussian distribution on $(0, \infty)$ with parameters $a > 0$, $b > 0$ has density with respect to Lebesgue measure

$$f(x; a, b) := \left(\frac{b}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{b(x-a)^2}{2a^2 x}\right), \quad \forall x > 0,$$

and $f(x; a, b) = 0$ elsewhere.

Proposition 3.1 – *The inverse-Gaussian distribution with parameters $b, a > 0$ satisfies equations (3.5) to (3.7) with $a_1 = b/(2a^2)$, $b_1 = 1$, $a_2 = b/4$, $b_2 = 1$, $b_3 = 1$, $b_4 = 1$ and $a_3 = b/2$.*

Proof. It suffices to write, for any $x \geq 1$

$$G_\sigma(\sigma > x) \leq \left(\frac{b}{2\pi x^3}\right)^{1/2} \int_x^\infty \exp\left(-\frac{b(t-a)^2}{2a^2 t}\right) dt$$
$$\leq \left(\frac{b}{2\pi}\right)^{1/2} \exp\left(\frac{b}{a} - \frac{b}{2}\right) \int_x^\infty \exp\left(-\frac{bt}{2a^2}\right) dt.$$

Also, for any $x \geq 1$

$$G_\sigma(\sigma \leq 1/x) \leq \left(\frac{b}{2\pi}\right)^{1/2} \int_0^{1/x} t^{-3/2} \exp\left(-\frac{b(t-a)^2}{2a^2 t}\right) dt$$
$$\leq \left(\frac{b}{2\pi}\right)^{1/2} e^{b/a} \int_0^{1/x} t^{-3/2} e^{-b/(2t)}\, dt$$
$$\leq 216(b\sqrt{e})^{-3} \left(\frac{b}{2\pi}\right)^{1/2} e^{b/a} \int_0^{1/x} e^{-b/(4t)}\, dt.$$

Finally, for any $x \geq 1$ and $0 < t < 1$,

$$G_\sigma\left(x^{-1} \leq \sigma \leq x^{-1}(1+t)\right) \geq \left(\frac{b}{2\pi}\right)^{1/2} e^{b/a - b/a^2} \int_{x^{-1}}^{x^{-1}(1+t)} e^{-b/(2t)}\, dt. \qquad \blacksquare$$

Location-scale mixtures of Gaussians

A symmetric Gamma process location-scale mixture of Gaussians prior Π is the distribution of the random function $f(x) := \int \varphi((x-\mu)/\sigma)\, dM(\sigma, \mu)$ where $M \sim \Pi_\alpha$, with α a finite positive measure on $(0, \infty) \times \mathbb{R}$ and $\varphi(x) := e^{-x^2/2}$ for all $x \in \mathbb{R}$. We focus the attention of the reader on the fact that althought we use the same notations (*i.e.* Π, α) as the previous section, these are different distributions and in the sequel we pay attention as making the context clear enough to avoid confusions.

We restrict our discussion to priors for which $\alpha := \overline{\alpha} G_\sigma \times G_\mu$, with $\overline{\alpha} > 0$ and G_σ, G_μ satisfying the same assumptions as in section 3.2.1.

Hybrid location-scale mixtures of Gaussians

The proof of the results given in the two preceeding sections suggests that neither location or location-scale mixtures can achieve the optimal rates, whatever the nature of the tails of Q_0. We show that we can get better upper bounds by introducing hybrid mixtures.

By a hybrid location-scale mixtures of Gaussians, we mean the distribution Π of the random function $f(x) := \int \varphi((x-\mu)/\sigma)\, dM(\sigma, \mu)$, where $M \sim \Pi_\alpha$, with $\alpha = \overline{\alpha} P_\sigma \times G_\mu$, $\overline{\alpha} > 0$, $P_\sigma \sim \Pi_\sigma$ and G_μ a probability measure satisfying equation (3.8). Here Π_σ is a prior distribution on the space of probability measures (endowed with Borel σ-algebra). We now formulate conditions on Π_σ that are the random analoguous to equations (3.5)

and (3.6). For the same constants a_1, a_2, b_1, b_2 as in section 3.2.1, we consider the existence of positive constants a_4, a_5 such that Π_σ satisfies for $x > 0$ large enough

$$\Pi_\sigma\Big(P_\sigma : P_\sigma(\sigma > x) \geq \exp(-a_1 x^{b_1}/2)\Big) \lesssim \exp(-a_4 x^{b_1}), \tag{3.9}$$

$$\Pi_\sigma\Big(P_\sigma : P_\sigma(\sigma < 1/x) \geq \exp(-a_2 x^{b_2}/2)\Big) \lesssim \exp(-a_5 x^{b_2}). \tag{3.10}$$

As a replacement of equation (3.7), we assume that for all $r \geq 1$ there are constants a_6, b_7 such that for any positive integer J large enough

$$\Pi_\sigma\left(\cap_{j=0}^{J}\{P_\sigma \,:\, P_\sigma[2^{-j}, 2^{-j}(1+2^{-Jr})] \geq 2^{-J}\}\right) \gtrsim \exp(-a_6 J^{b_7} 2^J). \tag{3.11}$$

Equations (3.9) to (3.11) are rather restrictive and it is not clear *a priori* whether or not such distribution exists. For example, if P_σ is chosen to be almost-surely an Inverse-Gaussian distribution with parameters b, μ then equation (3.11) is not satisfied. However, we now show that under conditions on the base measure, Π_σ can be chosen as a Dirichlet Process, hereafter referred to as DP.

We recall that if Π_σ is a Dirichlet Process distribution with base measure $\alpha_\sigma G(\cdot)$ on $(0, \infty)$ (Ferguson, 1973), then $P_\sigma \sim \Pi_\sigma$ is a random probability measure on $(0, \infty)$ such that for any Borel measurable partition $A_1, \ldots, A_k$ of $(0, \infty)$, the joint distribution of the random variables $P_\sigma(A_1), \ldots, P_\sigma(A_k)$ is the k-variate Dirichlet distribution with parameters $\alpha_\sigma G(A_1), \ldots, \alpha_\sigma G(A_k)$.

Proposition 3.2 – *Let $\alpha_\sigma > 0$, G_σ a probability measure on $(0, \infty)$ satisfying the same assumptions as in equations (3.5) to (3.7), and Π_σ be a Dirichlet Process with base measure $\alpha_\sigma G_\sigma(\cdot)$. Then Π_σ satisfies equations (3.9) to (3.11) with constants $a_4 = a_1$, $a_5 = a_2$, a constant $a_6 > 0$ eventually depending on r, and $b_7 = 0$.*

Proof. We first prove equation (3.9). It follows from the definition of the DP that $P_\sigma(x, \infty)$ has Beta distribution with parameters $\alpha_\sigma G_\sigma(x, \infty)$ and $\alpha_\sigma(1 - G_\sigma(x, \infty))$, then by Markov's inequality

$$\Pi_\sigma\Big(P_\sigma : P_\sigma(x, \infty) \geq t\Big) \leq \frac{G_\sigma(x, \infty)}{t}.$$

Likewise, if $t = \exp(-a_1 x^{b_1}/2)$ and G_σ satisfies equations (3.5) to (3.7), the conclusion follows. The same steps with $G_\sigma(0, 1/x)$ give the proof of equation (3.10). It remains to prove equation (3.11). Let $r \geq 1$ and define $V_{j,r} := \{\sigma \,:\, 2^{-j} \leq \sigma \leq 2^{-j}(1 + 2^{-Jr})\}$ for any integer $0 \leq j \leq J$. For all $r \geq 1$ the $V_{j,r}$'s are disjoint. Set $V_r^c := \cup_{j=0}^{J} V_{j,r}^c$. If $\alpha_\sigma G_\sigma(V_r^c) \leq 1$ let $V_{J+1,r} = V_r^c$ and $M = 1$; otherwise split V_r^c into $M > 1$ disjoint subsets $V_{1,r}^c, \ldots V_{M,r}^c$ such that $\exp(-2^J) \leq \alpha_\sigma G_\sigma(V_{k,r}^c) \leq 1$ for all $k = 1, \ldots, M$ and set $V_{J+1,r} = V_{1,r}^c$, $V_{J+2,r} = V_{2,r}^c$, $\ldots, V_{J+M,r} = V_{M,r}^c$ (since $G_\sigma(0, \infty) = 1$ this can be done with a number M independent of J). For J large enough (so that $(J + M)2^{-J+1} < 1$), acting as in Ghosal et al. (2000, lemma 6.1), it follows

$$\Pi_\sigma\Big(P_\sigma : P_\sigma[2^{-j}, 2^{-j}(1 + 2^{-Jr}] \geq 2^{-J} \quad \forall\, 0 \leq j \leq J\Big) \geq \frac{\Gamma(\alpha_\sigma) 2^{-J(J+M)}}{\prod_{j=0}^{J+M} \Gamma(\alpha_\sigma G_\sigma(V_{j,r}))},$$

Also, $\alpha_\sigma G_\sigma(V_{j,r}) \leq 1$ implies $\Gamma(\alpha_\sigma G_\sigma(V_{j,r})) \leq 1/(\alpha_\sigma G_\sigma(V_{j,r}))$, hence

$$\Pi_\sigma\Big(P_\sigma \ : \ P_\sigma[2^{-j}, 2^{-j}(1+2^{-Jr}] \geq 2^{-J} \quad \forall\ 0 \leq j \leq J\Big)$$
$$\geq \Gamma(\alpha_\sigma)\alpha_\sigma^{J+M+1} 2^{-J(J+M)} \prod_{j=0}^{J+M} G_\sigma(V_{j,r}).$$

Since M does not depend on J, one can find a constant $C > 0$ such that

$$\Pi_\sigma\Big(P_\sigma : P_\sigma[2^{-j}, 2^{-j}(1+2^{-Jr}] \geq 2^{-J} \quad \forall\ 0 \leq j \leq J\Big)$$
$$\geq \Gamma(\alpha_\sigma) \exp\left\{-CJ^2 + \sum_{j=0}^{J} \log G_\sigma(V_{j,r}) + \sum_{j=J+1}^{J+M} \log G_\sigma(V_{j,r})\right\}.$$

By construction, the second sum in the rhs of the last equation is lower bounded by $-M2^J$, whereas if G_σ satisfies equations (3.5) to (3.7), the first sum is lower bounded by $-C'2^J$ for a constant $C' > 0$ eventually depending on r. Then the proposition is proved. ■

3.2.2. Posterior concentration rates under the mixture priors

We let $\Pi(\cdot \mid \mathbf{y}^n, \mathbf{x}^n)$ denote the posterior distribution of $f \sim \Pi$ based on n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ modelled as in section 3.1. Let $(\epsilon_n)_{n\geq 1}$ be a sequence of positive numbers with $\lim_n \epsilon_n = 0$, and d_n denote the empirical L^2 distance, that is $nd_n(f,g)^2 = \sum_{i=1}^n |f(X_i) - g(X_i)|^2$.

The following theorem is proved in Section section 3.3 .

Theorem 3.1 – *Consider the model (3.3), and assume that $f_0 \in L^1 \cap \mathtt{C}^\beta$ and $Q_0|X|^p < +\infty$. Then there exist a constant $C > 0$ and $t > 0$ depending only on f_0 and Q_0 such that*

- *If the prior Π is the symmetric Gamma location mixture of Gaussians as defined in section 3.2.1*

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-2\beta/(3\beta+1)}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1)$$

when $0 < p \leq 2$, and

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-2\beta/(2\beta+1+2\beta/p)}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1)$$

when $p > 2$.

- *If the prior Π is the symmetric Gamma location-scale mixture of Gaussians defined in section 3.2.1*

$$\Pi\left(d_n(f,f_0)^2 > C[n^{-2\beta/(3\beta+2)} \wedge n^{-2\beta/(2\beta+1+2\beta/p)}](\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1)$$

when $0 < p \leq 2\beta$, and

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-\beta/(\beta+1)}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1),$$

when $p > 2\beta$.

- *If the prior* Π *is the hybrid symmetric Gamma location-scale mixture of Gaussians defined in section 3.2.1*

$$\Pi\left(d_n(f,f_0)^2 > C[n^{-2\beta/(3\beta+1)} \wedge n^{-p/(p+1)}](\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1),$$

when $0 < p \leq 2\beta$ *or*

$$\Pi\left(d_n(f,f_0)^2 > Cn^{-2\beta/(2\beta+1)}(\log n)^t \mid \mathbf{y}^n, \mathbf{x}^n\right) = o_p(1),$$

when $p > 2\beta$.

The upper bounds on the rates in the previous paragraph are no longer valid when $p = 0$. Indeed the constant $C > 0$ depends on p and might not be definite if $p = 0$; the reason is to be found in the fact that C heavily depends on the ability of the prior to draw mixture component in regions of observed data, which remains concentrated near the origin when $p > 0$. In section 3.2.3, we overcome this issue by making the prior covariate dependent ; this allows to derive rates under the assumption $p = 0$ (no tail assumption).

3.2.3. Relaxing the tail assumption : covariate dependent prior for location mixtures

Although the rates derived in section 3.3 do not depend on $p > 0$ when p is small, the assumption $Q_0|X|^p < +\infty$ is crucial in proving the Kullback-Leibler condition. Indeed, this condition ensures that the covariates belong to a set $\mathcal{X}_n$ which is not too large, which allows us to bound from below the prior mass of Kullback-Leibler neighbourhoods of the true distribution. Surprisingly, it seems very difficult to get rid of this assumption under a fully Bayesian framework without fancy assumptions, while making the prior covariates dependent allows to drop all tail conditions on Q_0. Doing so, we can adapt to the tail behaviour of Q_0, as shown in the following theorem, which is an adaptation of the general theorems of Ghosal et al. (2007b). For convenience, in the sequel we drop out the superscript n and we write $\mathbf{x}$, $\mathbf{y}$ for $\mathbf{x}^n$, $\mathbf{y}^n$, respectively. For $\epsilon > 0$ and any subset A of a metric space equipped with metric d, we let $N(\epsilon, A, d)$ denote the ϵ-covering number of A, *i.e.* $N(\epsilon, A, d)$ is the smallest number of balls of radius ϵ needed to cover A.

Theorem 3.2 – *Let* $\Pi_{\mathbf{x}}$ *be a prior distribution that depends on the covariate vector* $\mathbf{x}$*,* $0 < c_2 < 1/4$ *and* $\epsilon_n \to 0$ *with* $n\epsilon_n^2 \to \infty$*. Suppose that* $\mathcal{F}_n \subseteq \mathcal{F}$ *is such that* $Q_0^n \Pi_{\mathbf{x}}(\mathcal{F}_n^c) \lesssim \exp(-\frac{1}{2}(1+2c_2)n\epsilon_n^2)$ *and* $\log N(\epsilon_n/18, \mathcal{F}_n, d_n) \leq n\epsilon_n^2/4$ *for* n *large enough. If for any* $\mathbf{x} \in \mathbb{R}^n$ *it holds* $\Pi_{\mathbf{x}}(f \ : \ d_n(f,f_0) \leq s\epsilon_n) \gtrsim \exp(-c_2 n\epsilon_n^2)$*, then for all* $M > 0$ *we have* $\Pi_{\mathbf{x}}(f \ : \ d_n(f,f_0) > M\epsilon_n \mid \mathbf{y}, \mathbf{x}) = o_p(1)$.

We apply theorem 3.2 to symmetric Gamma process location mixtures of Gaussians in the following way. Let $\mathbb{Q}_{\mathbf{x}}^n$ denote the empirical measure of the covariate vector $\mathbf{x}$. Given a probability density function g, we let $G_{\mathbf{x}}$ the probability measure which density is $z \mapsto \int g(z - x_i)\, d\mathbb{Q}_{\mathbf{x}}^n(x)$.

Corollary 3.1 – *Then we let* $\Pi_{\mathbf{x}}$ *be the distribution of the random function* $f(x) := \int \varphi((x-\mu)/\sigma)\, dM(\mu)$*, where* $\sigma \sim G_\sigma$ *and* $M \sim \Pi_\alpha$ *with* $\alpha = \overline{\alpha} G_{\mathbf{x}}$ *for some* $\overline{\alpha} > 0$*. Assume that* G_σ *satisfies equations (3.5) to (3.7) and that there exists a constant* $b_8 > 0$ *such that* $\sup_{\mathbf{x}\in\mathbb{R}^n} G_{\mathbf{x}}(\mu \ : \ |\mu - s| \leq t) \lesssim t^{b_8}$ *for all* $0 < t, s \leq 1$*. Then* $\Pi_{\mathbf{x}}(f \ : \ d_n(f,f_0) > M\epsilon_n \mid \mathbf{y}, \mathbf{x}) = o_p(1)$ *with* $\epsilon_n^2 \lesssim n^{-2\beta/(3\beta+1)}(\log n)^{2-2\beta/(3\beta+1)}$.

To prove corollary 3.1, note that neither the proof of lemma 3.4 or lemma 3.5 involve the base measure α (indeed, it only involves $\overline{\alpha}$); thus we can use the sieve $\mathcal{F}_n$ constructed in section 3.4.1. To apply theorem 3.2 it is then sufficient to prove that for all $x \in \mathbb{R}^n$

$$\Pi_x(f \ : \ d_n(f, f_0) \leq s\epsilon_n) \gtrsim \exp(-c_2 n\epsilon_n^2). \tag{3.12}$$

This is done in lemma 3.1.

Lemma 3.1 – *Assume that there is a constant $b_8 > 0$ such that $\sup_{\mathbf{x}\in\mathbb{R}^n} G_{\mathbf{x}}(\mu \ : \ |\mu - s| \leq t) \lesssim t^{b_8}$ for all $0 < t, s \leq 1$. Also assume that G_σ satifies equations (3.5) to (3.7). Then equation (3.12) holds for the symmetric Gamma location mixture of Gaussians with base measure $\overline{\alpha}G_{\mathbf{x}}$ if $\epsilon_n^2 \leq Cn^{-2\beta/(3\beta+1)}(\log n)^{2-2\beta/(3\beta+1)}$ for an appropriate constant $C > 0$.*

The proof of lemma 3.1 is given in section 3.B.

3.3. Proofs

To prove theorem 3.1 we follow the lines of Ghosal et al. (2000); Ghosal and Van Der Vaart (2001); Ghosal et al. (2007a). Namely we need to verify the following three conditions

- Kullback-Leibler condition : For a constant $0 < c_2 < 1/4$,

$$\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq e^{-c_2 n\epsilon_n^2}, \tag{3.13}$$

  where

$$\mathtt{KL}(f_0, \epsilon_n) := \left\{ f \ : \ \frac{1}{2s^2} \int |f_0(x) - f(x)|^2 \, dQ_0(x) \leq \epsilon_n^2 \right\}.$$

- Sieve condition : There exists $\mathcal{F}_n \subset \mathcal{F}$ such that

$$\Pi(\mathcal{F}_n^c) \leq e^{-\frac{1}{2}(1+2c_2)n\epsilon_n^2} \tag{3.14}$$

- Tests : Let $N(\epsilon_n/18, \mathcal{F}_n, d_n)$ be the logarithm of the covering number of $\mathcal{F}_n$ with radius $\epsilon_n/18$ in the $d_n(\cdot, \cdot)$ metric.

$$N(\epsilon_n/18, \mathcal{F}_n, d_n) \leq \frac{n\epsilon_n^2}{4}. \tag{3.15}$$

The Kullback-Leibler condition is proved by defining an approximation of f by a discrete mixture under weak tail conditions. Although the general idea is close to Kruijer et al. (2010) or Scricciolo (2014), the construction remains quite different to be able to handle various tail behaviours. This is detailed in the following section.

3.3.1. Approximation theory

To describe the approximation of f_0 by a finite mixture, we first define a few notations.

Let $\widehat{\chi}$ be a $\mathtt{C}^\infty$ function that equals 1 on $[-1, 1]$ and 0 outside $[-2, 2]^c$ (think for instance as the convolution of $\mathbb{1}_{[-1,1]}$ with $x \mapsto \exp(-1/(1-x^2))\,\mathbb{1}_{[-1,1]}(x)$). For any $\sigma > 0$ we use the shortened notation $\widehat{\chi}_\sigma(\xi) := \widehat{\chi}(2\sigma\xi)$. Define η as the function which L^1

Fourier transform satisfies $\widehat{\eta}(\xi) = \widehat{\chi}(\xi)/\widehat{\varphi}(\xi)$ for all $\xi \in [-2,2]$ and $\widehat{\eta}(\xi) = 0$ elsewhere. For two positive real numbers h and σ, we define the kernel $K_{h,\sigma} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ such that

$$K_{h,\sigma}(x,y) := \frac{h}{\sigma}\sum_{k\in\mathbb{Z}} \varphi\left(\frac{x-h\sigma k}{\sigma}\right)\eta\left(\frac{y-h\sigma k}{\sigma}\right), \quad \forall (x,y)\in\mathbb{R}\times\mathbb{R}.$$

For a measurable function f we introduce the operator associated with the kernel : $K_{h,\sigma}f(x) = \int K_{h,\sigma}(x,y)f(y)\,dy$. The function $K_{h,\sigma}f$ will play the role of an *approximation* for the function f, and we will evaluate how this approximation becomes close to f given h and σ sufficiently close to zero.

More precisely, we will prove that, when choosing h appropriately, f can be approximated by $K_{h,\sigma}(\chi_\sigma * f_0)$ to the order σ^β. Moreover $K_{h,\sigma}(\chi_\sigma * f_0)$ can be written as $\sum_{k\in\mathbb{Z}} u_k\varphi((x-\mu_k)/\sigma))$. In a second step we approximate $K_{h,\sigma}(\chi_\sigma * f_0)$ by a truncated version of it, retaining only the k's such that $|u_k|$ is large enough and $|\mu_k|$ not too large. In the case of location-scale and hybrid location-scale mixtures we consider a modification of this approximation to control better the number of components for which σ needs to be small. We believe that these constructions have interest in themselves. In particular they shed light on the relations between Gaussian mixtures and wavelet approximations.

These approximation properties are presented in the following two Lemmas which are proved in section 3.A:

Lemma 3.2 – *There is $C > 0$ depending only on β such that for any $f_0 \in L^1 \cap \mathtt{C}^\beta$ and any $\sigma > 0$ we have $|\chi_\sigma * f_0(x) - f_0(x)| \le C\|f\|_{\mathtt{C}^\beta}\sigma^\beta$ for all $x \in \mathbb{R}$.*

Lemma 3.3 – *Let $f_\sigma := \chi_\sigma * f_0$ and $h \le 1$. Then there is a universal constant $C > 0$ such that $|K_{h,\sigma}f_\sigma(x) - f_\sigma(x)| \le C\|f_0\|_1\sigma^{-1}e^{-4\pi^2/h^2}$ for all $x \in \mathbb{R}$.*

We now present the approximation schemes in the context of location mixtures.

3.3.2. Construction of the approximation under location mixtures

Let $0 < \sigma \le 1$ and $h_\sigma\sqrt{\log\sigma^{-1}} := 2\pi\sqrt{\beta+1}$. Then combining the results of lemma 3.2 and lemma 3.3 we can conclude that $|K_{h_\sigma,\sigma}(\chi_\sigma * f_0)(x) - f_0(x)| \lesssim \sigma^\beta$. Now we define the coefficients u_k, $k \in \mathbb{Z}$ so that

$$K_{h_\sigma,\sigma}(\chi_\sigma * f_0)(x) =: \sum_{k\in\mathbb{Z}} u_k\,\varphi\left(\frac{x-\mu_k}{\sigma}\right), \quad \forall k \in \mathbb{Z},$$

where $\mu_k := h_\sigma\sigma k$ for all $k \in \mathbb{Z}$. Let define

$$\Lambda := \left\{k\in\mathbb{Z} \,:\, |u_k| > \sigma^\beta, \quad |\mu_k| \le \sigma^{-2\beta/p} + \sigma\sqrt{2(\beta+1)\log\sigma^{-1}}\right\},$$

$U_\sigma := \{\sigma' \,:\, \sigma \le \sigma' \le \sigma(1+\sigma^\beta)\}$, and for all $k \in \Lambda$ we define $V_k := \{\mu \,:\, |\mu-\mu_k| \le \sigma^{\beta+1}\}$ and $V = \cup_{k\in\Lambda}V_k$. We also denote

$$\mathcal{M}_\sigma := \left\{M \text{ signed measure on } \mathbb{R} \,:\, \begin{array}{l} |M(V_k) - u_k| \le \sigma^\beta, \\ \forall k \in \Lambda \,:\, |M|(V^c) \le \sigma^\beta \end{array}\right\},$$

and for any $M \in \mathcal{M}_\sigma$, we write $f_{M,\sigma}(x) := \int \varphi((x-\mu)/\sigma)\,dM(\mu)$.

Proposition 3.3 – *For $\sigma > 0$ small enough, it holds $|\Lambda| \lesssim \sigma^{-(\beta+1)} \wedge h_\sigma^{-1}\sigma^{-(2\beta/p+1)}$.*

Proof. Because there is a separation of $h_\sigma\sigma$ between two consecutive μ_k, it is clear that $|\Lambda| \leq 2h_\sigma^{-1}\sigma^{-(2\beta/p+1)}$. Moreover, from proposition 3.9 we have the following estimate.

$$\|f_0\|_1\sigma^{-1} \gtrsim \sum_{k\in\mathbb{Z}} |u_k| \geq \sum_{k\in\Lambda} |u_k| \geq \sigma^\beta|\Lambda|. \qquad \blacksquare$$

Proposition 3.4 – *For all $x \in \mathbb{R}$, all $\sigma > 0$ small enough and all $M \in \mathcal{M}_\sigma$ it holds $|f_{M,\sigma}(x) - f_0(x)| \lesssim h_\sigma^{-1}$.*

Proof. For any $M \in \mathcal{M}_\sigma$, we have that $|f_{M,\sigma}(x) - f_0(x)| \leq |f_{M,\sigma}(x)| + \|f_0\|_\infty$. But, with $\mathcal{I} \equiv \mathcal{I}(x) := \{k \in \mathbb{Z} : |x - \mu_k| \leq 2\sigma\}$,

$$f_{M,\sigma}(x) = \sum_{k\in\Lambda\cap\mathcal{I}} \int_{V_k} \varphi\left(\frac{x-\mu}{\sigma}\right) dM(\mu) + \sum_{k\in\Lambda\cap\mathcal{I}^c} \int_{V_k} \varphi\left(\frac{x-\mu}{\sigma}\right) dM(\mu) + \int_{V^c} \varphi\left(\frac{x-\mu}{\sigma}\right) dM(\mu). \quad (3.16)$$

Clearly the last term of this last expression is bounded above by $\|\varphi\|_\infty\sigma^\beta$. For the second term, we have for any $\mu \in V_k$ with $k \in \mathcal{I}^c$ that $|x - \mu| \geq |x - \mu_k| - |\mu - \mu_k| \geq |x - \mu_k|/2$. Then the second term of the rhs of equation (3.16) is bounded above by

$$\sup_{k\in\Lambda\cap\mathcal{I}^c} |M|(V_k) \sum_{k\in\mathbb{Z}} \varphi\left(\frac{x - h_\sigma\sigma k}{\sigma}\right).$$

Proceeding as in the proof of lemma 3.9, we deduce that the series in the last expression is bounded above by a constant times $1/h_\sigma$, whereas proposition 3.9 and Young's inequality yields $|M|(V_k) \leq |M(V_k) - u_k| + |u_k| \lesssim \sigma^\beta + \|\chi_\sigma * f_0\|_\infty \leq \sigma^\beta + \|\chi\|_1\|f_0\|_\infty$. Therefore the second term of the rhs in equation (3.16) is bounded by a constant multiple of h_σ^{-1}. Regarding the first term in equation (3.16), it is bounded by $\|\varphi\|_\infty|\mathcal{I}| \sup_{k\in\Lambda} |M|(V_k)$, which is in turn bounded by h_σ^{-1} times a constant. $\blacksquare$

Proposition 3.5 – *For all $\sigma > 0$ small enough, all $x \in \mathbb{R}$ with $|x| \leq \sigma^{-2\beta/p}$ and all $M \in \mathcal{M}_\sigma$ it holds $|f_{M,\sigma}(x) - f_0(x)| \lesssim h_\sigma^{-2}\sigma^\beta$.*

Proof. We define $A_\sigma(\beta) := \sqrt{2\log|\Lambda| + 2(\beta+1)\log\sigma^{-1}}$. Then for any $M \in \mathcal{M}_\sigma$, letting $\mathcal{J} \equiv \mathcal{J}(x) := \{k \in \mathbb{Z} : |x - \mu_k| \leq 2\sigma A_\sigma(\beta)\}$, we may write

$$\begin{aligned} f_{M,\sigma}(x) - K_{h_\sigma,\sigma}(\chi_\sigma * f_0)(x) &= \sum_{k\in\Lambda\cap\mathcal{J}} \int_{V_k} \left[\varphi\left(\frac{x-\mu}{\sigma}\right) - \varphi\left(\frac{x-\mu_k}{\sigma}\right)\right] dM(\mu) \\ &+ \sum_{k\in\Lambda\cap\mathcal{J}} [M(V_k) - u_k]\,\varphi\left(\frac{x-\mu_k}{\sigma}\right) + \sum_{k\in\Lambda\cap\mathcal{J}^c} \int_{V_k} \varphi\left(\frac{x-\mu}{\sigma}\right) dM(\mu) \\ &- \sum_{k\in\Lambda\cap\mathcal{J}^c} u_k\,\varphi\left(\frac{x-\mu_k}{\sigma}\right) - \sum_{k\in\Lambda^c} u_k\,\varphi\left(\frac{x-\mu_k}{\sigma}\right) + \int_{V^c} \varphi\left(\frac{x-\mu}{\sigma}\right) dM(\mu) \\ &:= r_1(x) + r_2(x) + r_3(x) + r_4(x) + r_5(x) + r_6(x). \quad (3.17) \end{aligned}$$

With the same argument as in proposition 3.3, we deduce that $|\mathcal{J}| \leq 2h_\sigma^{-1}A_\sigma(\beta)$. The same proposition implies $A_\sigma(\beta) \lesssim \sqrt{\log \sigma^{-1}}$. Recalling that $|M|(V_k) \lesssim 1+\|\chi\|_1\|f_0\|_\infty$ for all $k \in \Lambda$ and all $M \in \mathcal{M}_\sigma$, it follows from proposition 3.11 that $|r_1(x)| \lesssim A_\sigma(\beta)h_\sigma^{-1}\sigma^\beta$. From the definition of $\mathcal{M}_\sigma$, it comes $|r_2(x)| \leq \|\varphi\|_\infty|\mathcal{J}|\sigma^\beta \leq 2\|\varphi\|_\infty A_\sigma(\beta)h_\sigma^{-1}\sigma^\beta$. Whenever $k \in \Lambda \cap \mathcal{J}^c$ and $\mu \in V_k$, it holds $|x-\mu| \geq |x-\mu_k| - |\mu-\mu_k| \geq \sigma A_\sigma(\beta)$. Therefore, $|r_3(x)| \lesssim \varphi(A_\sigma(\beta))|\Lambda| \lesssim \sigma^{\beta+1}$. With the same argument, proposition 3.9 and Young's inequality we get $|r_4(x)| \lesssim \|\chi_\sigma * f_0\|_\infty\varphi(2A_\sigma(\beta))|\Lambda| \leq \|\chi\|_1\|f_0\|_\infty\sigma^\beta$. Regarding r_5, we rewrite $\Lambda^c = \Lambda_1^c \cup \Lambda_2^c$, with $\Lambda_1^c := \{k \in \mathbb{Z} : |u_k| \leq \sigma^\beta\}$ and $\Lambda_2^c := \{k \in \mathbb{Z} : |\mu_k| > \sigma^{-2\beta/p} + \sigma\sqrt{2(\beta+1)\log\sigma^{-1}}\}$. Then,

$$\begin{aligned}|r_5(x)| &\leq \sum_{k\in\Lambda_1^c} |u_k|\,\varphi\left(\frac{x-\mu_k}{\sigma}\right) + \sum_{k\in\Lambda_2^c} |u_k|\,\varphi\left(\frac{x-\mu_k}{\sigma}\right)\\ &\leq \sigma^\beta \sup_{x\in\mathbb{R}} \sum_{k\in\mathbb{Z}} \varphi\left(\frac{x-\mu_k}{\sigma}\right) + \sum_{k\in\Lambda_2^c} |u_k|\,\varphi\left(\frac{x-\mu_k}{\sigma}\right). \qquad (3.18)\end{aligned}$$

The first term of the rhs of equation (3.18) is bounded by a multiple constant of $h_\sigma^{-1}\sigma^\beta$, with the same argument as in the proof of lemma 3.9. By definition of Λ_2^c, $|x-\mu_k| \geq \sigma\sqrt{2(\beta+1)\log\sigma^{-1}}$ when $k \in \Lambda_2^c$ and $|x| \leq \sigma^{-2\beta/p}$. This implies, together with proposition 3.9 and Young's inequality, that the second term of the rhs of equation (3.18) is bounded by a constant multiple of $\sigma^{\beta+1}\sum_{k\in\mathbb{Z}}|u_k| \lesssim \|\chi_\sigma * f_0\|_1\sigma^\beta \leq \|\chi\|_1\|f_0\|_1\sigma^\beta$ for all $|x| \leq \sigma^{-2\beta/p}$. Finally, we have the trivial bound $|r_6(x)| \leq \|\varphi\|_\infty|M|(V^c) \leq \|\varphi\|_\infty\sigma^\beta$. ■

3.3.3. Construction of the approximation under location-scale and hybrid location-scale mixtures

Let $\sigma_0 := 1$ and define recursively $\sigma_{j+1} := \sigma_j/2$ for any $j \geq 0$. Let $\Delta_0 := f_0 - \chi_{\sigma_0} * f_0$, and define recursively $\Delta_{j+1} := \Delta_j - \chi_{\sigma_{j+1}} * \Delta_j$, for any $j \geq 0$.

The general idea of the construction is that $|\Delta_j| \lesssim \sigma_j^\beta$, as shown in proposition 3.10 in appendix, and that similarly to wavelet decomposition, we approximate a function f_0 Hölder β by

$$f_1 := K_0(\chi_{\sigma_0} * f_0) + \sum_{j=1}^{J} K_j(\chi_{\sigma_j} * \Delta_{j-1}).$$

where $J \geq 1$ is a large enough integer, $h_J\sqrt{J} := 2\pi/\sqrt{\beta\log 2}$, and $K_j := K_{h_J,\sigma_j}$. By induction, we get that $\Delta_j = \Delta_0 - \sum_{l=0}^{j-1}\chi_{\sigma_{l+1}} * \Delta_l$. It follows,

$$\begin{aligned}f_1 - f_0 &= K_0(\chi_{\sigma_0} * f_0) - f_0 + \sum_{j=1}^{J} K_j(\chi_{\sigma_j} * \Delta_{j-1})\\ &= \Delta_J + K_0(\chi_{\sigma_0} * f_0) - \chi_{\sigma_0} * f_0 + \sum_{j=1}^{J}\left[K_j(\chi_{\sigma_j} * \Delta_{j-1}) - \chi_{\sigma_j} * \Delta_{j-1}\right].\end{aligned}$$

Therefore, from lemma 3.3 and proposition 3.10 and Young's inequality, the error of

approximating f_0 by f_1 is

$$
\begin{aligned}
&|f_1(x) - f_0(x)| \\
&\qquad \le |\Delta_J| + |K_0(\chi_{\sigma_0} * f_0) - \chi_{\sigma_0} * f_0| + \sum_{j=1}^{J} |K_j(\chi_{\sigma_j} * \Delta_{j-1}) - \chi_{\sigma_j} * \Delta_{j-1}| \\
&\qquad \lesssim \|f_0\|_{\mathcal{C}^\beta} \sigma_J^\beta + \|\chi_{\sigma_0} * f_0\|_1 \sigma_0^{-1} e^{-4\pi^2/h_J^2} + e^{-4\pi^2/h_J^2} \sum_{j=1}^{J} \|\chi_{\sigma_j} * \Delta_{j-1}\| \sigma_j^{-1} \\
&\qquad \lesssim \|f_0\|_{\mathcal{C}^\beta} \sigma_J^\beta + \|f\|_1 e^{-4\pi^2/h_J^2} + \|f_0\|_1 e^{-4\pi^2/h_J^2} \sum_{j=1}^{J} 2^j \\
&\qquad \lesssim \|f_0\|_{\mathcal{C}^\beta} \sigma_J^\beta + \|f_0\|_1 (1 + 2^J) e^{-4\pi^2/h_J^2} \lesssim \sigma_J^\beta.
\end{aligned}
$$

The reason for considering different scale parameters in the construction, is to deal with fat tail, the heuristic being that in the tail we do not require as precise an approximation as in the center. In particular small values of j will be used to estimate the function far off in the tails. To formalize this, we define $\zeta_j := 2^{(J-j)(2\beta/p)}$, and $A_j := [-\zeta_j, \zeta_j]$, for all $j = 0, \dots J$. We also define $I_J = [-1, 1]$, and for all $j = 0, \dots, J-1$ we set $I_j := A_j \backslash A_{j+1}$. Notice that by definition of K_j, we can write,

$$
\begin{aligned}
K_0(\chi_{\sigma_0} * f_0)(x) &=: \sum_{k \in \mathbb{Z}} u_{0k}\, \varphi((x - h_J \sigma_0 k)/\sigma_0) \\
K_j(\chi_{\sigma_j} * \Delta_{j-1})(x) &=: \sum_{k \in \mathbb{Z}} u_{jk}\, \varphi((x - h_J \sigma_j k)/\sigma_j), \quad \forall j \ge 1.
\end{aligned}
$$

To ease notation, we define $\mu_{jk} := h_J \sigma_j k$ for all $j \ge 0$ and all $k \in \mathbb{Z}$. In the sequel we shall need the following subset of indexes,

$$
\Lambda := \left\{ (j,k) \in \{0, \dots, J\} \times \mathbb{Z} \,:\, |u_{jk}| > \sigma_J^\beta, \quad |\mu_{jk}| \le \zeta_j + \sqrt{2(\beta+1)\log \sigma_J^{-1}} \right\}.
$$

We prove below that we can approximate f_1 by a finite mixture corresponding to retaining only the components associated to indices in Λ and that we can bound the cardinality of Λ by $O(J \log J \sigma_J^{-2\beta/p})$

To any $(j,k) \in \Lambda$ we associate $U_j := \{\sigma \,:\, \sigma_j \le \sigma \le \sigma_j(1 + \sigma_J^\beta)\}$, $V_{jk} := \{\mu \,:\, |\mu - \mu_{jk}| \le \sigma_j \sigma_J^\beta\}$ and $W_{jk} := U_j \times V_{jk}$. We denote by $\mathcal{M}$ the set of signed measures M on $(0, \infty) \times \mathbb{R}$ such that $|M(W_{jk}) - u_{jk}| \le \sigma_J^\beta$ for all $(j,k) \in \Lambda$, and $|M|(W^c) \le \sigma_J^\beta$, where W^c is the relative complement of the union of all W_{jk} for $(j,k) \in \Lambda$. For any $M \in \mathcal{M}$, we write

$$
f_M(x) := \int \varphi((x - \mu)/\sigma)\, dM(\sigma, \mu).
$$

In proposition 3.6 we control the cardinality of Λ while in proposition 3.8 we control the error between f_M and f_1 on the decreasing sequence of intervals $[-\zeta_j, \zeta_j]$. Proposition 3.7 provides a crude uniform upper bound on f_M and f_0.

Proposition 3.6 – *There is a constant $C > 0$ depending only on f_0 and Q_0 such that $|\Lambda| \le C[\sigma_J^{-(\beta+1)} \wedge (J \log J)\sigma_J^{-2\beta/p}]$ if $p \le 2\beta$, and $|\Lambda| \le C(J \log J)\sigma_J^{-1}$ if $p > 2\beta$.*

Proof. First notice that because of propositions 3.9 and 3.10, we always have the bound

$$4\|f_0\|_1\sigma_J^{-1} \geq 2\|f_0\|_1 \sum_{j=0}^{J}\sigma_j^{-1} \geq \sum_{j=0}^{J}\sum_{k\in\mathbb{Z}}|u_{jk}| \geq \sum_{(j,k)\in\Lambda}|u_{jk}| \geq \sigma_J^{\beta}|\Lambda|. \tag{3.19}$$

If $p \leq 2\beta$, we define $B := \sqrt{2(\beta+1)\log 2}$, so that $\sqrt{2(\beta+1)\log\sigma_J^{-1}} = B\sqrt{J}$. Now consider those indexes j with $\zeta_j \leq B\sqrt{J}$. An elementary computation shows that there are at most $\lesssim \log J$ such indexes. Therefore, recalling that there is a separation of $h_J\sigma_j$ between two consecutive μ_{jk} and that there are at most J indexes j with $\zeta_j > B\sqrt{J}$

$$\begin{aligned}|\Lambda| &\lesssim \sum_{j=0}^{J}\frac{4\zeta_j}{h_J\sigma_j} + \log J \times \frac{2B\sqrt{J}}{h_J\sigma_J} \\ &\leq 4h_J^{-1}\sigma_J^{-2\beta/p}\sum_{j=0}^{J}2^{-j(\frac{2\beta}{p}-1)} + 2B(\sqrt{J}\log J)h_J^{-1}\sigma_J^{-1}. \end{aligned}\tag{3.20}$$

Because $h_J\sqrt{J} \lesssim 1$ by definition, and because $p \leq 2\beta$, the result follows from the last equation and equation (3.19). If $p > 2\beta$, the reasoning is the same as in the first part, but we can rewrite in this situation the equation (3.20) as

$$|\Lambda| \leq 4h_J^{-1}\sigma_J^{-1}\sum_{j=0}^{J}2^{(j-J)(1-\frac{2\beta}{p})} + 2B(\sqrt{J}\log J)h_J^{-1}\sigma_J^{-1}.$$

Since $p > 2\beta$, the conclusion is immediate. ■

Proposition 3.7 – *For all $x \in \mathbb{R}$, all $J > 0$ large enough and all $M \in \mathcal{M}$, it holds $|f_M(x) - f_0(x)| \lesssim J^{3/2}$.*

Proof. Let $\mathcal{I} \equiv \mathcal{I}(x) := \{(j,k) \in \{0,\dots,J\}\times\mathbb{Z} : |x-\mu_{jk}| \leq 2\sigma_j\}$. Then the proof is almost identical to proposition 3.4. It suffices to notice that

- $|M|(W_{jk}) \leq |M(W_{jk}) - u_{jk}| + |u_{jk}|$ is always bounded above by a constant, because of the definition of $\mathcal{M}$, of propositions 3.9 and 3.10.
- $|x-\mu|/\sigma \geq (1/4)|x-\mu_{jk}|/\sigma_j$ whenever $(\sigma,\mu) \in W_{jk}$ and $(j,k) \in \Lambda \cap \mathcal{I}^c$, as soon as J is large enough.
- $|\mathcal{I}| \leq 5Jh_J^{-1}$ for $J \geq 1$. ■

Proposition 3.8 – *If $f_0 \in \mathcal{C}_\beta$, for all $J > 0$ large enough, all $0 \leq j \leq J$, all $x \in [-\zeta_j, \zeta_j]$ and all $M \in \mathcal{M}$, it holds $|f_M(x) - f_0(x)| \lesssim J^{3/2}\sigma_j^{\beta}$.*

The proof of proposition 3.8 is given in section 3.C.

3.4. Proof of theorem 3.1

As mentioned earlier, the proof of theorem 3.1 boils down to verifying conditions (3.13), (3.14) and (3.15) for the three types of priors.

3.4.1. Case of the location mixture

Kullback-Leibler condition for location mixtures

In this Section we verify condition (3.13) in the case of the location mixture prior, using the results of section 3.3.2

By Chebychev inequality, we have $Q_0[-\sigma^{-2\beta/p}, \sigma^{2\beta/p}]^c \leq \sigma^{2\beta} Q_0|X|^p$. Then by bringing together results from propositions 3.4 and 3.5, we can find a constant $C > 0$ such that for all $M \in \mathcal{M}_\sigma$

$$\begin{aligned}\int |f_{M,\sigma}(x) - f_0(x)|^2 \, dQ_0(x) \leq & \sup_{|x|>\sigma^{-2\beta/p}} |f_{M,\sigma}(x) - f_0(x)|^2 \, Q_0[-\sigma^{-2\beta/p}, \sigma^{-2\beta/p}]^c \\ & + \sup_{|x|\leq\sigma^{-2\beta/p}} |f_{M,\sigma}(x) - f_0(x)|^2 \\ & \leq C\sigma^{2\beta}(\log\sigma^{-1})^2.\end{aligned}$$

By equation (3.7), we have $G_\sigma(U_\sigma) \gtrsim \sigma^{-b_3}\sigma^{b_4\beta}\exp(-a_3/\sigma)$. Moreover, there is a separation of $h_\sigma\sigma$ between two consecutive μ_k and $h_\sigma\sigma \ll \sigma$, thus all the V_k with $k \in \Lambda$ are disjoint. By assumptions on G_μ (see equation (3.8)), $\alpha_k := \overline{\alpha} G_\mu(V_k) \gtrsim \sigma^{b_5(\beta+1)}(1+|\mu_k|)^{-b_6}$ for all $k \in \Lambda$. We also define $\alpha^c := \alpha(V^c)$. For σ small enough, there is a constant $C' > 0$ not depending on σ such that $\alpha^c > C'$. Moreover, since α has finite variation we can assume without loss of generality that $C' \leq \alpha^c \leq 1$, otherwise we split V^c into disjoint parts, each of them having α-measure smaller than one. With $\epsilon_n^2 := C\sigma^{2\beta}(\log\sigma^{-1})^2$, using that $\Gamma(\alpha) \leq 2\alpha^{\alpha-1}$ for $\alpha \leq 1$, it follows the lower bound

$$\begin{aligned}\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq G_\sigma(U_\sigma)\Pi_\alpha(\mathcal{M}_\sigma) &\gtrsim \sigma^{-b_3+b_4\beta} e^{-a_3\sigma^{-1}} \frac{\sigma^\beta}{3e\Gamma(\alpha^c)} \prod_{k\in\Lambda} \left(\frac{\sigma^\beta e^{-2|u_k|}}{3e\Gamma(\alpha_k)}\right) \\ &\gtrsim \exp\left\{-K|\Lambda|\log\sigma^{-1} - a_3\sigma^{-1} - 2\sum_{k\in\Lambda}|u_k| - \sum_{k\in\Lambda}\log\frac{1}{\alpha_k}\right\} \\ &\gtrsim \exp\left\{-K|\Lambda|\log\sigma^{-1} - K\sigma^{-1} - \sum_{k\in\Lambda}\log\frac{1}{\alpha_k}\right\},\end{aligned}$$

for a generic constant $K > 0$. From the definition of α_k, it holds

$$\sum_{k\in\Lambda}\log\frac{1}{\alpha_k} \lesssim |\Lambda|\log\sigma^{-1} + \sum_{k\in\Lambda}\log(1+|\mu_k|),$$

when σ is small enough. Also,

$$\begin{aligned}\sum_{k\in\Lambda}\log(1+|\mu_k|) &= \sum_{k\in\Lambda}\log(1+|\mu_k|)\,\mathbb{1}\{|\mu_k| \leq 1\} + \sum_{k\in\Lambda}\log(1+|\mu_k|)\,\mathbb{1}\{|\mu_k| > 1\} \\ &\leq |\{k\in\Lambda : |\mu_k| \leq 1\}| + |\Lambda|\log 2 + \sum_{k\in\Lambda}\log|\mu_k| \\ &\leq 2h_\sigma^{-1}\sigma^{-1} + 4|\Lambda|\frac{2\beta}{p}\log\sigma^{-1} \lesssim |\Lambda|\log\sigma^{-1} + \sigma^{-1}\end{aligned}$$

Because $|\Lambda| > \sigma^{-1}$ for σ small enough, it follows from all of the above the existence of a constant $K' > 0$, depending only on f, φ and Π, such that

$$\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq \exp\left\{-K'|\Lambda| \log \sigma^{-1}\right\}.$$

Then for an appropriate constant $C''' > 0$, as a consequence of proposition 3.3, we can have $\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq e^{-c_2 n \epsilon_n^2}$ if

$$\epsilon_n^2 = \begin{cases} C''' n^{-2\beta/(3\beta+1)} (\log n)^{2-2\beta/(3\beta+1)} & 0 < p \leq 2, \\ C''' n^{-2\beta/(2\beta+1+2\beta/p)} (\log n)^{2-3\beta/(2\beta+1+2\beta/p)} & p > 2. \end{cases}$$

Sieve construction for location mixtures

We construct the following sequence of subsets of $\mathcal{F}$, also called *a sieve*. With the notation $f_{M,\sigma}(x) := \int \varphi((x-\mu)/\sigma)\, dM(\mu)$,

$$\mathcal{F}_n(H, \epsilon) := \left\{ f = f_{M,\sigma} \,:\, \begin{array}{l} M = \sum_{i=1}^{\infty} u_i \delta_{\mu_i}, \quad n^{-1/b_2} < \sigma \leq n^{1/b_1} \\ \sum_{i=1}^{\infty} |u_i| \leq n, \quad \sum_{i=1}^{\infty} |u_i| \, \mathbb{1}\{|u_i| \leq n^{-1}\} \leq \epsilon \\ |\{i \,:\, |u_i| > n^{-1}\}| \leq H n \epsilon^2 / \log n \end{array} \right\}.$$

The next two lemmas show that $\mathcal{F}_n(H, \epsilon)$ defined as above satisfies all the condition stated in equations (3.14) and (3.15) if H and δ are chosen small enough.

Lemma 3.4 – *Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be arbitrary and d_n be the empirical L^2-distance associated with $\mathbf{x}$. Then for any $n^{-1/2} < \epsilon_n \leq 1$, $0 < H \leq 1$ and n sufficiently large there is a constant $C > 0$ not depending on n such that $\log N(\epsilon_n, \mathcal{F}_n(H, \epsilon_n), d_n) \leq C H n \epsilon_n^2$.*

Proof. We write $\mathcal{F}_n \equiv \mathcal{F}_n(H, \epsilon_n)$ to ease notations. The proof is based on arguments from Shen et al. (2013), it uses the fact that the covering number $N(\epsilon_n, \mathcal{F}_n, d_n)$ is the minimal cardinality of an ϵ_n-net over $(\mathcal{F}_n, d_n)$. We recall that $(\mathcal{F}_n, d_n)$ has ϵ_n-net $\mathcal{F}_{n,\epsilon}$, if for any $f \in \mathcal{F}_n$ we have $m \in \mathcal{F}_{n,\epsilon}$ such that $d_n(f, m) < \epsilon_n$. Let $S_n := \cup_{i=1}^n \{x \,:\, |x - x_i| \leq n^{1/b_1}\sqrt{6 \log n}\}$, $R_n := \{\mu \in \mathbb{R} \,:\, \mu = k/n^{3/2+1/b_2},\ k \in \mathbb{Z},\ \mu \in S_n\}$ and,

$$\mathcal{F}_{n,\epsilon} := \left\{ f = \textstyle\sum_{i \in \mathcal{I}} u_i \, \varphi\left(\frac{\cdot - \mu_i}{\sigma}\right) \,:\, \begin{array}{l} |\mathcal{I}| \leq H n \epsilon_n^2 / \log n,\ n^{-1/b_2} \leq \sigma \leq n^{1/b_1} \\ \forall i \in \mathcal{I} : |u_i| \leq n,\ \mu_i \in R_n \\ u_i = k n^{-3/2} H^{-1},\ k \in \mathbb{Z}, \\ \sigma = k/n^{3/2+1/b_2},\ k \in \mathbb{N}, \end{array} \right\}.$$

We claim that there is a constant $\delta > 0$ such that $\mathcal{F}_{n,\epsilon}$ is a $\delta\epsilon$-net over $(\mathcal{F}_n, d_n)$. Indeed, let $f \in \mathcal{F}_n$ be arbitrary, so that $f = \sum_{i=1}^{\infty} u_i \varphi((\cdot - \mu_i)/\sigma)$. We define $\mathcal{J} := \mathbb{N} \cup \{\infty\}$, $\mathcal{K} := \{i \,:\, |u_i| > n^{-1}\}$, and $\mathcal{L} := \{i \,:\, \mu_i \in S_n\}$. Now choose $\mathcal{I} = \mathcal{J} \cap \mathcal{K} \cap \mathcal{L}$, and notice that $|\mathcal{I}| \leq |\mathcal{K}| \leq H n \epsilon_n^2 / \log n$. Hence we can pick a $m \in \mathcal{F}_{n,\epsilon}$ with $m(x) = \sum_{i \in \mathcal{I}} u_i' \varphi((x - \mu_i')/\sigma')$. Moreover, for any $j = 1, \ldots, n$

$$\begin{aligned} |f(x_j) - m(x_j)| \leq & \sum_{\mathcal{J} \cap \mathcal{K} \cap \mathcal{L}^c} |u_i| \varphi((x_j - \mu_i)/\sigma) + \sum_{\mathcal{J} \cap \mathcal{K}^c} |u_i| \varphi((x_j - \mu_i)/\sigma) \\ & + \sum_{i \in \mathcal{I}} |u_i| |\varphi((x_j - \mu_i)/\sigma) - \varphi((x_j - \mu_i')/\sigma')| \\ & + \sum_{i \in \mathcal{I}} |u_i - u_i'| \, \varphi((x_j - \mu_i')/\sigma'). \end{aligned}$$

The fourth term in the rhs of the last equation is bounded above by ϵ_n. Regarding the third term, for any $i \in \mathcal{L}^c$ we have $|x_j - \mu_i|/\sigma > \sqrt{6 \log n}$ for all $j = 1, \ldots, n$. Then the third term is bounded by $|\mathcal{K}| n \varphi(\sqrt{6 \log n}) \leq H n \epsilon_n^2 n^{-2} / \log n \leq \epsilon_n$. Since we can always choose $m \in \mathcal{F}_{n,\epsilon}$ with $|u_i - u_i'| \leq n^{-3/2} H^{-1}$ for all $i \in \mathcal{I}$, $|\mu_i - \mu_i'| \leq n^{-3/2-1/b_2}$ for all $i \in \mathcal{I}$, and $|\sigma - \sigma'| \leq n^{-3/2-1/b_2}$, it follows from proposition 3.11

$$
\begin{aligned}
|f(x_j) - m(x_j)| &\leq 2\epsilon_n + \sum_{i \in \mathcal{I}} |u_i - u_i'| + \sum_{i \in \mathcal{I}} |u_i| |\varphi((x_j - \mu_i)/\sigma) - \varphi((x_j - \mu_i')/\sigma')| \\
&\leq 2\epsilon_n + \sum_{i \in \mathcal{I}} |u_i - u_i'| + 4 \sum_{i \in I} |u_i| \frac{|\sigma_i - \sigma_i'|}{\sigma_i \vee \sigma_i'} + \sum_{i \in \mathcal{I}} |u_i| \frac{|\mu_i - \mu_i'|}{\sigma_i \vee \sigma_i'} \leq 8\epsilon_n,
\end{aligned}
$$

for all $j = 1, \ldots, n$. Therefore $d_n(f, m) \leq 8\epsilon_n$, and the claim is proved with $\delta := 8$. To finish the proof, it suffices to compute the cardinality of $\mathcal{F}_{n,\epsilon}$. A straightforward computation shows that $|R_n| \leq n^{5/2+1/b_1+1/b_2} \sqrt{6 \log n} \leq n^{4+1/b_1+1/b_2}$ for all $n \geq 1$, then

$$
\begin{aligned}
\log N(c_3 \epsilon_n, \mathcal{F}_n, d_n) &\leq |\mathcal{I}| \log \left(\frac{n}{n^{-3/2}} \times n^{4+1/b_1+1/b_2} \right) + \log \left(\frac{n^{1/b_1}}{n^{-3/2-1/b_2}} \right) \\
&\leq H \left(\frac{11}{2} + \frac{2}{b_1} + \frac{2}{b_2} \right) n \epsilon_n^2,
\end{aligned}
$$

where the last line holds when n becomes large enough. Then the lemma is proved with $C := (11/2 + 2/b_1 + 2/b_2)/64$. ■

Lemma 3.5 – *Assume that there is $n_0 \in \mathbb{N}$, and $0 < \gamma_1 \leq \gamma_2 < 1$ such that $n^{-\gamma_2/2} \leq \epsilon_n \leq n^{-\gamma_1/2}$ for all $n \geq n_0$. Then $\Pi(\mathcal{F}_n(H, \epsilon_n)^c) \lesssim \exp(-\frac{H}{4}(1 - \gamma_2) n \epsilon_n^2)$ for all $n \geq n_0$.*

Proof. We use the fact that $M \sim \Pi_\alpha$ is almost surely purely-atomic (Kingman, 1992)). Then from the definition of $\mathcal{F}_n$ it follows

$$
\begin{aligned}
\Pi(\mathcal{F}_n^c) \leq G_\sigma(\sigma \leq n^{-1/b_2}) + G_\sigma(\sigma > n^{1/b_1}) + \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| > n \Big) \\
+ \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{|u_i| \leq n^{-1}\} > \epsilon_n \Big) + \\
\Pi_\alpha\Big(|\{i : |u_i| > n^{-1}\}| > H n \epsilon_n^2 / \log n \Big).
\end{aligned}
$$

We bound each of the term as follows. By assumption $G_\sigma(\sigma \leq n^{-1/b_2}) \lesssim e^{-a_2 n}$ and $G_\sigma(\sigma > n^{1/b_1}) \lesssim e^{-a_1 n}$. Notice that $\sum_{i=1}^\infty |u_i| = |M|$, where $|M|$ denote the total variation of the measure M. Since by definition we have $M \stackrel{d}{=} M_1 - M_2$, with M_1, M_2 independent Gamma random measures with same base measure $\alpha(\cdot)$, it follows that $|Q|$ has the distribution of a Gamma random variable with shape parameter $2\overline{\alpha}$. Then by Markov's inequality,

$$
\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| > n \Big) = \Pi_\alpha \left(e^{\frac{1}{2}|M|} > e^{\frac{1}{2}n} \right) \leq 2^{2\overline{\alpha}} e^{-\frac{1}{2}n}.
$$

Also, by the superposition theorem (Kingman, 1992, section 2), for any $M \sim \Pi_\alpha$ we have $M \stackrel{d}{=} M_3 + M_4$, where M_3 and M_4 are independent random measures with total

variation $|M_3|$ and $|M_4|$ having Laplace transforms (for all $t \in \mathbb{R}$ for which the integrals in the expressions converge)

$$Ee^{t|M_3|} := \exp\left\{2\overline{\alpha}\int_{1/n}^{\infty}(e^{tx}-1)x^{-1}e^{-x}\,dx\right\},$$
$$Ee^{t|M_4|} := \exp\left\{2\overline{\alpha}\int_{0}^{1/n}(e^{tx}-1)x^{-1}e^{-x}\,dx\right\}.$$

M_3 and M_4 are almost-surely purely atomic, M_3 has only jumps greater than $1/n$ (almost surely) which number is distributed according to a Poisson distribution with intensity $2\overline{\alpha}E_1(n^{-1})$, where E_1 denotes the exponential integral E_1 function: $E_1(x) = \int_x^\infty \frac{e^{-t}}{t}dt$. Likewise, M_4 has only jumps smaller or equal to $1/n$ (almost-surely) which number is almost-surely infinite. Recalling that $E_1(x) = \gamma + \log(1/x) + o(1)$ for x small, it holds $2\overline{\alpha}\gamma \leq 2\overline{\alpha}E_1(1/n) \leq 6\overline{\alpha}\log n \leq x_n$ for n sufficiently large, with $x_n := Hn\epsilon_n^2/\log n$. Thus using Chernoff's bound on Poisson distribution, we get

$$\begin{aligned}\Pi_\alpha\Big(|\{i\,:\,|u_i|>n^{-1}\}| > Hn\epsilon_n^2/\log n\Big) &\leq e^{-2\overline{\alpha}E_1(1/n)}\frac{(e2\overline{\alpha}E_1(1/n))^{x_n}}{x_n^{x_n}}\\ &\leq \exp\left\{-\frac{1}{2}x_n\log x_n\right\}.\end{aligned}$$

But, $\log x_n = \log n + \log H - 2\log\epsilon_n^{-1} - \log\log n \geq (1-\gamma_2)\log n + \log H - \log\log n \geq \frac{1}{2}(1-\gamma_2)\log n$ for large n. Therefore, as $n \to \infty$

$$\Pi_\alpha\Big(|\{i\,:\,|u_i|>n^{-1}\}| > Hn\epsilon_n^2/\log n\Big) \leq \exp\left\{-\frac{H}{4}(1-\gamma_2)n\epsilon_n^2\right\}.$$

Finally, we use again Markov's inequality to get

$$\begin{aligned}\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\,\mathbb{1}\{|u_i|\leq n^{-1}\} > \epsilon_n\Big) &= \Pi_\alpha\Big(e^{n\epsilon_n|M_4|} > e^{n\epsilon_n^2}\Big)\\ &\leq e^{-n\epsilon_n^2}\exp\left\{2\overline{\alpha}\int_0^{1/n}(e^{n\epsilon_n x}-1)x^{-1}e^{-x}\,dx\right\}.\end{aligned}$$

But for $x \in (0, 1/n)$, we have $e^{n\epsilon_n x} - 1 \leq n(e^{n\epsilon_n\delta_n}-1)x$, thus the integral in the previous expression is bounded by $2\overline{\alpha}(e^{\epsilon_n}-1)$, which is in turn bounded by $2\overline{\alpha}(e-1)$ because $\epsilon_n \leq 1$ if $n \geq n_0$. ■

3.4.2. Case of the location-scale mixture

Kullback-Leibler condition

By Chebychev inequality, we have $Q_0[-\zeta_j,\zeta_j]^c \le \zeta_j^{-p} Q_0|X|^p$. Therefore, bringing together results from propositions 3.7 and 3.8,

$$\begin{aligned}\int |f_M(x)-f_0(x)|^2\,dQ_0(x) &\\ &= \sum_{j=0}^{J}\int_{I_j}|f_M(x)-f_0(x)|^2\,dQ_0(x) + \int_{A_0^c}|f_M(x)-f(x)|^2\,dQ_0(x)\\ &\lesssim J^3\sum_{j=0}^{J}\sigma_j^{2\beta}Q_0(I_j) + J^3 Q_0(A_0^c).\end{aligned}$$

Then we can find a constant $C>0$ such that $\int |f_M(x)-f_0(x)|^2\,dQ_0(x) \le CJ^4\sigma_J^{2\beta}$ for all $M\in\mathcal{M}$ and J large enough.

By equation (3.7), we have $G_\sigma(U_j) \gtrsim \sigma_j^{-b_3}\sigma_J^{b_4\beta}\exp(-a_3/\sigma_j)$ for all $j=0,\dots J$. Moreover, there is a separation of $h_J\sigma_j$ between two consecutive μ_{jk} and $h_J\sigma_j \ll \sigma_j$, thus all the W_{jk} with $(j,k)\in\Lambda$ are disjoint. By equation (3.8), we have $\alpha_{jk} := \overline{\alpha}G_\sigma(U_j)G_\mu(V_{jk}) \gtrsim \sigma_j^{b_5(\beta+1)+b_4\beta}\exp(-a_3/\sigma_j)(1+|\mu_{jk}|)^{-b_6}$ for all $(j,k)\in\Lambda$. We also define $\alpha^c := \alpha(W^c)$. For J large enough, there is a constant $C'>0$ not depending on J such that $\alpha^c > C'$. Moreover, since α has finite variation we can assume without loss of generality that $C'\le\alpha^c\le 1$, otherwise we split W^c into disjoint parts, each of them having α-measure smaller than one. With $\epsilon_n^2 := CJ^4\sigma_J^{2\beta}$, using that $\Gamma(\alpha)\le 2\alpha^{\alpha-1}$ for $\alpha\le 1$ and $\mathcal{M}\subset \mathtt{KL}(f_0,\epsilon_n)$, it follows the lower bound

$$\begin{aligned}\Pi(\mathtt{KL}(f_0,\epsilon_n)) &\ge \frac{\sigma_J^\beta}{3e\Gamma(\alpha^c)}\prod_{(j,k)\in\Lambda}\left(\frac{\sigma_J^\beta e^{-2|u_{jk}|}}{3e\Gamma(\alpha_{jk})}\right)\\ &\ge \frac{\sigma_J^\beta}{3e\Gamma(\alpha^c)}\prod_{(j,k)\in\Lambda}\exp\left\{-2|u_{jk}| - \beta\log\sigma_J^{-1} + \log\frac{1}{6e} + (\alpha_{jk}-1)\log\alpha_{jk}\right\}\\ &\ge \exp\left\{-KJ|\Lambda| - 2\textstyle\sum_{(j,k)\in\Lambda}|u_{jk}| - \sum_{(j,k)\in\Lambda}\log\alpha_{jk}^{-1}\right\},\end{aligned} \tag{3.21}$$

for a constant $K>0$ depending only on C and β. We now evaluate the sums involved in the rhs of equation (3.21). As before, be have that $\sum_{(j,k)\in\Lambda}|u_{jk}| \le 4\|f_0\|_1\sigma_J^{-1}$ (see for instance the proof of proposition 3.8). Act as in section 3.4.1 to find that

$$\sum_{(j,k)\in\Lambda}\log\alpha_{jk}^{-1} \lesssim J|\Lambda| + J^{3/2}\sigma_J^{-1} + |\Lambda|\sigma_J^{-1}.$$

The term proportional to $|\Lambda|\sigma_J^{-1}$ is entirely responsible for the bad rates in location-scale mixtures, and the aim of the hybridation of next section is to get rid of it. For a constant $K'>0$,

$$\Pi(\mathtt{KL}(f_0,\epsilon_n)) \ge \exp\left\{-K'|\Lambda|\sigma_J^{-1}\right\}.$$

Then for an appropriate constant $C'>0$ we can have $\Pi(\mathtt{KL}(f_0,\epsilon_n)) \ge e^{-c_2 n\epsilon_n^2}$ if

$$\epsilon_n^2 = \begin{cases} C'[n^{-2\beta/(3\beta+2)}(\log n)^{t_1} \wedge n^{-2\beta/(2\beta+1+2\beta/p)}(\log n)^{t_2}], & p\le 2\beta,\\ C'n^{-\beta/(\beta+1)}(\log n)^{t_3}, & p>2\beta,\end{cases}$$

where $t_1 := 4 - 8\beta/(3\beta+2)$, $t_2 := 4 - 4\beta/(2\beta+1+2\beta/p)$ and $t_3 := 4 - 2\beta/(\beta+1)$.

Sieve construction

Using the notation $f_M(x) := \int \varphi((x-\mu)/\sigma)\, dM(\sigma,\mu)$, we construct the following sieve.

$$\mathcal{F}_n(H,\epsilon) := \left\{ f = f_M \begin{array}{l} M = \sum_{i=1}^\infty u_i \delta_{\sigma_i,\mu_i}, \quad \sum_{i=1}^\infty |u_i| \le n, \\ |\{i \,:\, |u_i| > n^{-1},\ n^{-1/b_2} < \sigma_i \le n^{1/b_1}\}| \le Hn\epsilon^2/\log n, \\ \sum_{i=1}^\infty |u_i|\, \mathbb{1}\{|u_i| \le n^{-1}\} \le \epsilon, \\ \sum_{i=1}^\infty |u_i|\, \mathbb{1}\{\sigma_i \le n^{-1/b_2}\} \le \epsilon,\ \sum_{i=1}^\infty |u_i|\, \mathbb{1}\{\sigma_i > n^{1/b_1}\} \le \epsilon \end{array} \right\}. \quad (3.22)$$

Lemma 3.6 – *Let $x = (x_1,\dots,x_n) \in \mathbb{R}^n$ be arbitrary and d_n be the empirical L^2-distance associated with x. Then for any $n^{-1/2} < \epsilon_n \le 1$, $0 < H \le 1$ and n sufficiently large there is a constant $C > 0$ not depending on n such that $\log N(\epsilon_n, \mathcal{F}_n(H,\epsilon_n), d_n) \le CHn\epsilon_n^2$.*

The proof is almost identical to lemma 3.4, with the same constant $C > 0$.

Lemma 3.7 – *Assume that there is $n_0 \in \mathbb{N}$, and $0 < \gamma_1 \le \gamma_2 < 1$ such that $n^{-\gamma_2/2} \le \epsilon_n \le n^{-\gamma_1/2}$ for all $n \ge n_0$. Then $\Pi(\mathcal{F}_n(H,\epsilon_n)^c) \lesssim \exp(-\frac{H}{4}(1-\gamma_2)n\epsilon_n^2)$ for all $n \ge n_0$.*

Proof. We first write the estimate

$$\begin{aligned} \Pi(\mathcal{F}_n^c) \le{}& \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| > n\Big) + \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\, \mathbb{1}\{|u_i| \le \delta\} > \epsilon_n\Big) \\ &+ \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\, \mathbb{1}\{\sigma_i \le n^{-1/b_2}\} > \epsilon_n\Big) + \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\, \mathbb{1}\{\sigma_i > n^{1/b_1}\} > \epsilon_n\Big) \\ &+ \Pi_\alpha\Big(|\{i \,:\, |u_i| > \delta,\ n^{-1/b_2} < \sigma_i \le n^{1/b_1}\}| > Hn\epsilon_n^2/\log n\Big). \end{aligned}$$

The first three terms in the rhs above obeys the same bounds as in the proof of lemma 3.5, using the same arguments. The last two term are bounded using the same trick, thus we simply bound the last term and left the other to the reader. Notice that the random variable $U := \sum_{i=1}^\infty |u_i|\, \mathbb{1}\{\sigma_i > n^{1/b_1}\}$ has Gamma distribution with parameters $2\alpha(A_n), 1$, with $A_n := \{(\sigma,\mu) \,:\, \sigma > n^{1/b_1}\}$. For n large, by assumptions on P_σ, it holds $\alpha(A_n) \ll \epsilon_n$. Then by Chebychev inequality, for n large enough

$$\begin{aligned} \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\, \mathbb{1}\{\sigma_i > n^{1/b_1}\} > \epsilon_n\Big) &\le \mathbb{P}(U - EU > \epsilon_n - EU) \\ &\le \mathbb{P}(U - EU > \epsilon_n/2) \le 16\epsilon_n^{-2}\alpha(A_n)^2. \end{aligned} \quad (3.23)$$

The conclusion follows from the assumptions on G_σ which imply $\alpha(A_n) = \overline{\alpha} G_\sigma(\sigma > n^{1/b_1}) \lesssim \exp(-a_1 n)$. ∎

3.4.3. Hybrid location-scale mixtures

Obviously, given the definition of hybrid mixtures (see section 3.4.3), most of the proof is redundant with the location-scale case, and in the sequel we deal only with the parts that differ.

Kullback-leibler condition

Let $\mathcal{M} \equiv \mathcal{M}(\beta, J, f, \Lambda)$ be the set of signed measures constructed in section 3.3.3. For any integer $J > 0$ let Ω_J be the event

$$\Omega_J := \Big\{ P_\sigma \,:\, P_\sigma[2^{-j}, 2^{-j}(1 + 2^{-J\beta})] \geq 2^{-J} \quad \forall\, 0 \leq j \leq J \Big\}.$$

Then with arguments and constant $C > 0$ from section 3.4.2, letting $\epsilon_n^2 := CJ^4\sigma_J^{2\beta}$, we have

$$\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq \Pi(\mathcal{M}) \geq \Pi(\mathcal{M} \mid \Omega_J)\Pi_\sigma(\Omega_J).$$

But by equation (3.11) we have $\Pi_\sigma(\Omega_J) \gtrsim \exp(-a_6 J^{b_7} 2^J)$ and on Ω_J it holds $\alpha(W_{jk}) = \overline{\alpha} P_\sigma(U_j) G_\mu(V_{jk}) \geq \overline{\alpha} 2^{-J} G_\mu(V_{jk})$ for all $(j,k) \in \Lambda$. Then act as in equation (3.21) to find a constant $K > 0$ such that (recalling that $\sigma_J = 2^{-J}$)

$$\Pi(\mathtt{KL}(f_0, \epsilon_n)) \gtrsim \exp\Big\{ -K(J^{b_7} \vee J^{1/2})\sigma_J^{-1} - KJ|\Lambda| \Big\}.$$

Because of proposition 3.6 we can have $\Pi(\mathtt{KL}(f_0, \epsilon_n)) \geq e^{-c_2 n \epsilon_n^2}$ if for an appropriate constant $C' > 0$

$$\epsilon_n^2 = \begin{cases} C'[n^{-2\beta/(3\beta+1)}(\log n)^{4-6\beta/(3\beta+1)} \wedge n^{-p/(p+1)}(\log n)^{4-p/(p+1)}] & p \leq 2\beta, \\ C' n^{-2\beta/(2\beta+1)}(\log n)^{4-2\beta(4-b_7 \vee 3)/(2\beta+1)} & p > 2\beta. \end{cases}$$

Sieve construction

We use the same sieve $\mathcal{F}_n(H, \epsilon)$ as in equation (3.22). The definition of $\mathcal{F}_n(j, \epsilon)$ is independent of Π thus the conclusion of lemma 3.4 holds for hybrid location-scale mixtures. It remains to show that $\Pi(\mathcal{F}_n(H, \epsilon)^c) \leq \exp(-2c_2 n \epsilon_n^2)$, which is the object of the next lemma.

Lemma 3.8 – *Assume that there is $n_0 \in \mathbb{N}$, and $0 < \gamma_1 \leq \gamma_2 < 1$ such that $n^{-\gamma_2/2} \leq \epsilon_n \leq n^{-\gamma_1/2}$ for all $n \geq n_0$. Then there is a constant a constant $\gamma_2 < \gamma < 1$ such that $\Pi(\mathcal{F}_n(H, \epsilon_n)^c) \lesssim \exp(-\frac{H}{4}(1-\gamma)n\epsilon_n^2)$ for all $n \geq n_0$.*

Proof. We proceed as in the proof of lemma 3.7. Following the same steps, we deduce that it is sufficient to prove that

$$\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{\sigma_i > n^{1/b_1}\} > \epsilon_n \Big) \lesssim e^{-2c_2 n},$$
$$\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{\sigma_i \leq n^{-1/b_2}\} > \epsilon_n \Big) \lesssim e^{-2c_2 n}.$$

Since the proofs are almost identical for the two previous conditions, we only prove the first and left the second to the reader. Notice that by equation (3.23) we have

$$\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i| \, \mathbb{1}\{\sigma_i > n^{1/b_1}\} > \epsilon_n \;\Big|\; P_\sigma \Big) \leq 16\overline{\alpha}\epsilon_n^{-2} P_\sigma(\sigma > n^{1/b_1})^2.$$

Letting $\Omega := \{P_\sigma \,:\, P_\sigma(\sigma > n^{1/b_1}) < \exp(-a_1 n/2)\}$, with a slight abuse of notation, it follows from equation (3.9)

$$\begin{aligned}\Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\,\mathbb{1}\{\sigma_i > n^{1/b_1}\} > \epsilon_n\Big) &\\ \leq \Pi_\alpha\Big(\textstyle\sum_{i=1}^\infty |u_i|\,\mathbb{1}\{\sigma_i > n^{1/b_1}\} > \epsilon_n \;\Big|\; \Omega\Big) + \Pi_\sigma(\Omega^c) &\\ \lesssim \epsilon_n^{-2}\exp(-a_1 n) + \exp(-a_4 n). &\end{aligned}$$

■

3.5. Proof of theorem 3.2

The proof follows the same lines as Ghosal et al. (2007b) with additional cares. The first step consists on rewriting expectation of the posterior distribution as follows. Let $(\phi_n(\cdot \mid \cdot))_{n\geq 0}$ be a sequence of test functions such that for n large enough

$$\begin{aligned} Q_0^n\left[P_0^n[\phi_n(\mathbf{y} \mid \mathbf{x}) \mid \mathbf{x}]\right] &\lesssim N(\epsilon/18, \mathcal{F}_n, d_n)\exp\left(-\frac{n\epsilon_n^2}{2}\right), \\ \sup_{\{f: d_n(f,f_0)\geq 17\epsilon_n/18\}\cap\mathcal{F}_n} Q_0^n\left[P_f^n[1-\phi_n(\mathbf{y} \mid \mathbf{x})] \mid \mathbf{x}\right] &\lesssim \exp\left(-\frac{n\epsilon_n^2}{2}\right). \end{aligned}$$

The existence of such test functions is standard and follows for instance from Birgé (2006, proposition 4), or Ghosal and van der Vaart (2007, section 7.7). From here, we bound the posterior distribution in a standard fashion,

$$\begin{aligned} Q_0^n\left[P_0^n[\Pi_{\mathbf{x}}(\{f \,:\, d_n(f,f_0) > \epsilon_n\} \mid \mathbf{y},\mathbf{x}) \mid \mathbf{x}]\right] &\leq Q_0^n[P_0^n[\Pi_{\mathbf{x}}(\mathcal{F}_n^c \mid \mathbf{y},\mathbf{x}) \mid \mathbf{x}]] \\ &\quad + Q_0^n[P_0^n[\Pi(\{f \,:\, d_n(f,f_0) > \epsilon_n\}\cap\mathcal{F}_n \mid \mathbf{y},\mathbf{x}) \mid \mathbf{x}]]. \end{aligned}$$

So that,

$$\begin{aligned} Q_0^n\left[P_0^n[\Pi_{\mathbf{x}}(\{f \,:\, d_n(f,f_0) > \epsilon_n\} \mid \mathbf{y},\mathbf{x}) \mid \mathbf{x}]\right] &\leq Q_0^n[P_0^n[\Pi_{\mathbf{x}}(\mathcal{F}_n^c \mid \mathbf{y},\mathbf{x}) \mid \mathbf{x}]] \\ &\quad + Q_0^n\left[P_0^n\left[\phi_n(\mathbf{y} \mid \mathbf{x})\Pi_{\mathbf{x}}(\{f \,:\, d_n(f,f_0) > \epsilon_n\}\cap\mathcal{F}_n \mid \mathbf{y},\mathbf{x}) \mid \mathbf{x}\right]\right] \\ &\quad + Q_0^n\left[P_0^n\left[(1-\phi_n(\mathbf{y} \mid \mathbf{x}))\Pi_{\mathbf{x}}(\{f \,:\, d_n(f,f_0) > \epsilon_n\}\cap\mathcal{F}_n \mid \mathbf{y},\mathbf{x}) \mid \mathbf{x}\right]\right]. \end{aligned} \tag{3.24}$$

Now, to any $\mathbf{x} \in \mathbb{R}^n$, we associate the event

$$E_n(\mathbf{x}) := \left\{ y \in \mathbb{R}^n \,:\, \int_{\mathcal{F}} \prod_{i=1}^n \frac{p_f(x_i,y_i)}{p_{f_0}(x_i,y_i)}\, d\Pi_{\mathbf{x}}(f) \geq \exp\left(-(1+4c_2)\frac{n\epsilon_n^2}{4}\right)\right\}. \tag{3.25}$$

Consider the first term of the rhs of equation (3.24). We can rewrite,

$$
\begin{aligned}
&Q_0^n[P_0^n[\Pi_{\mathbf{x}}(\mathcal{F}_n^c \mid \mathbf{y}, \mathbf{x}) \mid \mathbf{x}]] \\
&\qquad \leq e^{\frac{1}{4}(4c_2+1)n\epsilon_n^2} \int_{\mathbb{R}^n} \int_{E_n(\mathbf{x})} \int_{\mathcal{F}_n^c} \prod_{i=1}^{n} \frac{p_f(x_i, y_i)}{p_{f_0}(x_i, y_i)} \, d\Pi_{\mathbf{x}}(f) dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) \\
&\qquad\quad + \int_{\mathbb{R}^n} \int_{E_n(\mathbf{x})^c} dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) \\
&\qquad = e^{\frac{1}{4}(4c_2+1)n\epsilon_n^2} \int_{\mathbb{R}^n} \int_{\mathcal{F}_n^c} \int_{E_n(\mathbf{x})} dP^n(\mathbf{y} \mid \mathbf{x}) d\Pi_{\mathbf{x}}(f) dQ_0^n(\mathbf{x}) \\
&\qquad\quad + \int_{\mathbb{R}^n} \int_{E_n(\mathbf{x})^c} dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) \\
&\qquad \leq e^{\frac{1}{4}(4c_2+1)n\epsilon_n^2} \int_{\mathbb{R}^n} \Pi_{\mathbf{x}}(\mathcal{F}_n^c) \, dQ_0^n(\mathbf{x}) + \int_{\mathbb{R}^n} \int_{E_n(\mathbf{x})^c} dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}),
\end{aligned}
$$

where the third line follows from Fubini's theorem. The same reasoning applies to the other terms of equation (3.24), using the test functions introduced above and $0 < c_2 < 1/4$. Hence the theorem is proved if we show that $\int_{\mathbb{R}^n} \int_{E_n(\mathbf{x})^c} dP_0^n(\mathbf{y} \mid \mathbf{x}) dQ_0^n(\mathbf{x}) = o(1)$. But under the condition of the theorem, Ghosal et al. (2007b, Lemma 10) implies that

$$
P_0^n \left(\int_{\mathcal{F}} \prod_{i=1}^{n} \frac{p_f(x_i, Y_i)}{p_{f_0}(x_i, Y_i)} \, d\Pi_{\mathbf{x}}(f) < \exp\left(-\frac{1}{4}(1 + 4c_2)\epsilon_n^2 \right) \;\middle|\; \mathbf{x} \right) = o(1).
$$

Appendix

3.A. Proofs of lemmas 3.2 and 3.3 and some technical results on the kernels

3.A.1. Proof of lemma 3.2

Clearly, $\|\chi_\sigma * f\|_1 \leq \|\chi_\sigma\|_1 \|f\|_1$ by Young's inequality, so that $\chi_\sigma * f \in L^1$ and $(\chi_\sigma * f)^\wedge(\xi) = \widehat{\chi}_\sigma(\xi)\widehat{f}(\xi)$, showing that the support of the Fourier transform of $\chi_\sigma * f$ is included in $[-1/\sigma, 1/\sigma]$. Moreover, using again Young's inequality we get that $\|\chi_\sigma * f\|_\infty \leq \|\chi_\sigma\|_1 \|f\|_\infty$, thus $\chi_\sigma * f \in L^\infty$.

Because $\widehat{\chi}$ is $\mathtt{C}^\infty$ and compactly supported, for any integer $q \geq 0$ we have $(iu)^q \chi(u) = (2\pi)^{-1} \int \widehat{\chi}^{(q)}(\xi) e^{i\xi u} d\xi$. Clearly $\widehat{\chi}$ is Schwartz, hence by Fourier inversion we have that

$$\int u^q \chi(u) e^{-i\xi u} du = (-i)^q \widehat{\chi}^{(q)}(\xi), \quad \forall \xi \in \mathbb{R}.$$

But, by construction $\widehat{\chi}(0) = 1$, and for any $q \geq 1$ we have $\widehat{\chi}^{(q)}(0) = 0$. It follows that $\int \chi(u) du = 1$, and $\int u^q \chi(u) du = 0$ for any $q \geq 1$. Whence, letting m be the largest integer smaller than β, and using Taylor's formula with exact remainder term

$$\begin{aligned}
\chi_\sigma * f(x) - f(x) &= \int \chi_\sigma(y) \left[f(x-y) - f(x)\right] dy = \int \chi(y) \left[f(x - \sigma y) - f(x)\right] dy \\
&= \sum_{k=1}^{m} \frac{(-1)^k \sigma^k}{k!} \int u^k \chi(u)\, du \\
&\quad + \int \chi(y) \int_0^1 (-\sigma y)^m \frac{(1-u)^{m-1}}{(m-1)!} \left[f^{(m)}(x - u\sigma y) - f^{(m)}(x)\right] du dy \\
&= \int \chi(y) \int_0^1 (-\sigma y)^m \frac{(1-u)^{m-1}}{(m-1)!} \left[f^{(m)}(x - u\sigma y) - f^{(m)}(x)\right] du dy.
\end{aligned}$$

Therefore, because $f \in \mathtt{C}^\beta$,

$$\begin{aligned}
|\chi_\sigma * f(x) - f(x)| &\leq \sigma^m \int |y^m \chi(y)| \int_0^1 \frac{(1-u)^{m-1}}{(m-1)!} |f^{(m)}(x - u\sigma y) - f^{(m)}(x)|\, du dy \\
&\leq \|f\|_{\mathtt{C}^\beta} \sigma^\beta \int |y^\beta \chi(y)|\, dy \int_0^1 \frac{(1-u)^{m-1}}{(m-1)!} u^{\beta - m}\, du.
\end{aligned}$$

3.A.2. Proof of lemma 3.3

We mostly follow the proof of Hangelbroek and Ron (2010, proposition 1). Writing,

$$\begin{aligned} K_{h,\sigma} f_\sigma(x) &= \int \frac{h}{\sigma} \sum_{k\in\mathbb{Z}} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) \eta\left(\frac{y-h\sigma k}{\sigma}\right) f_\sigma(y)\, dy \\ &= \frac{h}{\sigma} \sum_{k\in\mathbb{Z}} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) \int \eta\left(\frac{y-h\sigma k}{\sigma}\right) f_\sigma(y)\, dy \\ &= \frac{h}{2\pi} \sum_{k\in\mathbb{Z}} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) \int \widehat{\eta}(\sigma\xi)\widehat{f}_\sigma(\xi) e^{i\xi h\sigma k}\, d\xi \\ &= \int \widehat{\eta}(\sigma\xi)\widehat{f}_\sigma(\xi) \frac{h}{2\pi} \sum_{k\in\mathbb{Z}} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) e^{i\xi h\sigma k}\, d\xi. \end{aligned}$$

Then we can invoke the *Poisson summation formula* (Härdle et al., 1998, theorem 4.1), which is obviously valid for φ, and

$$\sum_{k\in\mathbb{Z}} \varphi\left(\frac{x-h\sigma k}{\sigma}\right) e^{i\xi h\sigma k} = \frac{1}{h} \sum_{m\in\mathbb{Z}} \widehat{\varphi}\left(\sigma\xi + \frac{2\pi m}{h}\right) e^{i(\sigma\xi + \frac{2\pi m}{h})x/\sigma}.$$

Therefore, recalling that $\widehat{f}_\sigma$ is supported on $[-1/\sigma, 1/\sigma]$ and $\widehat{\chi}$ equals 1 on $[-1,1]$,

$$\begin{aligned} K_{h,\sigma} f_\sigma(x) &= \frac{1}{2\pi} \int \widehat{\chi}(\sigma\xi)\widehat{f}_\sigma(\xi) \sum_{m\in\mathbb{Z}} \frac{\widehat{\varphi}(\sigma\xi + 2\pi m/h)}{\widehat{\varphi}(\sigma\xi)} e^{i(\sigma\xi + \frac{2\pi m}{h})x/\sigma}\, d\xi \\ &= f_\sigma(x) + \frac{1}{2\pi} \sum_{m\in\mathbb{Z}\setminus\{0\}} \int \widehat{f}_\sigma(\xi) \frac{\widehat{\varphi}(\sigma\xi + 2\pi m/h)}{\widehat{\varphi}(\sigma\xi)} e^{i(\sigma\xi + \frac{2\pi m}{h})x/\sigma}\, d\xi. \end{aligned}$$

It follows that,

$$|K_{h,\sigma} f_\sigma(x) - f_\sigma(x)| \leq \frac{1}{2\pi} \|\widehat{f}_\sigma\|_1 \sup_{\xi\in[-1,1]} \sum_{m\in\mathbb{Z}\setminus\{0\}} \left| \frac{\widehat{\varphi}(\xi + 2\pi m/h)}{\widehat{\varphi}(\xi)} \right|.$$

Now, $\|\widehat{f}_\sigma\|_1 \leq 2\sigma^{-1}\|\widehat{f}_\sigma\|_\infty \leq 2\sigma^{-1}\|f_\sigma\|_1 \leq 2\sigma^{-1}\|f\|_1$, which is finite because of lemma 3.2. Recalling that by assumption $\widehat{\varphi}$ is Gaussian, it follows for all $\xi \in [-1,1]$ and all $h \leq 1$,

$$\begin{aligned} \sum_{m\in\mathbb{Z}\setminus\{0\}} \left| \frac{\widehat{\varphi}(\xi + 2\pi m/h)}{\widehat{\varphi}(\xi)} \right| &\leq \exp\left\{ -\frac{1}{2}(\xi + 2\pi m/h)^2 + \frac{1}{2}\xi^2 \right\} \\ &\leq e^{-1/2} \sum_{m\in\mathbb{Z}\setminus\{0\}} e^{-4\pi^2 m^2/h^2} \leq 4e^{-1/2} e^{-4\pi^2/h^2}. \end{aligned}$$

Then the lemma is proved with $C := 8e^{-1/2}$.

3.A.3. Some other technical results on $K_{h,\sigma}$

Lemma 3.9 – *There is a universal constant $C > 0$ such that for all $x \in \mathbb{R}$, all $0 < h \leq 1$ and all $\sigma > 0$, $\sum_{k\in\mathbb{Z}} |\eta((x-h\sigma k)/\sigma)| \leq Ch^{-1}$. Moreover, $\eta \in \mathcal{S}$.*

Proof. We first prove that $\eta \in \mathcal{S}$. Obviously $\widehat{\varphi} \in \mathcal{S}$, and therefore so is $\widehat{\eta}$. Since the Fourier transform and the inverse Fourier transform are continuous mapping of $\mathcal{S}$ onto itself, it is immediate that $\eta \in \mathcal{S}$.

We finish the proof by remarking that $x \mapsto \sum_{k\in\mathbb{Z}} |\eta((x - h\sigma k)/\sigma)|$ is periodic with period $h\sigma$, hence it suffices to check that it is bounded for $x \in [0, h\sigma]$. If $x \in [0, h\sigma]$, then $|x - h\sigma k| \geq |h\sigma k|/2$ for any $|k| \geq 2$, so that

$$\begin{aligned}\sum_{k\in\mathbb{Z}} |\eta((x - h\sigma k)/\sigma)| &\leq 3 \sup_{u\in\mathbb{R}} |\eta(u)| + \sum_{|k|\geq 2} |\eta((x - h\sigma k)/\sigma)| \\ &\leq 3\|\eta\|_{0,0} + \|\eta\|_{2,0} \sum_{|k|\geq 2} (1 + |hk|/2)^{-2} \\ &\leq 3\|\eta\|_{0,0} + 4\|\eta\|_{2,0}/h,\end{aligned}$$

which concludes the proof of the first assertion with $C := 3\|\eta\|_{0,0} + 4\|\eta\|_{2,0}$, because of the assumption $h \leq 1$. ■

The following Lemma gives some control on the coefficients of f on η.

Proposition 3.9 – *Let $0 < h \leq 1$ and $a_k(f) := (h/\sigma) \int \eta((y - h\sigma k)/\sigma) f(y)\, dyx$. Then there are universal constants $C, C' > 0$, depending only on φ, such that $\sum_{k\in\mathbb{Z}} |a_k(f)| \leq C\|f\|_1\sigma^{-1}$, and for all $k \in \mathbb{Z}$, $|a_k(f)| \leq C'\|f\|_\infty$.*

Proof. For the first assertion of the proposition, we write,

$$\begin{aligned}\sum_{k\in\mathbb{Z}} |a_k(f)| &\leq \frac{h}{\sigma} \sum_{k\in\mathbb{Z}} \int |f(y)||\eta((y - h\sigma k)/\sigma)|\, dy \\ &\leq \sigma^{-1}\|f\|_1 \sup_{y\in\mathbb{R}} h \sum_{k\in\mathbb{Z}} |\eta((y - h\sigma k)/\sigma)|,\end{aligned}$$

and the conclusion follows from lemma 3.9. The proof of the second assertion is simpler. Indeed,

$$|a_k(f)| \leq \frac{h}{\sigma} \int |f(y)||\eta((y - h\sigma k)/\sigma)|\, dy \leq h\|f\|_\infty \int |\eta(u)|\, du,$$

where the last integral is bounded because $\eta \in \mathcal{S}$ by lemma 3.9. ■

3.B. Proof of lemma 3.1

Let $x \in \mathbb{R}^n$ arbitrary, $\sigma > 0$ and $h_\sigma\sqrt{\log \sigma^{-1}} := 2\pi\sqrt{\beta + 1}$. Recall that from lemmas 3.2 and 3.3 we have $\|K_{h_\sigma,\sigma}(\chi_\sigma * f_0) - f_0\|_\infty \lesssim \sigma^\beta$, where $K_{h_\sigma,\sigma}(\chi_\sigma * f_0)(z) := \sum_{k\in\mathbb{Z}} u_k\, \varphi((z - h_\sigma\sigma k)/\sigma)$. Define $S_n(x) := \cup_{i=1}^n \{z \in \mathbb{R} : |z - x_i| \leq \sigma\sqrt{2(\beta + 1)\log \sigma^{-1}}\}$ and

$$\Lambda(x) := \{k \in \mathbb{Z} : |u_k| > \sigma^\beta, \quad h_\sigma\sigma k \in S_n(x)\}.$$

Also define $U_\sigma := \{\sigma' : \sigma \leq \sigma' \leq \sigma(1 + \sigma^\beta)\}$, and for all $k \in \Lambda(x)$ define $V_k := \{\mu : |\mu - h_\sigma\sigma k| \leq \sigma^{\beta+1}\}$. We denote by $\mathcal{M}_\sigma$ the set of signed measures M on $\mathbb{R}$ such that $|M(V_k) - u_k| \leq \sigma^\beta$ for all $k \in \Lambda(x)$ and $|M|(V^c) \leq \sigma^\beta$, where V^c is the relative complement of the union of all V_k for $k \in \Lambda(x)$. For any $M \in \mathcal{M}_\sigma$, we write $f_{M,\sigma}(z) :=$

$\int \varphi((z-\mu)/\sigma)\, dM(\mu)$. Act as in proposition 3.5 to find that $d_n(f, f_0) \leq C h_\sigma^{-2} \sigma^\beta$ for any $M \in \mathcal{M}_\sigma$, with a constant $C > 0$ not depending on x. By construction of $S_n(x)$, for all $k \in \Lambda(x)$ there is at least one x_i such that $|h_\sigma \sigma k - x_i| \leq \sigma\sqrt{2(\beta+1)\log \sigma^{-1}}$. Then for any $k \in \Lambda(x)$, by definition of G_x

$$\overline{\alpha} G_x(V_k) \geq n^{-1} \int_{h_\sigma \sigma k - \sigma^{\beta+1}}^{h_\sigma \sigma k + \sigma^{\beta+1}} g(z - x_i)\, dz \geq a_{21} n^{-1} \sigma^{a_{22}(\beta+1)}.$$

Remarking that $|\Lambda(x)| \lesssim \sigma^{-(\beta+1)}$ independently of x (see proposition 3.3) and letting $\epsilon_n = C' h_\sigma^{-2} \sigma^\beta$ we can mimic the steps of section 3.4.1 to find that

$$\begin{aligned} \Pi_x(f \,:\, d_n(f, f_0) \leq s\epsilon) &\gtrsim \exp\left\{-C''|\Lambda(x)| \log \sigma^{-1} - C''|\Lambda(x)| \log n\right\} \\ &\gtrsim \exp(-c_2 n \epsilon_n^2), \end{aligned}$$

for a constant $C'' > 0$ not depending on x and ϵ_n^2 defined in the lemma.

3.C. Some technical results on the construction of the approximation in the case of location-scale mixtures

Proposition 3.10 – *Let $f_0 \in \mathcal{C}_\beta$. For any $j \geq 0$, we have $|\Delta_j(x)| \leq C \|f_0\|_{\mathcal{C}^\beta} \sigma_j^\beta$, with the same constant $C > 0$ as in lemma 3.2. Moreover, $\|\Delta_j\|_1 \leq 2\|f_0\|_1$ for all $j \geq 0$.*

Proof. Notice that $\|\Delta_{j+1}\|_1 \leq \|\Delta_j\|_1 + \|\chi_{\sigma_{j+1}} * \Delta_j\| \leq (1 + \|\chi\|_1)\|\Delta_j\|_1$, by Young's inequality. Since $f_0 \in L^1$, this implies $\Delta_j \in L^1$ for all $j \geq 0$. Since $\widehat{\Delta}_{j+1}(\xi) = \widehat{\Delta}_j(\xi) - \widehat{\chi}_{\sigma_{j+1}}(\xi)\widehat{\Delta}_j(\xi)$, we get $\widehat{\Delta}_j(\xi) = \widehat{f_0}(\xi) \prod_{l=1}^j (1 - \chi_{\sigma_l}(\xi))$, by induction. Because $\sigma_{j+1} = \sigma_j/2$, and by construction of χ_{σ_l} we have $\widehat{\chi}_{\sigma_m}(\xi)\widehat{\chi}_{\sigma_l}(\xi) = \widehat{\chi}_{\sigma_m}(\xi)$ for any $m > l$, hence the last equation can be rewritten as $\widehat{\Delta}_j(\xi) = \widehat{f_0}(\xi)(1 - \widehat{\chi}_{\sigma_j}(\xi))$. Then we deduce that $\Delta_j = f_0 - \chi_{\sigma_j} * f_0$. By lemma 3.2, this implies that $|\Delta_j(x)| \leq C\|f_0\|_{\mathcal{C}^\beta}\sigma_j^\beta$. From the same estimate, it is clear that $\|\Delta_j\| \leq \|f_0|_1 + \|\chi_{\sigma_j} * f_0\| \leq 2\|f_0\|_1$. ■

3.C.1. Proof of proposition 3.8

Let define $A(\beta, J) := (2\log|\Lambda| + 2\beta \log \sigma_J^{-1})^{1/2}$ and $\mathcal{J} \equiv \mathcal{J}(x) := \{(j,k) \in \{0, \ldots, J\} \times \mathbb{Z} : |x - \mu_{jk}| \leq 4A(\beta, J)\sigma_j\}$. For any $M \in \mathcal{M}$ we can write

$$\begin{aligned} f_M(x) - f_0(x) = &\sum_{(j,k)\in\Lambda\cap\mathcal{J}} \int_{W_{jk}} \left[\varphi\left(\frac{x-\mu}{\sigma}\right) - \varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right)\right] dM(\sigma,\mu) \\ &+ \sum_{(j,k)\in\Lambda\cap\mathcal{J}} [M(W_{jk}) - u_{jk}]\, \varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) + \sum_{(j,k)\in\Lambda\cap\mathcal{J}^c} \int_{W_{jk}} \varphi\left(\frac{x-\mu}{\sigma}\right) dM(\sigma,\mu) \\ &- \sum_{(j,k)\in\Lambda\cap\mathcal{J}^c} u_{jk}\, \varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) - \sum_{(j,k)\notin\Lambda} u_{jk}\, \varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) \\ &+ \int_{W^c} \varphi\left(\frac{x-\mu}{\sigma}\right) dM(\sigma,\mu) \\ &:= r_1(x) + r_2(x) + r_3(x) + r_4(x) + r_5(x) + r_6(x). \end{aligned}$$

The proof follows similar steps as the proof of proposition 3.5. From the definition of $A(\beta, J)$ and proposition 3.6, we deduce that $A(\beta, J) \lesssim \sqrt{J}$ for J large enough. Also, there is a separation of $h_J\sigma_j$ between two consecutive μ_{jk}. Then there are no more than $2A(\beta, J)\sigma_j/(h_J\sigma_j) = 2A(\beta, J)h_J^{-1}$ distinct values of μ_{jk} in an interval of length $2A(\beta, J)\sigma_j$. Thus the bound $|\Lambda \cap \mathcal{J}| \leq 2(J+1)A(\beta, J) \lesssim J^{3/2}$ holds. It follows from proposition 3.11 that $|r_1(x)| \lesssim |\Lambda \cap \mathcal{J}|\sigma_J^\beta \lesssim J^{3/2}\sigma_J^\beta$. Obviously, $|r_2(x)| \leq \|\varphi\|_\infty|\Lambda \cap \mathcal{J}|\sigma_J^\beta \lesssim J^{3/2}\sigma_J^\beta$. Whenever $(j,k) \in \Lambda \cap \mathcal{J}^c$ and $(\sigma, \mu) \in W_{jk}$, choosing J large enough so that $1/2 \leq \sigma_j/\sigma \leq 2$ and $|\mu - \mu_{jk}| \leq \sigma_j A(\beta, J)/2$, it holds $|x - \mu| \geq A(\beta, J)\sigma$. Therefore, $|r_3(x)| \lesssim \varphi(A(\beta, J))|\Lambda| \leq \sigma_J^\beta$. With the same reasoning we get $|r_4(x)| \lesssim \|f\|_\infty\sigma_J^\beta$. Regarding r_6, we have the obvious bound $|r_6(x)| \leq \|\varphi\|_\infty\sigma_J^\beta$. The r_5 term is more subtle and constitutes the remainder of the proof.

Let $\Lambda_1^c := \{(j,k) \in \{0, \dots, J\} \times \mathbb{Z} : |u_{jk}| \leq \sigma_J^\beta\}$ and $\mathcal{K}_j := \{k \in \mathbb{Z} : |\mu_{jk}| > \zeta_j + \sqrt{2(\beta+1)\log\sigma_J^{-1}}\}$. Assuming that $x \in [-\zeta_q, \zeta_q]$ for some $0 \leq q \leq J$, we can bound $r_5(x)$ as follows,

$$|r_5(x)| \leq \sum_{(j,k)\in\Lambda_1^c} |u_{jk}|\,\varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) + \sum_{j\leq q}\sum_{k\in\mathcal{K}_j} u_{jk}\,\varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) + \sum_{j>q}\sum_{k\in\mathcal{K}_j} u_{jk}\,\varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right), \quad (3.26)$$

where the third term of the rhs does not exist if $q = J$. The first term of the rhs of equation (3.26) is bounded by $\sigma_J^\beta \sup_{x\in\mathbb{R}} \sum_{j=0}^J \sum_{k\in\mathbb{Z}} \varphi((x-\mu_{jk})/\sigma)$, which is in turn bounded by a constant multiple of $J^{3/2}\sigma_J^\beta$ (see for instance the proof of lemma 3.9). Because of propositions 3.9 and 3.10, when $x \in [-\zeta_q, \zeta_q]$ we always have

$$\sum_{j\leq q}\sum_{k\in\mathcal{K}_j} |u_{jk}|\,\varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) \leq \sup_{\substack{j\leq q\\ k\in\mathcal{K}_j}} \varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) \sum_{j\leq J}\sum_{k\in\mathbb{Z}} |u_{jk}| \leq \sigma_J^{\beta+1} \sum_{j\leq J} 2\|f_0\|_1\sigma_j^{-1} \leq 4\|f_0\|_1\sigma_J^\beta.$$

Regarding the second term of the rhs of equation (3.26), we introduce the sets of indexes $\mathcal{L}_j \equiv \mathcal{L}_j(x) := \{k \in \mathcal{K}_j : |x - \mu_{jk}| \leq \sigma_j\sqrt{2(\beta+1)\log\sigma_J^{-1}}\}$. Then, we can split again the sum as

$$\sum_{j>q}\sum_{k\in\mathcal{K}_j} u_{jk}\,\varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) = \sum_{j>q}\sum_{k\notin\mathcal{L}_j} u_{jk}\,\varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right) + \sum_{j>q}\sum_{k\in\mathcal{L}_j} u_{jk}\,\varphi\left(\frac{x-\mu_{jk}}{\sigma_j}\right).$$

With exactly the same reasoning as before, we get that the first sum of the rhs of the last expression is bounded above by $4\|f\|_1\sigma_J^\beta$. Concerning the second term, for any $j \geq 1$ we get from propositions 3.9 and 3.10, together with the definition of u_{jk}, that $|u_{jk}| \lesssim \|f\|_{\mathcal{C}^\beta}\sigma_j^\beta$. Since there is $h_J\sigma_j$ separation between two consecutive μ_{jk}, we deduce

that $|\mathcal{L}_j| \leq 2h_J^{-1}\sqrt{2(\beta+1)\log\sigma_J^{-1}}$. Therefore, for J large enough and $x \in [-\zeta_q, \zeta_q]$ with $0 \leq q \leq J$,

$$|r_5(x)| \lesssim \|f_0\|_1 \sigma_J^\beta + \|f_0\|_{\mathcal{C}^\beta}\sqrt{2(\beta+1)\log\sigma_J^{-1}}\sum_{j>q}\sigma_j^\beta \lesssim \sqrt{J}\sigma_q^\beta.$$

The conclusion of the proposition follows by combining all the preceding points.

3.D. Elementary results

Proposition 3.11 – *Let $\varphi(x) = \exp(-x^2/2)$. Then, for all $\mu_1, \mu_2 \in \mathbb{R}$, and all $\sigma_1, \sigma_2 > 0$ with $1/2 \leq \sigma_1/\sigma_2 \leq 2$,*

$$\sup_{x\in\mathbb{R}}\left|\varphi\left(\frac{x-\mu_1}{\sigma_1}\right) - \varphi\left(\frac{x-\mu_2}{\sigma_2}\right)\right| \leq 4\frac{|\sigma_1-\sigma_2|}{\sigma_1\vee\sigma_2} + \frac{|\mu_1-\mu_2|}{\sigma_1\vee\sigma_2}.$$

Proof. Without loss of generality we can assume that $\sigma_1 \leq \sigma_2$. Using the triangle inequality, we first write

$$\begin{aligned}\left|\varphi\left(\frac{x-\mu_1}{\sigma_1}\right) - \varphi\left(\frac{x-\mu_2}{\sigma_2}\right)\right| &\\ \leq \left|\varphi\left(\frac{x-\mu_1}{\sigma_1}\right) - \varphi\left(\frac{x-\mu_2}{\sigma_1}\right)\right| &+ \left|\varphi\left(\frac{x-\mu_2}{\sigma_1}\right) - \varphi\left(\frac{x-\mu_2}{\sigma_2}\right)\right| \\ \leq \sup_{u\in\mathbb{R}}\left|\varphi\left(u + \frac{\mu_1-\mu_2}{\sigma_1}\right) - \varphi(u)\right| &+ \sup_{u\in\mathbb{R}}\left|\varphi\left(\frac{\sigma_1}{\sigma_2}u\right) - \varphi(u)\right|.\end{aligned} \tag{3.27}$$

The first term of the rhs of equation (3.27) is obviously bounded by $|\mu_1-\mu_2|/\sigma_1$. Regarding the second term of the rhs of equation (3.27),

$$\left|\varphi\left(\frac{\sigma_1}{\sigma_2}u\right) - \varphi(u)\right| \leq |\sigma_1/\sigma_2 - 1|\left(\frac{\sigma_1}{\sigma_2}\vee 1\right)^2 \sup_x x^2\varphi(x),$$

which terminates the proof. ■

Proposition 3.12 – *Let $X \sim \mathrm{SGa}(\alpha, 1)$, with $0 < \alpha \leq 1$. Then for any $x \in \mathbb{R}$ and any $0 < \delta \leq 1/2$ we have $\mathbb{P}\{|X-x| \leq \delta\} \geq \frac{\delta e^{-2|x|}}{3e\Gamma(\alpha)}$.*

Proof. Assume for instance that $x \geq 0$. Recalling that X is distributed as the difference of two independent $\mathrm{Ga}(\alpha, 1)$ distributed random variables, it follows

$$\mathbb{P}\{|X-x| \leq \delta\} \geq \frac{1}{\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}e^{-y}\frac{1}{\Gamma(\alpha)}\int_{x+y}^{x+y+\delta} z^{\alpha-1}e^{-z}\,dz\,dy.$$

Because $\alpha \leq 1$, the mapping $z \mapsto z^{\alpha-1}e^{-z}$ is monotonically decreasing on $\mathbb{R}^+$, then the last integral in the rhs of the previous equation is lower bounded by $\delta(x+y+\delta)^{\alpha-1}e^{-(x+y+\delta)} \geq \delta e^{-2(x+y+\delta)}$. Then

$$\mathbb{P}\{|X-x| \leq \delta\} \geq \frac{\delta e^{-2(x+\delta)}}{\Gamma(\alpha)^2}\int_0^\infty y^{\alpha-1}e^{-3y}\,dy \geq \frac{3^{-\alpha}e^{-2(x+\delta)}}{\Gamma(\alpha)}\delta \geq \frac{\delta e^{-2|x|}}{3e\Gamma(\alpha)}.$$

The proof when $x < 0$ is obvious. ■

Bibliography

L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques*, volume 42, pages 273–325, 2006.

N. Bochkina and J. Rousseau. Adaptive density estimation based on a mixture of Gammas. *ArXiv e-prints*, May 2016.

A. Canale and P. De Blasi. Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404, 02 2017. doi: 10.3150/15-BEJ746. URL http://dx.doi.org/10.3150/15-BEJ746.

R. de Jonge and J. H. van Zanten. Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38:3300–3320, 2010.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 02 2007. doi: 10.1214/009053606000001172. URL http://dx.doi.org/10.1214/009053606000001172.

S. Ghosal and A. W. Van Der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, pages 1233–1263, 2001.

S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

S. Ghosal, A. Van Der Vaart, et al. Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007a.

S. Ghosal, A. Van Der Vaart, et al. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007b.

A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on r^d. *Probability Theory and Related Fields*, 159(3):479–543, 2014. ISSN 1432-2064. doi: 10.1007/s00440-013-0512-1. URL http://dx.doi.org/10.1007/s00440-013-0512-1.

T. Hangelbroek and A. Ron. Nonlinear approximation using gaussian kernels. *Journal of Functional Analysis*, 259(1):203–219, 2010.

W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. Wavelets. In *Wavelets, Approximation, and Statistical Applications*, pages 1–16. Springer, 1998.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK, 2010.

J. F. C. Kingman. *Poisson processes*, volume 3. Oxford university press, 1992.

W. Kruijer, J. Rousseau, and A. van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257, 2010.

Z. Naulet and E. Barat. Some aspects of symmetric gamma process mixtures. *arXiv preprint arXiv:1504.00476*, 2015.

P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Adaptive density estimation: a curse of support? *Journal of Statistical Planning and Inference*, 141:115–139, 2011.

J.-B. Salomond. Bayesian testing for embedded hypotheses with application to shape constrains. *arXiv preprint arXiv:1303.6466*, 2013.

C. Scricciolo. Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Analysis*, 9:475–520, 2014.

W. Shen, S. T. Tokdar, and S. Ghosal. Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

R. L. Wolpert, M. A. Clyde, and C. Tu. Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *The Annals of Statistics*, pages 1916–1962, 2011.

CHAPTER 4

Bayesian nonparametric estimation for Quantum Homodyne Tomography

Résumé

Nous estimons l'état quantique d'un faisceau de lumière à partir des résultats bruités de mesures de tomographie quantique homodyne, réalisées sur des systèmes quantiques identiquement préparés. Nous proposons deux approches Bayésiennes non paramétriques. La première approche est basée sur les modèles de mélanges et est illustrée au travers d'exemples de simulation. La seconde approche est basée sur une expansion sur une base. Nous étudions les performances théoriques de la seconde approche en quantifiant la vitesse de contraction de la distribution a posteriori autour du vrai état quantique dans la métrique L^2.

Abstract

We estimate the quantum state of a light beam from results of quantum homodyne tomography noisy measurements performed on identically prepared quantum systems. We propose two Bayesian nonparametric approaches. The first approach is based on mixture models and is illustrated through simulation examples. The second approach is based on random basis expansions. We study the theoretical performance of the second approach by quantifying the rate of contraction of the posterior distribution around the true quantum state in the L^2 metric.

4.1. Introduction

Quantum Homodyne Tomography (QHT), is a technique for reconstructing the quantum state of a monochromatic light beam in cavity (Artiles et al., 2005). Unlike classical optics, the predictions of quantum optics are probabilistic so that we cannot in general infer the result of a single measurement, but only the distribution of possible outcomes. The quantum state of a monochromatic light beam in cavity is a positive, self-adjoint and trace-class operator ρ acting on the Hilbert space $L^2(\mathbb{R})$. We should here distinguish the *pure states* which are projection operators onto one-dimensional subspaces of $L^2(\mathbb{R})$, and *mixed-states* which are all the other possible states.

Having prepared a quantum system in state ρ, the aim of the physicist is to perform *measurement* of certain *observables*. Mathematically speaking, an observable A is a self-adjoint operator on $L^2(\mathbb{R})$. A measurement is a mapping which assigns to an observable A and a state ρ a probability measure μ_A on $\mathbb{R}$; this mapping is given by the so-called *Born-von Neumann formula* (Hall, 2013).

Two observables of interest in quantum optics correspond to the measurements of the *electric field* and the *magnetic field* of a light beam, and are given respectively by the operator $\mathbf{Q}$ and $\mathbf{P}$ with domains $D(\mathbf{Q}) := \{\psi \in L^2(\mathbb{R}) : x \mapsto x\psi(x) \in L^2(\mathbb{R})\}$ and $D(\mathbf{P}) := \{\psi \in L^2(\mathbb{R}) : x \mapsto \psi'(x) \in L^2(\mathbb{R})\}$. The operarors $\mathbf{Q}$ and $\mathbf{P}$ act on $D(\mathbf{Q})$, respectively $D(\mathbf{P})$, as

$$\mathbf{Q}\psi(x) = x\psi(x), \text{ and } \qquad \mathbf{P}\psi(x) = -i\psi'(x).$$

The derivative in the definitions of $D(\mathbf{P})$ and $\mathbf{P}$ is understood in the distributional sense.

By virtue of the *Heisenberg uncertainty principle* (Hall, 2013), the observables $\mathbf{P}$ and $\mathbf{Q}$ cannot be measured simultaneously; that is there is no joint probability distribution associated to the simultaneous measurement of $\mathbf{P}$ and $\mathbf{Q}$. Nevertheless, the *Wigner density* $W_\rho : \mathbb{R}^2 \to \mathbb{R}$, with respect to the Lebesgue measure on $\mathbb{R}^2$, as defined below, is the closest object to a joint probability density function associated to the joint measurement of $\mathbf{P}$ and $\mathbf{Q}$ on a system in state ρ. The Wigner distribution satisfies $\int_{\mathbb{R}^2} W_\rho = 1$, and its marginals on any direction are *bona-fide* probability density functions. In general, however, W_ρ fails to be a proper joint probability density function, as it can take negative values, reflecting the non classicality of the quantum state ρ. For a pure state ρ_ψ, $\psi \in L^2(\mathbb{R})$, the Wigner quasi-probability density of ρ_ψ is defined as

$$W_\psi(x,\omega) := \int_{\mathbb{R}} \psi(x + t/2)\overline{\psi(x - t/2)}e^{-2\pi i\omega t}dt, \qquad (x,\omega) \in \mathbb{R}^2. \tag{4.1}$$

We delay to later the definition of the Wigner distribution for mixed states, which will follow from the definition for pure states in a relatively straightforward fashion. Here we take the opportunity to say that whenever we will be concerned with pure states, we will identify the state ρ_ψ to the function $\psi \in L^2(\mathbb{R})$, and talk abusively about the state ψ.

Although we cannot measure simultaneously the observables $\mathbf{P}$ and $\mathbf{Q}$, it is possible to measure the *quadrature observables*, defined as $\mathbf{X}_\theta := \mathbf{Q}\cos\theta + \mathbf{P}\sin\theta$ for all $\theta \in [0,\pi]$. We denote by X_θ^ρ the random variable whose distribution is the measurement of $\mathbf{X}_\theta$ on the quantum system in state ρ. Assuming that θ is drawn uniformly from $[0,\pi]$, the joint probability density function (with respect to the Lebesgue measure on $\mathbb{R} \times [0,\pi]$) for (X_θ^ρ, θ) is given by the *Radon transform* of the Wigner distribution W_ρ, that is

$$p_\rho(x,\theta) := \frac{1}{\pi}\int_{\mathbb{R}} W_\rho(x\cos\theta - \xi\sin\theta, x\sin\theta + \xi\cos\theta)\,d\xi, \quad (x,\theta) \in \mathbb{R}\times[0,\pi]. \tag{4.2}$$

For a pure state $\psi \in L^2(\mathbb{R})$, there is a convenient way of rewriting the previous equation, as stated for example in (Markus et al., 2010, equation 4.14),

$$p_\psi(x,\theta) = \begin{cases} \frac{1}{2\pi|\sin\theta|}\left|\int_{\mathbb{R}} \psi(z)\,\exp\left(\pi i\frac{\cot\theta}{2}z^2 - \pi i\frac{xz}{\sin\theta}\right)\,dz\right|^2 & \theta \neq 0,\ \theta \neq \pi/2, \\ |\psi(x)|^2/\pi & \theta = 0, \\ |\widehat{\psi}(x)|^2/\pi & \theta = \pi/2, \end{cases} \tag{4.3}$$

where $\widehat{\psi}$ is the Fourier transform of ψ (according to the convention defined in the next section of the paper). Equation (4.3) emphasizes that for any (x,θ) we indeed have $p_\psi(x,\theta) \geq 0$, a fact that remains true for mixed states, but which is not so clear from the definition of equation (4.2).

Quantum homodyne tomography is an experiment that allow for measuring the quadrature observables $\mathbf{X}_\theta$ for a monochromatic light beam in cavity in state ρ. Here we consider the situation when we perform identical and independent measurements of $\mathbf{X}_\theta$ on n quantum systems in the same state ρ, with θ spread uniformly over $[0, \pi]$. Following Butucea et al. (2007), it turns out that a good model for a realistic quantum homodyne tomography must take into account noise on observations.

In practice, the noise is mostly due to the fact that a number of photons fails to be detected. The ability of the detector to detect photons is quantified by a parameter $\eta \in [0, 1]$, called the *efficiency* of the detector. When $\eta = 0$, then the detector fails to detect all photons, whereas $\eta = 1$ corresponds to the ideal case where all the photons are detected. In general, it is assumed that η is known ahead of the measurement process, and η is relatively close to one, according to the physicists. Then, from Butucea et al. (2007, section 2.4), a more realistic model for quantum homodyne tomography is to consider that we observe the random variables (given θ)

$$Y_\theta^\rho = X_\theta^\rho + \sqrt{\frac{1-\eta}{\eta}} X_\theta^{\mathrm{vac}},$$

where $X_\theta \sim p_\rho(\cdot \mid \theta)$, and X_θ^{vac} is the random variable whose distribution is the measurement of $\mathbf{X}_\theta$ on the *vacuum state* and is assumed independent of X_θ^ρ. Here we adopt the convention that the vacuum state is the projection operator onto $x \mapsto 2^{-1/4} \exp(-\pi x^2)$. It turns out from equations (4.1) and (4.4) that X_θ^{vac} has a normal distribution with mean zero and variance[1] $1/(4\pi)$. This leads to the following *efficiency corrected* probability density function of observations,

$$p_\psi^\eta(y, \theta) := \sqrt{\frac{2}{1-\eta}} \int_{\mathbb{R}} p_\psi(x, \theta) \exp\left[-\frac{2\pi\eta}{1-\eta} (x-y)^2 \right] dx. \tag{4.4}$$

To shorten notations, we define

$$\gamma := \frac{\pi(1-\eta)}{2\eta}, \text{ and } \quad G_\gamma(x) := \sqrt{\pi/\gamma} \exp\left[-\pi^2 x^2 / \gamma \right], \tag{4.5}$$

so that we have $p_\psi^\eta(y, \theta) = [p_\psi(\cdot, \theta) * G_\gamma](y)$, where $*$ denote the convolution product.

To summarize the statistical model we are considering in this paper, we aim at estimating the Wigner density function W_ρ, or better directly the state ρ, from n independent and indentically distributed noisy observations $(Y_1, \theta_1), \ldots, (Y_n, \theta_n)$ distributed according to the distribution that has the density function of equation (4.4) with respect to the Lebesgue measure on $\mathbb{R} \times [0, \pi]$.

The problem of QHT is a statistical nonparametric *ill-posed* inverse problem that has been relatively well studied from a frequentist point of view in the last few years, and now quite well understood. We mention here only papers with theoretical analysis of the performance of their estimation procedure. We should classify frequentist methods in two categories, depending on whether they are based on estimating the state ρ, or estimating W_ρ (although $\rho \mapsto W_\rho$ is one-to-one, methods based on estimating W_ρ don't permit to do the reverse path from $W_\rho \mapsto \rho$).

[1] Some readers may have noticed that the variance here is different that in Butucea et al. (2007). This comes from a different convention for defining the vacuum state.

The estimation of the state ρ from QHT measurements has been considered in the ideal situation ($\eta = 1$, no noise) by Artiles et al. (2005), while the noisy setting is investigated in Aubry et al. (2008) under Frobenius-norm risk. For smoothness class of realistic states $\mathcal{R}(C, B, r)$, an adaptive estimation procedure has been proposed by Alquier et al. (2013) and an upper bound for the Frobenius-norm risk is given. Goodness-of-fit testing is investigated in Méziani (2008).

Regarding frequentist methods for estimating W_ρ, the first result goes back to Guţă and Artiles (2007), where sharp minimax results are given over a class of smooth Wigner functions $\mathcal{A}(\beta, r = 1, L)$, under the pointwise risk. The noisy framework has been considered in Butucea et al. (2007); authors obtain the minimax rates of convergence under the pointwise risk and propose an adaptive estimator over the set of parameters $\beta > 0$, $r \in (0, 1)$ that achieve nearly minimax rates. In the same time Méziani (2007) explored the estimation of a quadratic functional of the Wigner function, as an estimator of the purity of the state. In, Aubry et al. (2008) an upper bound for the L^2-norm risk over the class $\mathcal{R}(C, B, r)$ is given. More recently, Lounici et al. (2015) established the first sup-norm risk upper bound over $\mathcal{A}(\beta, r, L)$, as well as the first minimax lower bounds for both sup-norm and L^2-norm risk; they also provide an adaptive estimator that achieve nearly minimax rates for both sup-norm and L^2-norm risk over $\mathcal{A}(\beta, r, L)$ for all $\beta > 0$ and $r \in (0, 2)$.

To our knowledge, no Bayesian nonparametric method has been proposed to address the problem of QHT with noisy data, a gap that we try to fill with this paper. In particular, after having introduced preliminary notions in the next section, we propose two families of prior distributions over pure states that can be useful in practice, namely *mixtures of coherent-states* and *random Wilson series*. Regarding mixed-states, we will discuss how we can straightforwardly extend the prior distributions over pure states onto prior distributions over mixed states. After presenting simulation results, we will investigate posterior rates of contraction for random Wilson series in the main section of the paper. Rates of contraction, or even consistency, is still challenging for coherent states mixtures, a fact that will be discussed more thoroughly in section 4.5.2.

4.2. Preliminaries

4.2.1. Notations

For $x, y \in \mathbb{R}^d$, xy denote the euclidean inner product of x and y, and $\|x\|$ is the euclidean norm of a vector $x \in \mathbb{R}^d$. For any function f, we denote by $\check{f}$ the involution $\check{f}(x) = f(-x)$. We use the notation $\|\cdot\|_p$ for the norm of the spaces $L^p(\mathbb{R}^d)$.

We use the following convention for the Fourier transform of a function $f \in L^1(\mathbb{R}^d)$.

$$\mathscr{F}f(\omega) := \widehat{f}(\omega) := \int_{\mathbb{R}^d} f(x) e^{-2\pi i x\omega}\, dx, \qquad \forall \omega \in \mathbb{R}^d.$$

Then, whenever $f \in L^1(\mathbb{R}^d)$ and $\mathscr{F}f \in L^1(\mathbb{R}^d)$, the inverse Fourier transform $\mathscr{F}^{-1}\mathscr{F}f = f$ is well defined and given by

$$f(x) = \int_{\mathbb{R}^d} \widehat{f}(\omega) e^{2\pi i \omega x}\, d\omega, \qquad \forall x \in \mathbb{R}^d.$$

Regarding the space $L^2(\mathbb{R}^d)$, we use the convention that the inner product $\langle \cdot, \cdot \rangle : L^2(\mathbb{R}^d) \times L^2(\mathbb{R}^d) \to \mathbb{C}$ is linear in the first argument and antilinear in the second argument, that is for two functions $f, g \in L^2(\mathbb{R}^d)$ we define $\langle f, g \rangle := \int_{\mathbb{R}^d} f(x)\overline{g(x)}\, dx$, where $\overline{z}$ is the complex conjugate of $z \in \mathbb{C}$. The *unit circle* of $L^2(\mathbb{R}^d)$ will be denoted by $\mathbb{S}^2(\mathbb{R}^d)$; that is $\mathbb{S}^2(\mathbb{R}^d) := \{f \in L^2(\mathbb{R}^d) : \|f\|_2 = 1\}$.

We shall sometimes encounter the *Schwartz* space $\mathcal{S}(\mathbb{R}^d)$; that is the space of all infinitely differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ for which $|x^\alpha D^\beta f(x)| < +\infty$ for all $\alpha, \beta \in \mathbb{N}^d$, with the convention $x^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d}$ and $D^\beta f = \partial^{\beta_1 + \dots + \beta_d} f / (\partial x_1^{\beta_1} \dots \partial x_d^{\beta_d})$.

Dealing with probability distributions, we consider the *Hellinger* distance $H^2(P, Q) := \frac{1}{2} \int (\sqrt{dP/d\lambda} - \sqrt{dQ/d\lambda})^2 \, d\lambda$, for any probability measures P, Q absolutely continuous with respect to a common measure λ.

We denote by P_ρ, respectively P_ρ^η, the distributions that admit equation (4.2), respectively equation (4.4), as density with respect to the Lebesgue measure on $\mathbb{R} \times [0, \pi]$. When $\rho \equiv \rho_\psi$ denote a pure state, we denote the previous distribution by P_ψ and P_ψ^η, respectively.

Finally, inequalities up to a generic constant are denoted by the symbols $\lesssim$ and $\gtrsim$, where $a \lesssim b$ means $a \leq Cb$ for a constant $C > 0$ with no consequence on the result of the proof.

4.2.2. Coherent states

In quantum optics, a coherent state refers to a state of the quantized electromagnetic field that describes a classical kind of behavior.

Let $T_x f(y) := f(y - x)$, $M_\omega f(y) = e^{2\pi i \omega y} f(y)$, denote the translation and modulation operators, respectively, and g a window function with $\|g\|_2 = 1$; most of time g is chosen as $g(x) = 2^{-1/4} \exp(-\pi x^2)$. Mathematically speaking, coherent states are pure states ρ_ψ, that is projection operators, described by a *wave-function* ψ belonging to

$$\left\{ \psi \in L^2(\mathbb{R}) : \psi = T_x M_\omega g \quad (x, \omega) \in \mathbb{R}^2 \right\}.$$

Note that the operators T_x and M_ω are isometric on $L^p(\mathbb{R}^d)$ and $\|f\|_p = \|T_x M_\omega f\|_p$ for any $1 \leq p \leq \infty$, all $f \in L^p(\mathbb{R}^d)$ and all $x, \omega \in \mathbb{R}$.

4.2.3. Wilson bases

Daubechies et al. (1991) proposed simple *Wilson bases* of exponential decay. They constructed a real-valued function φ such that for some $a, b > 0$,

$$|\varphi(x)| \lesssim e^{-a|x|}, \qquad |\widehat{\varphi}(\omega)| \lesssim e^{-b|\omega|},$$

and such that the φ_{lm}, $l \in \mathbb{N}$, $m \in \frac{1}{2}\mathbb{Z}$ defined by

$$\varphi_{lm}(x) := \begin{cases} \varphi(x - 2m) & \text{if } l = 0, \\ \sqrt{2}\varphi(x - m)\cos(2\pi l x) & \text{if } l \neq 0 \text{ and } 2m + l \text{ is even}, \\ \sqrt{2}\varphi(x - m)\sin(2\pi l x) & \text{if } l \neq 0 \text{ and } 2m + l \text{ is odd}, \end{cases}$$

constitute an orthonormal base for $L^2(\mathbb{R})$. Following Gröchenig (2001, section 8.5), we may rewrite φ_{lm} in a convenient form for the sequel, emphasizing the relationship with coherent states,

$$\varphi_{lm} = c_l T_m (M_l + (-1)^{2m+l} M_{-l})\varphi, \qquad (l,m) \in \mathbb{N} \times \tfrac{1}{2}\mathbb{Z}, \tag{4.6}$$

where $c_0 := 1/2$ and $c_l := 1/\sqrt{2}$ for $l \geq 1$.

4.3. Prior distributions

We recall that a pure state ρ_ψ is a projection operator onto a one-dimensional subspace of $L^2(\mathbb{R})$. Before giving the methodology for estimating general states, we introduce two types of prior distribution over pure-states. More precisely, we first define two probability distributions over $\mathbb{S}^2(\mathbb{R})$, that can be trivially identified with the set of pure-state through the mapping $\mathbb{S}^2(\mathbb{R}) \ni \psi \mapsto \rho_\psi$; then we will show how to enlarge these prior distributions to handle mixed states.

The first prior model is based on Gamma mixtures, whereas the second is based on the Wilson base of exponential decay.

4.3.1. Gamma Process mixtures of coherent states

For any finite positive measure α on the measurable space $(X, \mathcal{X})$, let Π_α denote the Gamma process distribution with parameter α; that is, a $Q \sim \Pi_\alpha$ is a measure on $(X, \mathcal{X})$ such that for any disjoints $B_1, \ldots, B_k \in \mathcal{X}$ the random variables $Q(B_1), \ldots, Q(B_k)$ are independent random variables with distributions $\mathrm{Ga}(\alpha(B_i), 1)$, $i = 1, \ldots, k$.

We suggest a mixture of coherent states as prior distribution on the wave function ψ. For a Gamma random measure Q on $\mathbb{R}^2 \times [0, 2\pi]$, our model may be summarized by the following hierarchical representation. Recall that P_ψ^η denote the probability distribution having the density of equation (4.4), with $\rho = \rho_\psi$ the projection operator onto ψ.

$$\begin{aligned} (Y_1, \theta_1), \ldots, (Y_n, \theta_n) &\overset{\text{i.i.d}}{\sim} P_\psi^\eta, \quad \text{with } \psi = \widetilde{\psi}/\|\widetilde{\psi}\|_2 \\ \widetilde{\psi}(z) &= \int_{\mathbb{R}^2 \times [0,2\pi]} e^{i\phi} T_x M_\omega g(z)\, Q(dx d\omega d\phi) \\ Q &\sim \Pi_\alpha. \end{aligned}$$

4.3.2. Random Wilson series

Let (φ_{lm}) be the orthonormal Wilson base with exponential decay of section 4.2.3. For any positive number Z, let Λ_Z be the spherical array

$$\Lambda_Z := \Big\{ (l,m) \in \mathbb{N} \times \tfrac{1}{2}\mathbb{Z} \,:\, l^2 + m^2 < Z^2 \Big\}.$$

Also define the simplex Δ_Z in the ℓ_2 metric as

$$\Delta_Z := \Big\{ \mathbf{p} = (p_{lm})_{(l,m)\in\Lambda_Z} \,:\, \textstyle\sum_{(l,m)\in\Lambda_Z} p_{lm}^2 = 1,\ p_{lm} \geq 0 \Big\}.$$

We consider the following prior distribution Π on $\mathbb{S}^2(\mathbb{R})$. Let P_Z be a distribution over $\mathbb{R}^+$ and draw $Z \sim P_Z$. Given Z, draw $\mathbf{p}$ from a distribution $G(\cdot \mid Z)$ over the simplex Δ_Z.

Independently of $\mathbf{p}$, draw $\zeta = (\zeta_{lm})_{(l,m)\in\Lambda_Z}$ from a distribution $P_\zeta(\cdot \mid Z)$ over $[0, 2\pi]^{|\Lambda_Z|}$ and set

$$\psi := \sum_{(l,m)\in\Lambda_Z} p_{lm} e^{i\zeta_{lm}} \varphi_{lm}.$$

Note that (φ_{lm}) is orthonormal, thus $\|\psi\|_2^2 = \sum_{(l,m)\in\Lambda_Z} p_{lm}^2 = 1$ almost-surely, that is $\psi \in \mathbb{S}^2(\mathbb{R})$ almost-surely.

4.3.3. Estimation of mixed states

The set of quantum states is a convex set. According to the Hilbert-Schmidt theorem on the canonical decomposition for compact self-adjoint operators, for every quantum state ρ there exists an orthonormal set $(\psi_n)_{n=1}^N$ in $L^2(\mathbb{R})$ (finite or infinite, in the latter case $N = \infty$), and $\alpha_n > 0$ such that

$$\rho = \sum_{n=1}^{N} \alpha_n \rho_{\psi_n}, \text{ and } \qquad \mathrm{Tr}\rho = \sum_{n=1}^{N} \alpha_n = 1.$$

The $(\alpha_n)_{n=1}^N$ are the non-zero eigenvalues of ρ and $(\rho_{\psi_n})_{n=1}^N$ projection operators onto $(\psi_n)_{n=1}^N$. Thus every mixed state is a convex linear combination of pure states. In particular, for any state ρ we have

$$W_\rho(x, \omega) = \sum_{n=1}^{N} \alpha_n W_{\psi_n}(x, \omega),$$

making relatively straightforward the extension of priors over pure states onto priors over general states. In other words, a prior distribution over general states can be constructed as a mixture of pure states by a random probability measure.

4.4. Simulations examples

4.4.1. Simulation procedure

We test the Gamma process mixtures of coherent states on two examples of quantum states, corresponding to the Schrödinger cat and 2-photons states, that are respectively described by the wave functions

$$\psi_{\mathrm{cat}}^{x_0}(x) := \frac{\exp(-\pi(x - x_0)^2) + \exp(-\pi(x + x_0)^2)}{2^{1/4}\sqrt{1 + \exp(-2\pi x_0^2)}},$$

$$\psi_2(x) := 2^{-1/4}(4\pi x^2 - 1)\exp(-\pi x^2).$$

Using equations (4.1) and (4.2), it is seen that the conditional density on $\theta \in [0, \pi]$ corresponding to the measurement of $\mathbf{X}_\theta$ on the systems in states $\psi_{\mathrm{cat}}^{x_0}$ and ψ_2 are respectively given by

$$p_{\mathrm{cat}}^{x_0}(x \mid \theta) \propto \sqrt{2}e^{-2\pi(x - x_0\cos\theta)^2} + \sqrt{2}e^{-2\pi(x + x_0\cos\theta)^2} + 2e^{-2\pi x_0^2}\frac{\sqrt{2}e^{-2\pi x^2}\cos(4\pi x x_0 \sin\theta)}{e^{-2\pi x_0^2 \sin^2\theta}},$$

and,

$$p_2(x \mid \theta) = 2^{-1/2}(4\pi x^2 - 1)^2 e^{-2\pi x^2}.$$

Note that $p_{\text{cat}}^{x_0}(\cdot \mid \theta)$ is not a mixture density, since one term can take negative values. Conditionally on θ drawn uniformly on $[0, \pi]$, we simulate $n = 2000$ observations from the Schrödinger cat state with $x_0 = 2$ using $p_{\text{cat}}^{x_0}(\cdot \mid \theta)$ and the rejection sampling algorithm with candidate distribution $\frac{1}{2}\mathcal{N}(-x_0 \cos\theta, 1/(4\pi)) + \frac{1}{2}\mathcal{N}(x_0 \cos\theta, 1/(4\pi))$. Similarly, we simulate $n = 2000$ observations from the 2-photons state using the rejection sampling algorithm with a Laplace candidate distribution. A Gaussian noise is added to observations according to equation (4.4), where we choose $\eta = 0.95$, a reasonable efficiency the physicists say.

4.4.2. Simulation results

We use the algorithm of Naulet and Barat (2015) for simulating samples from posterior distributions of Gamma process mixtures. The base measure α on $\mathbb{R}^2 \times [0, 2\pi]$ of the mixing Gamma process is taken as the independent product of a normal distribution on $\mathbb{R}^2$ with covariance matrix $\text{diag}(1/2, 1/2)$ and the uniform distribution on $[0, 2\pi]$.

We ran 3000 iterations of the algorithm with $p = 50$ particles, leading to an acceptance ratio of approximately 60% for the particle moves and the both datasets. All random-walk Metropolis-Hastings steps are Gaussians, with amplitudes chosen to achieve approximately 25% acceptance rates. All the statistics were computed using only the 2000 last samples provided by the algorithm.

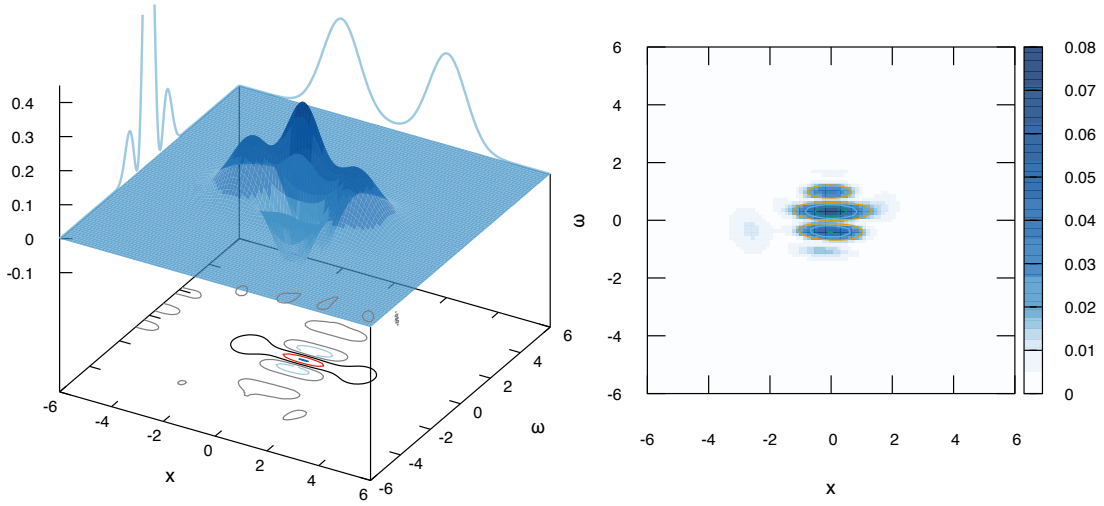

Figure 4.1: Left: Average of Wigner distribution samples from the posterior distribution of the mixture of coherent states prior given 2000 quantum homodyne tomography observations simulated from a Schrödinger cat state. Right: View map of the absolute value of the difference between the posterior mean estimate of the Wigner distribution and the true Wigner distribution.

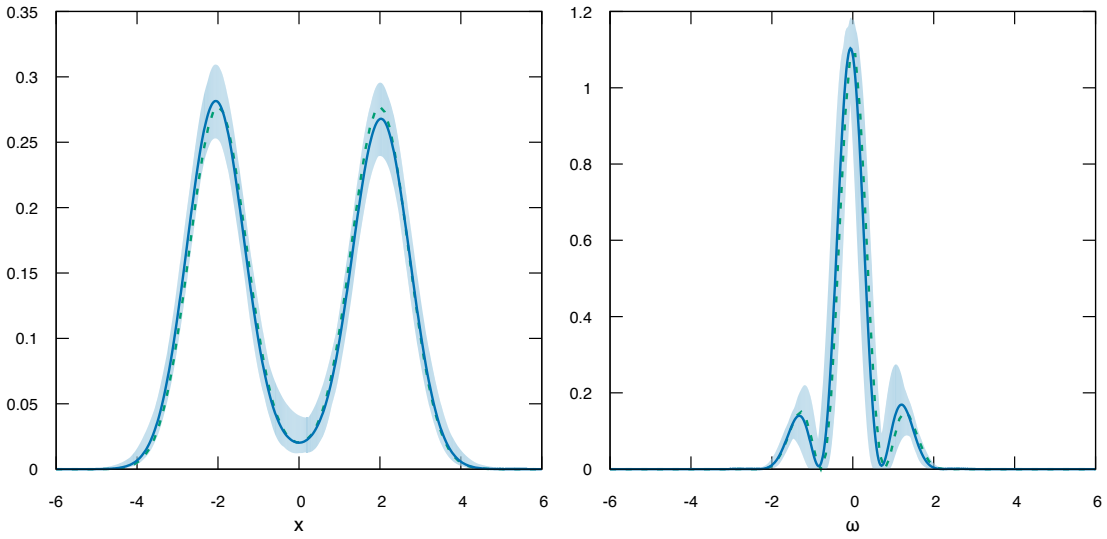

Figure 4.2: Marginals of the Wigner distribution samples from the posterior distribution of the mixture of coherent states prior given 2000 quantum homodyne tomography observations simulated from a Schrödinger cat state. In straight line the posterior mean estimate, whereas the dashed lines corresponds to the true marginals. The 95% credible intervals for the sup-norm distance are drawn in shading.

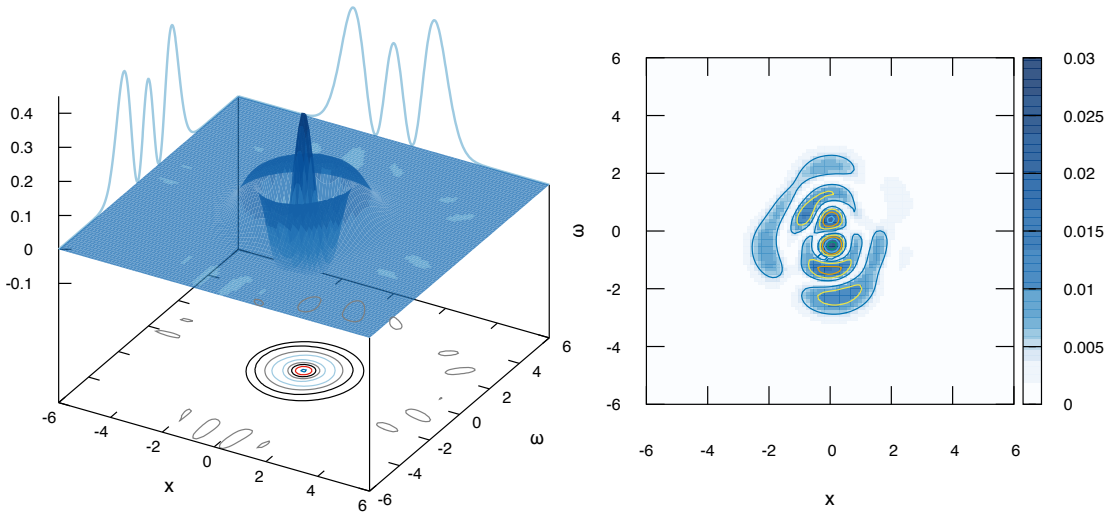

Figure 4.3: Left: Average of Wigner distribution samples from the posterior distribution of the mixture of coherent states prior given 2000 quantum homodyne tomography observations simulated from a 2-photons state. Right: View map of the absolute value of the difference between the posterior mean estimate of the Wigner distribution and the true Wigner distribution.

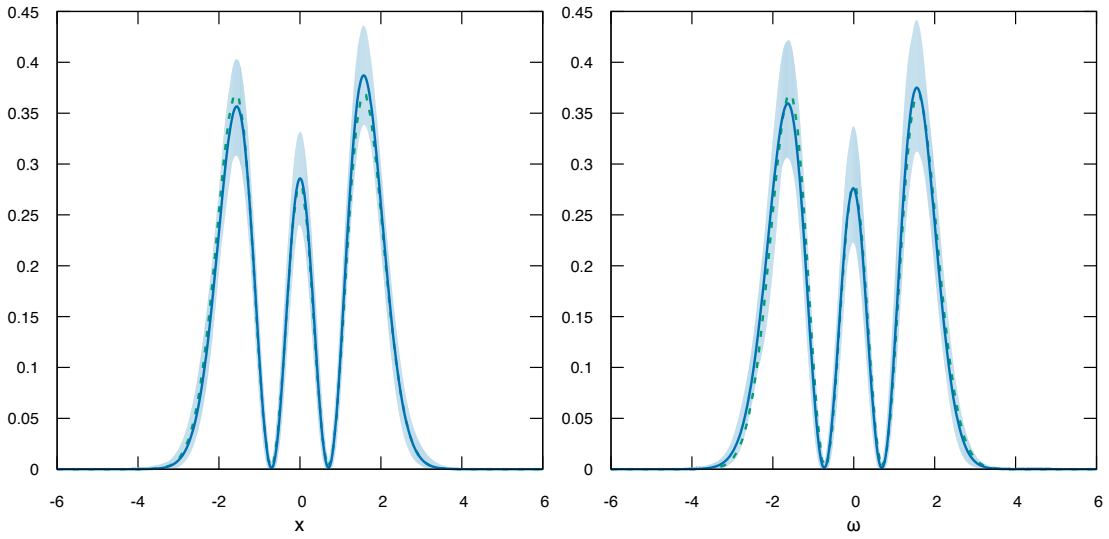

Figure 4.4: Marginals of the Wigner distribution samples from the posterior distribution of the mixture of coherent states prior given 2000 quantum homodyne tomography observations simulated from a 2-photons state. In straight line the posterior mean estimate, whereas the dashed lines corresponds to the true marginals. The 95% credible intervals for the sup-norm distance are drawn in shading.

Figures 4.1 and 4.3 represent the average of posterior samples of the Wigner distribution for the Schrödinger cat state, and the 2-photons state, respectively. Because it is hard to distinguish between the posterior mean estimator and the true Wigner distribution, we added to the figures a view map of the absolute value of the difference between the evaluated posterior mean and the true Wigner distribution.

Figures 4.2 and 4.4 show the marginals of the posterior mean estimates of Wigner distributions for our two examples. We represented the true marginals in dashed lines, as well as the posterior credible bands provided by the algorithm, which we computed by retaining the 95% samples with the smaller sup-norm distance from the posterior mean estimator of the marginals.

Compared to other classical methods in this area, our estimate is non linear, preventing easy computations. To our knowledge, however, none of the current approaches can preserve the physical properties of the true Wigner function (non negativity of marginal distributions, bounds) whereas our approach does guarantee preservation of all physical properties.

4.5. Rates of contraction for random series priors

In this section, we establish posterior convergence rates in the quantum homodyne tomography problem, for estimating pure states. Unfortunately, to get such result we need a fine control of the $L^2(\mathbb{R})$ norm of random functions drawn from the prior distribution, which remains challenging for mixtures of coherent states. However, dealing with Wilson bases, the control of the $L^2(\mathbb{R})$ norm is straightforward and we are able to obtain posterior

concentration rates.

4.5.1. Preliminaries on function spaces

To establish posterior concentration rates, we describe suitable classes of functions that can be well approximated by partial sums of Wilson bases elements; these functional classes are called *ultra-modulation spaces*. To this aim, we need the following ingredients: the short-time Fourier transform (STFT), a class of windows and a class of weights. For a non-zero window function $g \in L^2(\mathbb{R})$, the short-time Fourier transform of a function $f \in L^2(\mathbb{R})$ with respect to the window g is given by

$$V_g f(x,\omega) := \langle f,\, M_\omega T_x g\rangle = \int_{\mathbb{R}} f(t)\overline{g(t-x)}e^{-2\pi i\omega t}\,dt, \quad (x,\omega)\in\mathbb{R}^2. \tag{4.7}$$

We also need a class of analyzing windows g with sufficiently good time-frequency localization properties. Following, Cordero (2007); Cordero et al. (2005); Gröchenig and Zimmermann (2004), we use the *Gelfand-Shilov* space $\mathcal{S}_1^1(\mathbb{R})$. For any $d \geq 1$, a function $f : \mathbb{R}^d \to \mathbb{C}$ belongs to the Gelfand-Shilov space $\mathcal{S}_1^1(\mathbb{R}^d)$ if $f \in \mathcal{C}^\infty(\mathbb{R}^d)$ and there exist real constants $h > 0$ and $k > 0$ such that

$$\sup_{x\in\mathbb{R}^d} |f(x)e^{h\|x\|}| < +\infty, \qquad \sup_{\omega\in\mathbb{R}^d} |\widehat{f}(\omega)e^{k\|\omega\|}| < +\infty.$$

Next, for $\beta > 0$, $g \in \mathcal{S}_1^1(\mathbb{R})$, and $r \in [0,1)$, we consider the exponential weights on $\mathbb{R}^2$ defined by $x \mapsto \exp(\beta\|x\|^r)$, and we introduce the class of wave-functions

$$\mathcal{C}_g(\beta,r,L) := \left\{\psi \in \mathbb{S}^2(\mathbb{R}) \,:\, \int_{\mathbb{R}^2} |V_g\psi(z)|\,\exp(\beta\|z\|^r)dz \leq L\right\}. \tag{4.8}$$

The class $\mathcal{C}_g(\beta,r,L)$ is reminiscent to *modulation spaces* (Gröchenig, 2001, 2006). Note that it would be interesting to consider $\mathcal{C}_g(\beta,r,L)$ for $r \geq 1$, since most quantum states should fall in these classes. There is, however, at least two limitations for considering $r \geq 1$. First, we use repeatedly in the proofs that $\exp(\beta\|x+y\|^r) \leq \exp(\beta\|x\|^r)\exp(\beta\|y\|^r)$ for $r \leq 1$, which is no longer true when $r > 1$. The previous limitation is indeed not the more serious concerns, since for $r > 1$ we could use that $\exp(\beta\|x+y\|^r) \leq \exp(2^{r-1}\beta\|x\|^r)\exp(2^{r-1}\beta\|y\|^r)$. The more serious problem is that, to our knowledge, there is no Wilson base for $L^2(\mathbb{R})$ whose elements fall into $\mathcal{C}_g(\beta,r,L)$ for $r > 1$ and $\beta > 0$, $L > 0$. The case $r = 1$ is more delicate since it depends on the value of β. For sufficiently small $\beta > 0$, the results proved in this paper for $r < 1$ should also hold for $r = 1$.

Let also notice that, there is a fundamental limit on the growth of the weights in the definition of $\mathcal{C}_g(\beta,r,L)$, imposed by Hardy's theorem. If $r = 2$ and $\beta > \pi/2$, the the corresponding classes of smoothness $\mathcal{C}_g(\beta,r,L)$ are trivial for any $L > 0$ (Gröchenig and Zimmermann, 2001).

A critical point regarding the class $\mathcal{C}_g(\beta,r,L)$ is the dependence on g in the definition. We truly want that for two different windows g_0 and g_1 the corresponding smoothness are the same. Fortunately, we have the following theorem, proved in section 4.A.

Theorem 4.1 – *Let $g, g_0 \in \mathcal{S}_1^1(\mathbb{R})$. For all $\beta, L > 0$ and all $0 \leq r < 1$ there is a constant $C > 0$, depending only on g, g_0, such that embedding $\mathcal{C}_g(\beta,r,L) \subseteq \mathcal{C}_{g_0}(\beta,r,CL)$ holds.*

The STFT and the Wigner transform both aim at having a time-frequency representation of functions in $L^2(\mathbb{R})$, and are deeply linked to each other. However, contrarily to the Wigner transform, the STFT has the advantage of being a linear operator, which is one reason why we prefer to state the class $\mathcal{C}_g(\beta, r, L)$ in term of the STFT instead of the Wigner transform.

4.5.2. Assumptions and results

Before stating the main result of this paper, we need some further assumptions on the random Wilson base series prior, which we state now. To this aim, we need the following definition of the *weighted* simplex $\Delta_Z^w(\beta, r, M)$. For a constant $M > 0$, $\beta > 0$ and $r \in [0, 1)$ let

$$\Delta_Z^w(\beta, r, M) := \left\{ \mathbf{p} \in \Delta_Z \,:\, \textstyle\sum_{(l,m)\in\Lambda_Z} p_{lm} \exp\left(\beta(l^2+m^2)^{r/2}\right) < M \right\}.$$

Then, in the sequel, we assume that

- There is a constant $a_0 > 0$ such that for any sequence $(x_{lm})_{(l,m)\in\Lambda_Z} \in [0, 2\pi]^{|\Lambda_Z|}$,

$$P_\zeta\left(\textstyle\sum_{(l,m)\in\Lambda_Z} |\zeta_{lm} - x_{lm}|^2 \le t \mid Z\right) \gtrsim \exp\left(-a_0 Z^2 \log t^{-1}\right), \quad \forall t \in (0,1).$$

- $P_Z(Z < +\infty) = 1$ and there are constants $a_1, a_2 > 0$ and $b_1 > 2 + r$, such that for all k positive integer large enough

$$P_Z(Z = k) \gtrsim \exp(-a_1 k^{b_1}), \qquad P_Z(Z > k) \lesssim \exp(-a_2 k^{b_1}).$$

- For any constant $C > 0$ and any sequence $\mathbf{q} \in \Delta_Z^w(\beta, r, C)$, there is a constant $a_3 > 0$ such that the distribution $G(\cdot \mid Z)$ satisfy,

$$G\left(\textstyle\sum_{(l,m)\in\Lambda_Z} |p_{lm} - q_{lm}|^2 \le t \mid Z\right) \gtrsim \exp\left(-a_3 Z^{b_1 - r} \log t^{-1}\right), \ \forall t \in (0,1).$$

  We further assume that there exist constants $a_4 \ge 0$, $a_5, c_0 > 0$, and $b_5 > b_1/r$ such that for $x > 0$ large enough

$$G\left(\mathbf{p} \notin \Delta_Z^w(\beta, r, c_0 x^{a_4}) \mid Z \le x^{1/r}\right) \lesssim \exp\left(-a_5 x^{b_5}\right).$$

It is not clear whether or not we can find a distribution G for which the above conditions are satisfied simultaneously for all (β, r, L), eventually with constants a_3, a_4, a_5, b_5 depending on (β, r, L). If such distribution exists, then the rates stated below are easily seen to be adaptive on (β, r, L). In section 4.6, we show that for a given (β, r, L) it is easy to construct a distribution G that satisfies the above conditions, with $a_4 = 2/r$. However, we believe that the proof for adaptive rates must follow a different path, still to be found.

Under the hypothesis above, we will dedicate the rest of the paper to prove the following theorem.

Theorem 4.2 – *Let $\beta, L > 0$ and $r \in (0, 1)$. Let Π be the random Wilson series prior satisfying the assumptions above, and $(Y_1, \theta_1), \ldots, (Y_n, \theta_n)$ be observations coming from*

the statistical model described by equation (4.4), with $0 < \eta < 1$ *and* $\gamma > 0$ *defined in equation (4.5). Then for any* $\psi_0 \in \mathcal{C}_g(\beta, r, L)$*, there is* $M > 0$ *such that*

$$P_{\psi_0}^{\eta,n}\Pi(\|\psi - \psi_0\|_2 \geq M\epsilon_n \mid (Y_1, \theta_1), \ldots, (Y_n, \theta_n)) \to 0,$$
$$\epsilon_n^2 = (\log n)^{2a_4} \exp\left\{-\beta\left(\frac{\log n}{2\gamma}\right)^{r/2} + O(1)\right\}.$$

Note that the same result holds with $\|\psi - \psi_0\|_2$ replaced with $\|W_\psi - W_{\psi_0}\|_2$, because the Wigner transform is isometric from $L^2(\mathbb{R})$ onto $L^2(\mathbb{R}^2)$; see for instance Gröchenig (2001, proposition 4.3.2).

The rates of contraction are relatively slow, a fact that is also pointed out in Butucea et al. (2007). Indeed, the rates are faster than $(\log n)^{-a}$ but slower than n^{-a}, for all $a > 0$. The reason for such bad rates of convergence is to be found in the deconvolution of the Gaussian noise. If one does not carry about deconvoluting the noise, then all the steps in the proof of theorem 4.2 can be mimicked to get weaker a result. In particular, we infer from the results of the paper that the posterior distribution should contracts at nearly parametric rates, *i.e.* at rate $\epsilon_n \approx n^{-1/2}(\log n)^t$ for some $t > 0$, around balls of the form

$$\left\{\psi \in \mathbb{S}^2(\mathbb{R}) \,:\, \int_{\mathbb{R}^2} |\widehat{W}_\psi(z) - \widehat{W}_{\psi_0}(z)|^2\, \widehat{G}_\gamma(\|z\|)^2 dz \leq \epsilon_n^2\right\}, \tag{4.9}$$

whenever $\psi_0 \in \mathcal{C}_g(\beta, r, L)$ for some $\beta, L > 0$ and $r \in (0, 1)$. Moreover, we've made many restrictive assumptions on the prior distribution that can be easily released for those interested only in posterior contraction around balls of the form (4.9).

A natural question regarding the rates found in theorem 4.2 concerns optimality. We do not know yet the minimax lower bounds over the class $\mathcal{C}_g(\beta, r, L)$ for the L^2 risk. However, Butucea et al. (2007); Aubry et al. (2008); Lounici et al. (2015) consider a class $\mathcal{A}(\alpha, r, L)$ that resembles to $\mathcal{C}_g(\beta, r, L)$. More precisely, they define

$$\mathcal{A}(\alpha, r, L) := \left\{W_\rho \,:\, \int |\widehat{W}_\rho(z)|^2 \exp(2\alpha\|z\|^r) dz \leq L^2\right\}.$$

Identifying ρ_ψ with ψ, our proposition 4.7 state the embedding $\mathcal{C}_g(\beta, r, L) \subseteq \mathcal{A}(\beta/2, r, L)$. Hence $\mathcal{C}_g(\beta, r, L)$ is certainly contained in the intersection of a class $\mathcal{A}(\beta/2, r, L)$ with the set of pure states, and it makes sense to compare the rates. To our knowledge, the only minimax lower bound for the quadratic risk known is for the estimation of a state in $\mathcal{A}(\alpha, r = 2, L)$, stated in Lounici et al. (2015). For $r \in (0, 1)$, however, upper bounds for the quadratic risk over $\mathcal{A}(\beta/2, r, L)$ are established in Aubry et al. (2008), and coincide with the rates found here. Therefore, we believe that the rates we found in this paper are optimal.

Let conclude with a few points that are still challenging at this time. First, the rates (or even consistency) for the coherent states mixtures priors appears difficult to establish with the method employed here; the reason comes from the difficulty to control the norm $\|\widetilde{\psi}\|_2$ when $\widetilde{\psi}$ is a coherent states mixture. Regarding Wilson based priors, we already discussed the lack of adaptivity, which clearly deserved to be dug in a near future. Finally, it would be interesting to consider priors based on Gabor frames expansions, as they are more flexible than Wilson bases, and should be computationally more efficient than coherent states mixtures. However, Gabor frames suffer from the same evil that coherent

states, namely the expansions are not unique and it is hard to control from below the L^2 norm of random Gabor expansions.

4.6. Example of priors on the simplex

In this section, we construct a prior on the simplex Δ_Z that satisfy the assumptions of section 4.5.2 for a given (β, r). For all $k \geq 1$, and a constant $M > 0$ to be defined later, we define the sets

$$\mathcal{I}_k := \Big\{(l,m) \in \mathbb{N} \times \tfrac{1}{2}\mathbb{Z} \,:\, (k-1)M \leq \sqrt{l^2+m^2} < kM\Big\}.$$

We assume without loss of generality that $Z = KM$ for an integer $K > 0$; then $\Lambda_Z = \cup_{k=1}^K \mathcal{I}_k$. We then construct the distribution $G(\cdot \mid Z)$ over the simplex Δ_Z as follows. For $k = 2, \dots, K$, let H_k be the uniform distribution over $[0, \sqrt{2}L\exp(-\beta(k^r-1)M^r)]$. Let $\theta_1 := 1$ and for $k = 2, \dots, K$ draw θ_k from H_k independently. The next step is to introduce distributions F_k over the $\mathcal{I}_k$-simplex

$$\mathcal{S}_k := \Big\{(\eta_{lm})_{(l,m)\in\mathcal{I}_k} \,:\, \textstyle\sum_{(l,m)\in\mathcal{I}_k} \eta_{lm}^2 = 1, \quad \eta_{lm} \geq 0\Big\},$$

and draw independently sequences $(\eta_{lm})_{(l,m)\in\mathcal{I}_1}, (\eta_{lm})_{(l,m)\in\mathcal{I}_2}, \dots, (\eta_{lm})_{(l,m)\in\mathcal{I}_K}$, according to distributions $F_1, F_2, \dots, F_K$. Finally, the sequence $\mathbf{p} = (p_{lm})_{(l,m)\in\Lambda_Z}$ drawn from $G(\cdot \mid Z)$ is defined to be such that

$$p_{lm} := \frac{\eta_{lm}\theta_k \mathbb{1}\left((l,m) \in \mathcal{I}_k\right).}{\sum_{k=1}^K \theta_k^2}.$$

Now we prove that we can chose reasonably $M > 0$ and the distributions $F_1, F_2 \dots$ to met the assumptions of section 4.5.2. The proofs of the next two propositions are to be found in section 4.D.

Proposition 4.1 – *There is a constant $c_0 > 0$ such that for any $Z \geq 0$ the sequence $(p_{lm})_{(l,m)\in\Lambda_Z}$ belongs to $\Delta_Z^w(\beta, r, c_0 Z^2)$ with $G(\cdot \mid Z)$ probability one.*

Proposition 4.2 – *Let $M > 0$ be large enough, $K \geq 0$ integer, and $Z = KM$. Assume that there is a constant $c_0 > 0$ and a sequence $(d_k)_{k=1}^K$ such that $\sum_{k=1}^K d_k \leq c_0 K$, and for any sequence $(e_{lm})_{(l,m)\in\mathcal{S}_k}$ it holds $F_k(\sum_{(l,m)\in\mathcal{I}_k} |\eta_{lm} - e_{lm}|^2 \leq t) \gtrsim \exp(-d_k K^{b_1-r-1}\log t^{-1})$. Then there is a constant $a_3 > 0$ such that $G(\sum_{(l,m)\in\Lambda_Z} |p_{lm} - q_{lm}|^2 \leq 12t \mid Z) \gtrsim \exp(-a_3 Z^{b_1-r}\log t^{-1})$.*

In the previous proposition, some conditions are required on $F_1, F_2, \dots$; these conditions are indeed really mild. For instance, it follows from Ghosal et al. (2000, lemma 6.1) that the conclusion of proposition 4.2 is valid if $\eta_{lm} := \sqrt{u_{lm}}$ where $(u_{lm})_{(l,m)\in\mathcal{I}_k}$ are drawn from Dirichlet distributions with suitable parameters.

4.7. Proof of theorem 4.2

The proof of theorem 4.2 follows the classical approach of Ghosal et al. (2000); Ghosal and van der Vaart (2007) for which the prior mass of Kullback-Leibler type neighborhoods need to be bounded from below and tests constructed. See details in section 4.E.

Throughout the document, we let $D_n^{\beta,r} := (\log(n)/\beta)^{1/r}$. Then we introduce the following events, which we'll use several times in the proof of posterior contraction rates.

$$E_n := \Big\{(y,\theta) \in \mathbb{R} \times [0,2\pi] \,:\, |y| \leq D_n^{\beta,r}\Big\}, \tag{4.10}$$

$$\Omega_n := \{((y_1,\theta_1),\dots,(y_n,\theta_n)) \,:\, (y_i,\theta_i) \in E_n \quad \forall i = 1,\dots,n\}. \tag{4.11}$$

4.7.1. Prior mass of Kullback-Leibler neighborhoods

We introduce a new variation around the basic lines of Ghosal et al. (2000); Ghosal and van der Vaart (2007), permitting to slightly weaken the so-called *Kullback-Leibler* (KL) condition. We show that we can trade the KL condition for a restricted KL condition; that is prior positivity of the sets

$$B_n(\delta_n) := \left\{\psi \,:\, \int_{E_n} p_{\psi_0}^\eta \log \frac{p_{\psi_0}^\eta}{p_\psi^\eta} \leq \delta_n^2, \quad \int_{E_n} p_{\psi_0}^\eta \left(\log \frac{p_{\psi_0}^\eta}{p_\psi^\eta}\right)^2 \leq \delta_n^2\right\}. \tag{4.12}$$

Although looking trivial, this will ease the proof of our main theorem, since the prior positivity of $B_n(\delta_n)$ is simpler to prove than the classical positivity of KL balls of Ghosal et al. (2000); Ghosal and van der Vaart (2007).

Decay estimates of the true density

It is a classical fact that in Bayesian nonparametrics we often require tails assumptions on the density of observations to be able to state rates of convergence. Here, the density of observations is quite complicated, as being the convolution of a Gaussian noise with the Radon-Wigner transform of ψ. Since the Wigner transform of ψ interpolates ψ and its Fourier transform, we definitively have to take care about fancy tails assumptions on the density that could be non compatible with the requirements of a Wigner transform. Instead, we show that the decay assumptions on the STFT stated in the definition of $\mathcal{C}_g(\beta,r,L)$ directly translate onto the tails of the joint density of observations. We have the following theorem, whose proof is given in section 4.B.1.

Lemma 4.1 – *For all $\beta, L > 0$ and all $r \in (0,1)$ there is a constant $C(\beta,r,\eta) > 0$ such such that $P_\psi^\eta(E_n^c) \leq 2\pi C(\beta,r,\eta) L^2 n^{-2}$ and $P_\psi^{\eta,n}(\Omega_n^c) \leq 2\pi C(\beta,r,\eta) L^2 n^{-1}$ for all $\psi \in \mathcal{C}_g(\beta,r,L)$.*

Approximation theory

In order to prove the prior positivity of the sets $B_n(\delta_n)$, we need to construct a family $\mathcal{M}_n$ of functions in $\mathbb{S}^2(\mathbb{R})$ that approximate well ψ_0 in the $L^2(\mathbb{R})$ distance. We will show later that the sets $B_n(\delta_n)$ contains suitable closed balls around ψ_0 in the norm of $L^2(\mathbb{R})$.

In the sequel, we need to relate the parameters β, r, L to the decay of the coefficients $\langle \psi_0, \varphi_{lm}\rangle$ of $\psi_0 \in \mathcal{C}_g(\beta,r,L)$ expressed in the Wilson base. Fortunately, Wilson bases are unconditional bases for the ultra-modulation spaces, and $\mathcal{C}_g(\beta,r,L)$ is a subset of the ultra-modulation space $M^1_{\beta,r}$. It follows the following lemma (Gröchenig, 2001, theorem 12.3.1).

Lemma 4.2 – *Let $\psi \in \mathcal{C}_g(\beta, r, L)$ for some $\beta, L > 0$ and $0 \leq r < 1$. Then there is a constant $0 < C(\beta, r) < +\infty$ such that*

$$\sum_{(l,m)\in\Lambda_\infty} |\langle \psi,\, \varphi_{lm}\rangle| \exp\left(\beta(l^2+m^2)^{r/2}\right) \leq C(\beta, r)L.$$

Having characterized the decay of Gabor coefficients for those $\psi \in \mathcal{C}_g(\beta, r, L)$, we are now in position to construct functions ψ_Z which degree of approximation to $\psi_0 \in \mathcal{C}_g(\beta, r, L)$ is indexed by the value of Z. In view of section 4.2.3, ψ_0 has the formal decomposition $\psi_0 = \sum_{l,m}\langle \psi_0,\, \varphi_{lm}\rangle\, \varphi_{lm}$, with unconditional convergence of the series in $L^2(\mathbb{R})$. We define $\widetilde{\psi}_Z$ such that

$$\widetilde{\psi}_Z := \sum_{(l,m)\in\Lambda_Z} \langle \psi_0,\, \varphi_{lm}\rangle\, \varphi_{lm}.$$

Since (φ_{lm}) constitutes an orthonormal base for $L^2(\mathbb{R})$, lemma 4.2 implies that for any $\beta > 0$ and $r \in (0, 1)$,

$$\begin{aligned}
\|\psi_0 - \widetilde{\psi}_Z\|_2^2 &= \sum_{(l,m)\notin\Lambda_Z} |\langle \psi_0,\, \varphi_{lm}\rangle|^2 \\
&\leq \exp(-\beta Z^r) \sum_{l,m} |\langle \psi_0,\, \varphi_{lm}\rangle| \exp\left(\beta(l^2+m^2)^{r/2}\right) \\
&\leq C(\beta, r)L \exp\left(-\beta Z^r\right),
\end{aligned}$$

because on Λ_Z^{c} we have $l^2 + m^2 \geq Z^2$ and $|\langle \psi_0,\, \varphi_{lm}\rangle| \leq \|\psi_0\|_2 \|\varphi_{lm}\|_2 = 1$. Note that $\widetilde{\psi}_Z$ is not necessarily in $\mathbb{S}^2(\mathbb{R})$, that is in general $\|\widetilde{\psi}_Z\|_2 \neq 1$, whence it is not a proper wave-function. We now trade $\widetilde{\psi}_Z$ for a version ψ_Z with $\|\psi_Z\|_2 = 1$, keeping the same order of approximation. Indeed, let $\psi_Z := \widetilde{\psi}_Z / \|\widetilde{\psi}_Z\|_2$, then since $\|\psi_0\|_2 = 1$,

$$\begin{aligned}
\|\psi_Z - \psi_0\|_2 &\leq \|\psi_Z - \widetilde{\psi}_Z\|_2 + \|\widetilde{\psi}_Z - \psi_0\|_2 \\
&\leq \|\widetilde{\psi}_Z\| \left|1 - \frac{1}{\|\widetilde{\psi}_Z\|_2}\right| + \|\widetilde{\psi}_Z - \psi_0\|_2 \leq 2\|\widetilde{\psi}_Z - \psi_0\|_2 \\
&\leq 2\sqrt{C(\beta, r)L} \exp\left(-\frac{\beta Z^r}{2}\right). \qquad (4.13)
\end{aligned}$$

A lower bound on $\Pi(B_n(\delta_n))$

The proof of the lemmas and theorem of this section are to be found in sections 4.B.2 and 4.B.3. To prove the Kullback-Leibler condition, we first construct a suitable set $\mathcal{M}_n \subset B_n(\delta_n)$, and we'll lower bound $\Pi(B_n(\delta_n)) \geq \Pi(\mathcal{M}_n)$. Let ψ_Z be the function constructed in section 4.7.1 and $c_{lm} := \langle \psi_Z,\, \varphi_{lm}\rangle$, so that $\psi_Z = \sum_{(l,m)\in\Lambda_Z} c_{lm}\varphi_{lm}$. Then, we define the set $\mathcal{M}_n \equiv \mathcal{M}_n(Z, U)$ as follows, and we'll prove that Z, U can be chosen so that $\mathcal{M}_n(Z, U) \subset B_n(\delta_n)$.

$$\mathcal{M}_n(Z,U) := \left\{ \psi \in \mathbb{S}^2(\mathbb{R}) \;:\; \begin{array}{l} \psi = \sum_{(l,m)\in\Lambda_Z} p_{lm} e^{i\zeta_{lm}} \varphi_{lm}, \\ \sum_{(l,m)\in\Lambda_Z} |p_{lm} - |c_{lm}||^2 \leq U^2 \\ \sum_{(l,m)\in\Lambda_Z} |\zeta_{lm} - \arg c_{lm}|^2 \leq U^2 \end{array} \right\}. \qquad (4.14)$$

Lemma 4.3 – *For all $\psi \in \mathcal{M}_n(Z,U)$, it holds with the constant $C(\beta,r)$ of lemma 4.2,*

$$\|\psi - \psi_0\|_2 \leq 2U + 2\sqrt{C(\beta,r,g)L}\exp\left(-\frac{\beta Z^r}{2}\right).$$

The fact that $\mathcal{M}_n(Z,U)$ is included into a suitable $L^2(\mathbb{R})$ ball around ψ_0 is not enough to prove the inclusion $\mathcal{M}_n(Z,U) \subset B_n(\delta_n)$. The next lemma states sufficient conditions for which the inclusion $\mathcal{M}_n(Z,U) \subset B_n(\delta_n)$ actually holds true.

Lemma 4.4 – *There are constants $0 < C_1, C_2 < \infty$ depending only on $\gamma, \beta, r, A, B, L$ such that if $U \leq C_1(\log n)^{-4/r}\delta_n^2$ and $Z \geq C_2(\log \delta_n^{-1})^{1/r}$, then for n large enough $\mathcal{M}_n(Z,U) \subset B_n(\delta_n)$ for every $\delta_n^2 \geq 4\sqrt{2\pi C(\beta,r,\eta)}Ln^{-1}$, where $C(\beta,r,\eta)$ is the constant of lemma 4.1.*

Now that we have shown that $\mathcal{M}_n(Z,U) \subseteq B_n(\delta_n)$ for suitable choice of Z and U, it is clear that the prior mass of $B_n(\delta_n)$ is lower bounded by the prior mass of $\mathcal{M}_n(Z,U)$, the one is relatively easy to compute. This statement is made formal in the next theorem.

Theorem 4.3 – *Let $\psi_0 \in \mathcal{C}_g(\beta,r,L)$, and $b_1 > 2 + r$. Then there is a constant $C > 0$ such that for $n\delta_n^2 = C(\log n)^{b_1/r}$ it holds $\Pi(B_n(\delta_n)) \gtrsim \exp(-n\delta_n^2)$ for n large enough.*

4.7.2. Construction of tests

The approach for constructing tests is reminiscent to Knapik and Salomond (2014), where authors provide a general setup to establish posterior contraction rates in nonparametric inverse problems. We define the following sieve. For positive constants c, h to be determined later, and the constant $a_4 > 0$ of the assumptions

$$\mathcal{F}_n := \left\{ \psi \in \mathbb{S}^2(\mathbb{R}) \; : \; \begin{array}{l} \psi = \sum_{(l,m)\in\Lambda_Z} p_{lm} e^{i\zeta_{lm}} \varphi_{lm}, \quad 0 \leq Z \leq h(\log n)^{1/r}, \\ \mathbf{p} \in \Delta_Z^w(\beta, r, c(\log n)^{a_4}) \end{array} \right\}.$$

Then, we construct test functions with rapidly decreasing type I and type II errors, for testing the hypothesis $H_0 : \psi = \psi_0$ against the alternative $H_1 : \psi \in U_n \cap \mathcal{F}_n$, with $U_n := \{\psi \in \mathbb{S}^2(\mathbb{R}) \; : \; \|\psi - \psi_0\|_2 \geq \epsilon_n\}$, for a sequence $(\epsilon_n)_{n\geq 0}$ to be determined later. To this aim, we need the following series of propositions about $\mathcal{F}_n$, which are proved in section 4.C.1.

Proposition 4.3 – *Let $n\delta_n^2 = C(\log n)^{b_1/r}$ for some constant $C > 0$. Then $\Pi(\mathcal{F}_n^c) \lesssim \exp(-6n\delta_n^2)$ whenever $h > (6C/a_2)^{1/b_1}$ and $c > 0$ large enough.*

Proposition 4.4 – *Let $b_1 > 2 + r$ and assume that $n\delta_n^2 = C(\log n)^{b_1/r}$ for some constant $C > 0$. Then $N(\sqrt{2}\delta_n^2, \mathcal{F}_n, \|\cdot\|_2)\exp(-6n\delta_n^2) = o(1)$.*

Proposition 4.5 – *There is a constant $M > 0$, depending only on φ and η, such that for all $\psi \in \mathcal{F}_n$ it holds $\|p_\psi^\eta\|_\infty \leq Mh^2(\log n)^{2/r}$.*

Proposition 4.6 – *For all $\beta > 0$ and $r \in (0,1)$ there is a constant $R > 0$ such that for any $u > 0$ it holds $\sup_{\psi\in\mathcal{F}_n} \int_{\{\|z\|>u\}} |\widehat{W}_\psi(z)|^2 \, dz \leq R(\log n)^{2a_4}\exp(-\beta u^r)$.*

The first step in the tests construction consists on bounding, both from below and from above, the Hellinger distance $H^2(P_\psi^\eta, P_{\psi_0}^\eta)$ by a multiple constant of $\|\psi - \psi_0\|_2$, at least for those $\psi_0 \in \mathcal{C}_g(\beta,r,L)$ and those $\psi \in \mathcal{F}_n$. To this aim, we need to estimate the decay of $\widehat{W}_{\psi_0}$, stated in the next proposition. The remaining proofs for this section can be found in sections 4.C.2 and 4.C.3.

Proposition 4.7 – *Let $\psi \in \mathcal{C}_g(\beta, r, L)$ for some $\beta, L > 0$ and $r \in (0,1)$. Then*

$$\int_{\mathbb{R}^2} |\widehat{W}_\psi(z)|^2 \exp(\beta \|z\|^r)\, dz \leq L^2.$$

The practical proposition 4.7 allows to upper bound $\|\psi - \psi_0\|_2$ by $H(P^\eta_\psi, P^\eta_{\psi_0})$, provided ψ and ψ_0 are sufficiently separated from each other.

Lemma 4.5 – *Let $\beta, L > 0$, $r \in (0,1)$, $C_0 := \|p^\eta_{\psi_0}\|_\infty$, $M, R > 0$ be the constants of propositions 4.5 and 4.6, and assume n large enough. Then for all $u > 0$, all $\psi \in \mathcal{F}_n$ and all $\psi_0 \in \mathcal{C}_g(\beta, r, L)$ such that $\|\psi - \psi_0\|_2^2 \geq 8R(\log n)^{2a4} \exp(-\beta u^r)$, it holds*

$$\sqrt{2} H^2(P^\eta_\psi, P^\eta_{\psi_0}) \leq \|\psi - \psi_0\|_2 \leq 2\sqrt{C_0 + Mh^2(\log n)^{2/r}} e^{\gamma u^2} H(P^\eta_\psi, P^\eta_{\psi_0}).$$

From the last lemma, we are in position to construct test functions with rapidly decreasing type I and type II error for testing $H_0 : \psi = \psi_0 \in \mathcal{C}_g(\beta, r, L)$ against $H_1 : \|\psi - \psi_1\|_2 \leq \sqrt{2}\delta_n^2$ for any $\psi_1 \in \mathcal{F}_n$ such that $\|\psi_1 - \psi_0\|_2 \geq \epsilon_n^2$, with

$$\delta_n^2 := \frac{\epsilon_n^2 \exp(-2\gamma u_n^2)}{48[C_0 + Mh^2(\log n)^{2/r})]}, \qquad \epsilon_n^2 := 8R(\log n)^{2a4} \exp(-\beta u_n^r), \tag{4.15}$$

where $(u_n)_{n \geq 0}$ is an increasing sequence of positive numbers to be determined later and $M, R > 0$ the constants of propositions 4.5 and 4.6.

Proposition 4.8 – *Let δ_n, ϵ_n be as in equation (4.15). Then there exist test functions $(\phi_n)_{n \geq 0}$ for testing $H_0 : \psi = \psi_0 \in \mathcal{C}_g(\beta, r, L)$ against $H_1 : \|\psi - \psi_1\|_2 \leq \sqrt{2}\delta_n^2$ for any $\psi_1 \in \mathcal{F}_n$ such that $\|\psi_1 - \psi_0\|_2 \geq \epsilon_n$, with type I and type II errors satisfying*

$$P^{\eta,n}_{\psi_0} \phi_n \leq \exp(-6n\delta_n^2), \qquad \sup_{\psi \in \mathbb{S}^2 : \|\psi - \psi_1\|_2 \leq \sqrt{2}\delta_n^2} P^{\eta,n}_\psi (1 - \phi_n) \leq \exp(-6n\delta_n^2).$$

Proof. By lemma 4.5, we deduce that $H(P^\eta_{\psi_1}, P^\eta_{\psi_0}) \geq \sqrt{12}\delta_n$. From lemma 4.11, for any $\psi \in \mathbb{S}^2(\mathbb{R})$ with $\|\psi - \psi_1\|_2 \leq \sqrt{2}\delta_n^2$ (ψ not necessarily in $\mathcal{F}_n$), we have the estimate $H(P^\eta_\psi, P^\eta_{\psi_1}) \leq \delta_n \leq H(P^\eta_{\psi_1}, P^\eta_{\psi_0})/2$. Then the conclusion follows from Ghosal et al. (2000, section 7). ■

The small balls estimate of proposition 4.8 allows to build the desired test functions, using the classical approach of the covering of $\mathcal{F}_n$ with balls of radius $\sqrt{2}\delta_n^2$ in the $L^2(\mathbb{R})$ norm (Ghosal et al., 2000).

Theorem 4.4 – *Assume that $\psi_0 \in \mathcal{C}_g(\beta, r, L)$ for $\beta, L > 0$ and $r \in (0,1)$, and let ϵ_n, δ_n be as in equation (4.15). Let $N(\sqrt{2}\delta_n^2, \mathcal{F}_n, \|\cdot\|_2)$ be the number of $L^2(\mathbb{R})$ balls of radius $\sqrt{2}\delta_n^2$ needed to cover $\mathcal{F}_n$. Then there exist test functions $(\phi_n)_{n \geq 0}$ such that*

$$P^{\eta,n}_{\psi_0} \phi_n \leq N(\sqrt{2}\delta_n^2, \mathcal{F}_n, \|\cdot\|_2) \exp(-6n\delta_n^2), \text{ and}$$
$$\sup_{\psi \in \mathcal{F}_n : \|\psi - \psi_0\|_2 \geq \epsilon_n} P^{\eta,n}_\psi (1 - \phi_n) \leq \exp(-6n\delta_n^2).$$

4.7.3. Conclusion of the proof

Let summarize what we've done so far, and finalize the proof of theorem 4.2. In lemma 4.10 in appendix, we state sufficient conditions to finish the proof of our main theorem; these conditions involve two parts. First, proving that for a suitable sequence $\delta_n \to 0$ with $n\delta_n^2 \to$ our prior puts enough probability mass on the balls $B_n(\delta_n)$ and; the construction of tests functions with sufficiently rapidly decreasing type I and type II errors for testing $H_0 : \psi = \psi_0$ against $H_1 : \|\psi - \psi_0\|_2 \geq \epsilon_n$, for those ψ in a set $\mathcal{F}_n$ of prior probability $1 - \exp(-6n\delta_n^2)$.

For the prior considered here, we found in theorem 4.3 that δ_n must satisfy $n\delta_n^2 \geq C(\log n)^{b_1/r}$ for some $C > 0$, otherwise the so-called Kullback-Leilbler condition is not met. Regarding the construction of tests, this involved to build explicitly the sets $\mathcal{F}_n$ in section 4.7.2. From that construction and equation (4.15), we deduce that the required test functions exist, if for some constants $K_1, K_2 > 0$ and a sequence $u_n \to \infty$

$$\delta_n^2 \leq \frac{K_1 \exp(-2\gamma u_n^2)\epsilon_n^2}{(\log n)^{2/r}}, \quad \epsilon_n^2 \geq K_2(\log n)^{2a_4} \exp(-\beta u_n^r). \tag{4.16}$$

Since we must also have $n\delta_n^2 \geq C(\log n)^{b_1/r}$, we deduce that the sequence $(u_n)_{n\geq 1}$ should satisfy, for a suitable constant $C' > 0$,

$$\beta u_n^r + 2\gamma u_n^2 - 2a_5(\log n)^{s/2} \leq \log C' + \log n - r^{-1}(2 + b_1 - 2ra_4)\log\log n.$$

Finally, we can take,

$$u_n^2 = \frac{\log n}{2\gamma} - O((\log n)^{r/2})$$

and the conclusion of the proof follows by equation (4.16).

Appendix

4.A. Proof of theorem 4.1

We need some subsidiaries results to prove the theorem 4.1.

Proposition 4.9 – *For all $\beta > 0$, all $0 \leq r \leq 1$ and all $x, y \in \mathbb{R}^2$, it holds* $\exp(\beta\|x+y\|^r) \leq \exp(\beta\|x\|^r)\exp(\beta\|y\|^r)$.

Proof. This follows from the trivial estimate

$$\begin{aligned}\|x+y\|^r \leq (\|x\|+\|y\|)^r &= \|x\|(\|x\|+\|y\|)^{r-1} + \|y\|(\|x\|+\|y\|)^{r-1} \\ &\leq \|x\|\|x\|^{r-1} + \|y\|\|y\|^{r-1} = \|x\|^r + \|y\|^r.\end{aligned}$$ ■

The next lemma is about the change of window in the STFT; its proof is given for arbitrary $g \in \mathcal{S}(\mathbb{R})$ and $\psi \in \mathcal{S}'(\mathbb{R})$ in Gröchenig (2001, lemma 11.3.3). The proof is identical when $g, \psi \in L^2(\mathbb{R})$, since it essentially rely on a duality argument. Note, however, that the class of windows and functions that we are considering are subset of $\mathcal{S}(\mathbb{R})$.

Lemma 4.6 – *Let $g_0, g, h \in L^2(\mathbb{R})$ such that $\langle h, g\rangle \neq 0$ and let $\psi \in L^2(\mathbb{R})$. Then $|V_{g_0}\psi(x,\omega)| \leq |\langle h, g\rangle|^{-1}(|V_g\psi| * |V_{g_0}h|)(x,\omega)$ for all $(x,\omega) \in \mathbb{R}^d$.*

Proof. From Gröchenig (2001, corollary 3.2.3), for those $g, h \in L^2(\mathbb{R})$ with $\langle h, g\rangle \neq 0$, we have the *inversion formula* $\psi = \langle h, g\rangle^{-1} \int V_g\psi(x,\omega)\, M_\omega T_x h \, d\omega dx$ for all $\psi \in L^2$. Applying V_{g_0} both sides

$$V_{g_0}\psi(x',\omega') = \frac{1}{\langle h, g\rangle}\int_{\mathbb{R}^2} V_g\psi(x,\omega) V_{g_0}(M_\omega T_x h)(x',\omega')\, d\omega dx.$$

The conclusion follows because $|V_{g_0}(M_\omega T_x h)(x',\omega')| = |V_{g_0}h(x'-x,\omega'-\omega)|$. ■

Finally, we have the sufficient material to establish the independence of the class $\mathcal{C}_g(\beta, r, L)$ with respect to the choice of the window function g, as soon as g is suitably well behaved.

Proof of theorem 4.1. Using lemma 4.6, we have that $|V_{g_0}\psi| \leq \|g\|_2^{-2}|V_g\psi| * |V_{g_0}g|$. Then,

because $r < 1$ by assumption,

$$\begin{aligned}\int_{\mathbb{R}^2} |V_{g_0}\psi(z)| \exp(\beta\|z\|^r)\, dz \\ &\leq \int_{\mathbb{R}^2} (|V_g\psi(z)| * |V_{g_0}g(z)| \exp(\beta\|z\|^r)\, dz \\ &\leq \iint_{\mathbb{R}^2}\int_{\mathbb{R}^2} |V_g\psi(u)\exp(\beta\|u\|^r)|V_{g_0}g(z-u)| \exp(\beta\|z-u\|^r)\, dudz \\ &\leq \int_{\mathbb{R}^2} |V_g\psi(u)| \exp(\beta\|u\|^r)\, du \int_{\mathbb{R}^2} |V_{g_0}g(u)| \exp(\beta\|u\|^r)\, du,\end{aligned}$$

where we've used Young's inequality and the first estimate of proposition 4.9. We have by Gröchenig and Zimmermann (2004, corollary 3.10) that $V_{g_0}g \in \mathcal{S}_1^1(\mathbb{R}^2)$, thus the second integral in the rhs of the last equation is bounded. ■

4.B. Proofs of Kullback-Leibler neighborhoods prior mass

4.B.1. Proof of lemma 4.1

To prove lemma 4.1, we need the following intermediate lemmas, relating the smoothness of ψ to the tails of the Wigner density of ψ.

Lemma 4.7 – *Let $\psi \in \mathcal{C}_g(\beta, r, L)$ with $\beta, L > 0$ and $r \in (0,1)$. Then,*

$$\int_{\mathbb{R}^2} |W_\psi(z)| \exp(\beta\|2z\|^r)dz \leq L^2.$$

Proof. Let $\breve{\psi}(x) = \psi(-x)$. Then from the definition of $V_g\psi$ and W_ψ we have that $W_\psi(x,\omega) = 2e^{4\pi i\omega x}V_{\breve{\psi}}\psi(2x, 2\omega)$. By lemma 4.6 (with $|\langle g, g\rangle| = \|g\|_2^2 = 1$), proposition 4.9, and Young's inequality,

$$\begin{aligned}\int |W_\psi(z/2)| \exp(\beta\|z\|^r)dz \\ &\leq 2\int (|V_g\psi| * |V_{\breve{\psi}}g|)(z) \exp(\beta\|z\|^r)dz \\ &\leq 2\iint |V_g\psi(u)| \exp(\beta\|u\|^r)|V_{\breve{\psi}}g(z-u)| \exp(\beta\|z-u\|^r)\, dudz \\ &\leq 2\int |V_g\psi(z)| \exp(\beta\|z\|^r)dz \times \int |V_{\breve{\psi}}g(z)| \exp(\beta\|z\|^r)dz.\end{aligned}$$

Moreover, a straightforward computation shows that

$$V_{\breve{\psi}}g(x,\omega) = e^{-2\pi i\omega x}\overline{V_g\psi(x,-\omega)},$$

which concludes the proof. ■

Lemma 4.8 – *Let $\psi \in \mathcal{C}_g(\beta, r, L)$, with $\beta, L > 0$ and $r \in (0,1)$. Then,*

$$\sup_\theta \int_{\mathbb{R}} p_\psi(x,\theta) \exp(2\beta|x|^r)dx \leq L^2.$$

Proof. From the definition of p_ψ,

$$\int_{\mathbb{R}} p_\psi(x,\theta)\, e^{2\beta|x|^r} dx = \int_{\mathbb{R}^2} W_\psi(x\cos\theta - \xi\sin\theta, x\sin\theta + \xi\cos\theta)\, e^{2\beta|x|^r} d\xi dx.$$

Performing the change of variable $(x,\xi) \mapsto (x\cos\theta + \xi\sin\theta, -x\sin\theta + \xi\cos\theta)$, we arrive at

$$\int_{\mathbb{R}} p_\psi(x,\theta)\, e^{2\beta|x|^r} dx = \int_{\mathbb{R}^2} W_\psi(x,\xi)\, e^{2\beta|x\cos\theta+\xi\sin\theta|^r} d\xi dx.$$

But for all $r \in (0,1)$, by the triangle inequality and Hölder's inequality

$$|x\cos\theta + \xi\sin\theta|^r \le (|x\cos\theta| + |\xi\sin\theta|)^r \le (|x| + |\xi|)^r \le 2^{r/2}(x^2+\xi^2)^{r/2}.$$

Then

$$\int_{\mathbb{R}} p_\psi(x,\theta)\, e^{2\beta|x|^r} dx \le \int_{\mathbb{R}^2} |W_\psi(z)| \exp\left(\beta\|2z\|^r\right)\, dz,$$

and the conclusion follows from lemma 4.7. ■

Lemma 4.9 – *For all $\beta, L > 0$ and $r \in (0,1)$ there is a constant $C(\beta,r,\eta) > 0$ such that if $\psi \in \mathcal{C}_g(\beta,r,L)$ we have $\sup_\theta \int_{\mathbb{R}} p_\psi^\eta(y,\theta)\, \exp(2\beta|y|^r) dy \le C(\beta,r,\eta)L^2$.*

Proof. Using Fubini's theorem twice and the estimate $|u+x|^r \le |u|^r + |x|^r$,

$$\begin{aligned}
\int p_\psi^\eta(y,\theta)\, e^{2\beta|y|^r} dy &= \sqrt{\frac{\pi}{\gamma}} \iint p_\psi(x,\theta) \exp\left\{-\frac{\pi^2(x-y)^2}{\gamma}\right\} dx\, e^{2\beta|y|^r} dy \\
&= \sqrt{\frac{\pi}{\gamma}} \iint p_\psi(x,\theta) \exp\left\{-\frac{\pi^2 u^2}{\gamma}\right\} \exp\left(2\beta|u+x|^r\right)\, du dx \\
&\le \sqrt{\frac{\pi}{\gamma}} \int p_\psi(x,\theta)\, e^{2\beta|x|^r} dx \int \exp\left\{-\frac{\pi^2 u^2}{\gamma} + 2\beta|u|^r\right\} du.
\end{aligned}$$

The conclusion follows from lemma 4.8. ■

From the lemmas above the proof of lemma 4.1 is relatively straightforward, we give it here for the sake of completeness.

Proof of lemma 4.1. We begin with the obvious estimate $P_\psi^{\eta,n}(\Omega_n^c) \le nP_\psi^\eta(E_n^c)$. The proof is finished by noticing that

$$\begin{aligned}
P_\psi^\eta(E_n^c) &= \int_{E_n^c} p_\psi^\eta(y,\theta)\, e^{2\beta|y|^r} e^{-2\beta|y|^r} dy d\theta \\
&\le n^{-2} \int p_\psi^\eta(y,\theta)\, e^{2\beta|y|^r} dy d\theta \\
&\le 2\pi C(\beta,r,\eta) L^2 n^{-2},
\end{aligned}$$

because of lemma 4.9. ■

4.B.2. Proofs regarding approximation theory

Proof of lemma 4.3. For all $\psi \in \mathcal{M}_n(Z,U)$ we have the following estimate. Because (φ_{lm}) is an orthonormal base of $L^2(\mathbb{R})$,

$$\begin{aligned}\|\psi - \psi_Z\|_2^2 &= \sum_{(l,m)\in\Lambda_Z} |p_{lm}e^{i\zeta_{lm}} - c_{lm}|^2 \\ &\leq 2\sum_{(l,m)\in\Lambda_Z} |p_{lm} - |c_{lm}||^2 + 2\sum_{(l,m)\in\Lambda_Z} |\zeta_{lm} - \arg c_{lm}|^2 \leq 4U^2.\end{aligned}$$

Then the conclusion follows using $\|\psi-\psi_0\|_2 \leq \|\psi-\psi_Z\|_2+\|\psi_Z-\psi_0\|_2$ and equation (4.13). ■

Proof of lemma 4.4. Recall that $p_\psi^\eta(y,\theta) = [p_\psi(\cdot,\theta) * G_\gamma](y)$. We have the obvious bound

$$p_\psi^\eta(y,\theta) = \int_{-\infty}^{+\infty} p_\psi(x,\theta)G_\gamma(y-x)\,dx \geq \int_{-D_n^{\beta,r}}^{+D_n^{\beta,r}} p_\psi(x,\theta)G_\gamma(y-x)\,dx.$$

Then for all $(y,\theta) \in E_n$ (*i.e.* $|y| \leq D_n^{\beta,r}$) it follows from the definition of G_γ that $p_\psi^\eta(y,\theta) \geq G_\gamma(2D_n^{\beta,r})P_\psi(|X| \leq D_n^{\beta,r} \mid \theta)/(2\pi)$. From proposition 4.10 in appendix, the latter implies for n large enough that for all $\psi \in \mathcal{M}_n$ it holds $p_\psi^\eta(y,\theta) \geq G_\gamma(2D_n^{\beta,r})/(4\pi)$ whenever $(y,\theta) \in E_n$. Since $\psi_0 \in \mathcal{C}_g(\beta,r,L)$, which is a subset of the Schwartz space $\mathcal{S}(\mathbb{R})$, and since the Radon transform maps $\mathcal{S}(\mathbb{R})$ onto a subset of $\mathcal{S}(\mathbb{R}\times[0,2\pi])$ by Helgason (2011, theorem 2.4), we deduce that there is a constant $C = C(\psi_0,\eta) > 0$ such that for all $\psi \in \mathcal{M}_n(Z,U)$,

$$\frac{p_{\psi_0}^\eta(y,\theta)}{p_\psi^\eta(y,\theta)} \leq C\exp\left\{\frac{4\pi^2}{\gamma}\left(\frac{\log n}{\beta}\right)^{2/r}\right\} =: \lambda_n^{-1}, \quad \forall (y,\theta) \in E_n.$$

The proof now follows similar lines as Shen et al. (2013, lemma B2). The function $r : (0,\infty) \to \mathbb{R}$ defined implicitly by $\log x = 2(x^{1/2}-1) - r(x)(x^{1/2}-1)^2$ is nonnegative and decreasing. Thus we obtain,

$$\begin{aligned}&\int_{E_n} p_{\psi_0}^\eta \log\frac{p_{\psi_0}^\eta}{p_\psi^\eta} \\ &\quad= -2\int_{E_n} p_{\psi_0}^\eta\left(\sqrt{\frac{p_\psi^\eta}{p_{\psi_0}^\eta}} - 1\right) + \int_{E_n} p_{\psi_0}^\eta r\left(\frac{p_\psi^\eta}{p_{\psi_0}^\eta}\right)\left(\sqrt{\frac{p_\psi^\eta}{p_{\psi_0}^\eta}} - 1\right)^2 \\ &\quad\leq 2\left(1 - \int\sqrt{p_{\psi_0}^\eta p_\psi^\eta}\right) - 2P_{\psi_0}^\eta(E_n^c) \\ &\qquad\qquad + 2\int_{E_n^c}\sqrt{p_{\psi_0}^\eta p_\psi^\eta} + r(\lambda_n)\int_{E_n}\left(\sqrt{p_\psi^\eta} - \sqrt{p_{\psi_0}^\eta}\right)^2 \\ &\quad\leq 2H^2(P_\psi^\eta, P_{\psi_0}^\eta)\,(1 + r(\lambda_n)) + 2P_{\psi_0}^\eta(E_n^c)^{1/2}P_\psi^\eta(E_n^c)^{1/2},\end{aligned} \tag{4.17}$$

where the last line follows from Hölder's inequality. Also, proceeding as in the proof of Shen et al. (2013, lemma B2) we find that

$$\int_{E_n} p_{\psi_0}^\eta\left(\log\frac{p_{\psi_0}^\eta}{p_\psi^\eta}\right)^2 \leq H^2(P_\psi^\eta, P_{\psi_0}^\eta)\left(12 + 2r(\lambda_n)^2\right). \tag{4.18}$$

Note that $r(x) \leq \log x^{-1}$ for x small enough, and by lemma 4.1,

$$P_{\psi_0}^{\eta}(E_n^c)^{1/2} P_{\psi}^{\eta}(E_n^c)^{1/2} \leq P_{\psi_0}^{\eta}(E_n^c)^{1/2} \leq \sqrt{2\pi C(\beta, r, \eta)} L n^{-1}. \tag{4.19}$$

Then we deduce from equations (4.17) to (4.19) and lemma 4.11 that for n large enough, provided $\delta_n^2 \geq 4\sqrt{2\pi C(\beta, r, \eta)} L n^{-1}$,

$$B_n(\delta_n) \supset \left\{ P_{\psi}^{\eta} \,:\, \psi \in \mathbb{S}^2, \quad \|\psi - \psi_0\|_2 \leq \frac{\gamma^2}{48\sqrt{2}\pi^4} \left(\frac{\beta}{\log n}\right)^{4/r} \delta_n^2 \right\}.$$

Then the conclusion follows from lemma 4.3. ■

4.B.3. Proof of the lower bound

Proof of theorem 4.3. Let $C_1, C_2 > 0$ be the constants of lemma 4.4, and let $U_n = C_1 (\log n)^{-4/r} \delta_n^2$ and Z_n be the smaller integer larger than $C_2 (\log \delta_n^{-1})^{1/r}$. Then by lemma 4.4 $\Pi(B_n(\delta_n)) \geq \Pi(\mathcal{M}_n(Z_n, U_n))$, and

$$\begin{aligned}\Pi(\mathcal{M}_n(Z_n, U_n)) \geq P_Z(Z = Z_n) G\left(\textstyle\sum_{(l,m)\in\Lambda_Z} |p_{lm} - |c_{lm}||^2 \leq U_n^2 \mid Z\right) \\ \times P_\zeta\left(\textstyle\sum_{(l,m)\in\Lambda_Z} |\zeta_{lm} - \arg c_{lm}|^2 \leq U^2 \mid Z\right).\end{aligned}$$

Note that by lemma 4.2 the sequence $(|c_{lm}|)_{(l,m)\in\lambda_Z}$ is in $\Delta_Z^w(\beta, r, C(\beta, r)L)$. Hence, using the assumptions of section 4.5.2, we have for n large enough

$$\Pi(\mathcal{M}_n(Z_n, U_n)) \gtrsim \exp\left\{-a_1 Z_n^{b_1} - (a_0 + a_3) Z_n^{b_1 - r} \log U_n^{-2}\right\}.$$

We deduce from the above the existence of a constant $K > 0$ not depending on n, such that for n large enough,

$$\begin{aligned}\Pi(B_n(\delta_n)) &\gtrsim \exp\left\{-K(\log \delta_n^{-1})^{b_1/r} - K(\log \delta_n^{-1})^{b_1/r - 1}\left(\log \delta_n^{-1} + \log\log n\right)\right\} \\ &\gtrsim \exp(-n\delta_n^2).\end{aligned}$$

Then the conclusion of the theorem follows since we assume $n\delta_n^2 = C(\log n)^{b_1/r}$ for a suitable constant $C > 0$. ■

4.C. Proofs of tests construction

4.C.1. Proofs regarding the sieve

Proof of proposition 4.3. Let Z_n be the smaller integer larger than $h(\log n)^{1/r}$. Clearly $\psi \sim \Pi$ is almost-surely in $\mathbb{S}^2(\mathbb{R})$. Then if $c > 0$ is large enough we have the bound

$$\begin{aligned}\Pi(\mathcal{F}_n^c) &\leq P_Z\left(Z > h(\log n)^{1/r}\right) + G\left(\mathbf{p} \notin \Delta_Z^w(\beta, r, c(\log n)^{a_4}) \mid Z \leq Z_n\right) \\ &\lesssim \exp\left(-a_2 h^{b_1} (\log n)^{b_1/r}\right) + \exp\left(-a_5 (\log n)^{b_5}\right)\end{aligned}$$

which is trivially smaller than a multiple constant of $\exp(-6n\delta_n^2)$ when h is as large as in the proposition, and because $b_5 > b_1/r$ by assumption. ■

Proof of proposition 4.4. We use the argument that $N(\sqrt{2}\delta_n^2, \mathcal{F}_n, \|\cdot\|_2)$ is bounded by the cardinality of a $\sqrt{2}\delta_n^2$-net over $\mathcal{F}_n$ is the $\|\cdot\|_2$ distance (Shen et al., 2013). We compute the cardinality of such $\sqrt{2}\delta_n^2$-net as follows. Let $Z_n := h(\log n)^{1/r}$, $\widehat{\mathcal{P}}$ be a δ_n^2-net over the simplex Δ_{Z_n} in the ℓ_2 distance, and let $\widehat{\mathcal{O}}$ be a δ_n^2-net over $[0, 2\pi]$ in the euclidean distance. Then define

$$\mathcal{N}_n := \left\{ \psi \in \mathbb{S}^2(\mathbb{R}) \ : \ \begin{array}{l} \widetilde{\psi} = \sum_{(l,m)\in\Lambda_{Z_n}} v_{lm} e^{i\zeta_{lm}} \varphi_{lm}, \\ (v_{lm})_{(l,m)\in\Lambda_{Z_n}} \in \widehat{\mathcal{P}}, \quad \forall (l,m) \in \Lambda_{Z_n} \ : \ \zeta_{lm} \in \widehat{\mathcal{O}} \end{array} \right\}.$$

For all $\psi \in \mathcal{F}_n$ we have $\psi = \sum_{(l,m)\in\Lambda_{Z_n}} q_{lm} e^{i\zeta_{lm}} \varphi_{lm}$, with $q_{lm} = p_{lm}$ for those $(l,m) \in \Lambda_Z$, $Z \le Z_n$, and $q_{lm} = 0$ otherwise. Since (φ_{lm}) is an orthonormal base of $L^2(\mathbb{R})$, we have $\sum_{(l,m)\in\Lambda_{Z_n}} q_{lm}^2 = 1$, and we can find a function $\mathcal{N}_n \ni \psi' = \sum_{(l,m)\in\Lambda_{Z_n}} q'_{lm} e^{i\zeta'_{lm}} \varphi_{lm}$ such that $\sum_{(l,m)\in\Lambda_{Z_n}} |q'_{lm} - q_{lm}|^2 \le \delta_n^4$, and $|\zeta'_{lm} - \zeta_{lm}| \le \delta_n^2$ for all $(l,m) \in \Lambda_{Z_n}$. Using standard arguments, we have

$$\begin{aligned} \|\psi' - \psi\|_2^2 &= \sum_{(l,m)\in\Lambda_Z} \left| q'_{lm} e^{i\zeta'_{lm}} - q_{lm} e^{i\zeta_{lm}} \right|^2 \\ &\le 2 \sum_{(l,m)\in\Lambda_Z} |q'_{lm} - q_{lm}|^2 + 2 \sum_{(l,m)\in\Lambda_Z} q_{lm}^2 \left| e^{i\zeta'_{lm}} - e^{i\zeta_{lm}} \right|^2 \le 4\delta_n^4. \end{aligned}$$

Thus $\mathcal{N}_n$ is a $2\delta_n^2$ over $\mathcal{F}_n$ in the $\|\cdot\|_2$ norm. Moreover, the cardinality of $\mathcal{N}_n$ is upper bounded by $|\widehat{\mathcal{P}}| \times |\widehat{\mathcal{O}}|^{|\Lambda_{Z_n}|}$, which is in turn bounded by

$$C \left(\frac{1}{\delta_n^4} \right)^{|\Lambda_{Z_n}|} \left(\frac{2\pi}{\delta_n^2} \right)^{|\Lambda_{Z_n}|},$$

for a constant $C > 0$. Clearly, the cardinality of a $\sqrt{2}\delta_n^2$-net over $\mathcal{F}_n$ in the $\|\cdot\|_2$ distance satisfy the same bound, eventually for a different constant C. Therefore, for a suitable constant $K > 0$, when n is large enough.

$$N(\sqrt{2}\delta_n^2, \mathcal{F}_n, \|\cdot\|_2) \lesssim \exp\left\{ K |\Lambda_{Z_n}| \log \frac{1}{\delta_n} \right\} \lesssim \exp\left\{ K h^2 (\log n)^{1+2/r} \right\}.$$

The conclusion follows because $b_1 > 2 + r$. ■

Proof of proposition 4.5. The bound is obvious for those $\psi \in \mathcal{F}_n$ with $Z = 0$. For $Z \ge 1$, we have from the definition of the Wigner transform (equation (4.1)), for an arbitrary $\psi \in \mathcal{F}_n$,

$$W_\psi(x,\omega) = \sum_{(l,m)\in\Lambda_Z} \sum_{(j,k)\in\Lambda_Z} p_{lm} p_{jk} e^{i(\zeta_{lm} - \zeta_{jk})} \int_{\mathbb{R}} \varphi_{lm}(x + t/2) \overline{\varphi_{jk}(x - t/2)} e^{-2\pi i \omega t} \, dt.$$

Using the expression of φ_{lm} from equation (4.6), it follows

$$\begin{aligned} \varphi_{lm}(x + t/2)\overline{\varphi_{jk}(x - t/2)} &= c_l c_j T_m M_l \varphi(x + t/2) \overline{T_k M_j \varphi(x - t/2)} \\ &\quad + (-1)^{2k+j} c_l c_j T_m M_l \varphi(x + t/2) \overline{T_k M_{-j} \varphi(x - t/2)} \\ &\quad + (-1)^{2m+l} c_l c_j T_m M_{-l} \varphi(x + t/2) \overline{T_k M_j \varphi(x - t/2)} \\ &\quad + (-1)^{2m+l} (-1)^{2k+j} c_l c_j T_m M_{-l} \varphi(x + t/2) \overline{T_k M_{-j} \varphi(x - t/2)}. \end{aligned}$$

Recalling that $T_x\varphi(y) = \varphi(y-x)$ and $M_\omega\varphi(y) = e^{2\pi i\omega y}\varphi(y)$, it follows

$$\begin{aligned}
&\int_{\mathbb{R}} T_m M_l\varphi(x+t/2)\overline{T_k M_j\varphi(x-t/2)}e^{-2\pi i\omega t}\,dt \\
&\qquad = \int_{\mathbb{R}} e^{2\pi il(x+t/2-m)}\varphi(x+t/2-m)e^{-2\pi ij(x-t/2-k)}\overline{\varphi(x-t/2-k)}e^{-2\pi i\omega t}\,dt \\
&\qquad = 2e^{4\pi i\omega(x-m)-2\pi ij(2x-m-k)}\int_{\mathbb{R}} \varphi(u)\overline{\varphi(-u+2x-m-k)}e^{-2\pi iu(2\omega-l-j)}\,du \\
&\qquad = 2e^{4\pi i\omega(x-m)-2\pi ij(2x-m-k)}V_{\breve{\varphi}}\varphi(2x-m-k,2\omega-l-j).
\end{aligned}$$

Thus, we deduce the following expression for the Wigner transform of an arbitrary function $\psi \in \mathcal{F}_n$.

$$\begin{aligned}
W_\psi(x,\omega) = \sum_{(l,m)\in\Lambda_Z}\sum_{(j,k)\in\Lambda_Z} p_{lm}p_{jk}e^{i(\zeta_{lm}-\zeta_{jk})} \times 2c_lc_je^{4\pi i\omega(x-m)}\Big[\\
e^{-2\pi ij(2x-m-k)}V_{\breve{\varphi}}\varphi(2x-m-k,2\omega-l-j) \\
+(-1)^{2k+j}e^{2\pi ij(2x-m-k)}V_{\breve{\varphi}}\varphi(2x-m-k,2\omega-l+j) \\
+(-1)^{2m+l}e^{-2\pi ij(2x-m-k)}V_{\breve{\varphi}}\varphi(2x-m-k,2\omega+l-j) \\
+(-1)^{2m+l}(-1)^{2k+j}e^{2\pi ij(2x-m-k)}V_{\breve{\varphi}}\varphi(2x-m-k,2\omega+l+j)\Big].
\end{aligned}$$

To ease notations, let

$$f(x,\omega;l,m,j,k) := e^{4\pi i\omega(x-m)-2\pi ij(2x-m-k)}V_{\breve{\varphi}}\varphi(2x-m-k,2\omega-l-j).$$

Let $\mathscr{R}f(z,\theta)$ denote the Radon transform of f. A straightforward calculus show that $\mathscr{F}[\mathscr{R}f(\cdot,\theta)](u) = \widehat{f}(u\cos\theta, u\sin\theta)$, where $\widehat{f}$ is the Fourier transform with respect to both variables of f, and $\mathscr{F}$ the L^1 Fourier operator. Note that,

$$\begin{aligned}
&\int V_{\breve{\varphi}}\varphi(x,y)e^{\pi ixy}e^{-2\pi i(x\xi_1+y\xi_2)}\,dxdy \\
&\qquad = \int_{\mathbb{R}^2}\int_{\mathbb{R}} \varphi(u)\overline{\varphi(x-u)}e^{-2\pi iuy}\,du\,e^{\pi ixy}e^{-2\pi i(x\xi_1+y\xi_2)}dxdy \\
&\qquad = \iint \varphi(u)e^{-\pi iuy}\int \overline{\varphi(t)}e^{2\pi it(\xi_1-y/2)}dt\,e^{-2\pi iu\xi_1-2\pi iy\xi_2}dy\,du \\
&\qquad = 2e^{4\pi i\xi_1\xi_2}\int \widehat{\varphi}(t)\overline{\widehat{\varphi}(t-2\xi_1)}e^{-4\pi it\xi_2}\,dt \\
&\qquad = 2e^{4\pi i\xi_1\xi_2}V_{\widehat{\varphi}}\widehat{\varphi}(2\xi_1,2\xi_2).
\end{aligned}$$

Hence,

$$|\widehat{f}(u\cos\theta,u\sin\theta;l,m,j,k)| = \frac{1}{2}|V_{\widehat{\varphi}}\widehat{\varphi}(u\cos\theta+j-l,u\sin\theta+m-k)|$$

By Fourier duality, this implies that

$$\sup_x|\mathscr{R}f(\cdot;l,m,j,k)(x,\theta)| \le \frac{1}{2}\int|V_{\widehat{\varphi}}\widehat{\varphi}(u\cos\theta+j-l,u\sin\theta+m-k)|\,du$$

The function φ is in $\mathcal{S}_1^1(\mathbb{R})$ by construction. From Gröchenig and Zimmermann (2004, corollary 3.10) we can then find a constant $a > 0$ such that it holds $|V_{\widehat{\varphi}}\widehat{\varphi}(x,\omega)| \lesssim \exp(-a\sqrt{x^2+\omega^2})$. Moreover,

$$\begin{aligned}(u\cos\theta + j - l)^2 + (u\sin\theta + m - k)^2 \\ &= (u + (j-l)\cos\theta + (m-k)\sin\theta)^2 + ((m-k)\cos\theta - (j-l)\sin\theta)^2 \\ &\geq (u + (j-l)\cos\theta + (m-k)\sin\theta)^2 .\end{aligned}$$

Therefore,

$$\sup_{x,\theta} |\mathscr{R}f(\cdot; l, m, j, k)(x,\theta)| \lesssim \int \exp(-a|u|)\, du = 2a^{-1}.$$

Since the Radon transform is a linear map, we deduce that

$$|p_\psi(x,\theta)| \lesssim 8a^{-1}\left(\textstyle\sum_{(l,m)\in\Lambda_Z} p_{lm}\right)^2 \leq 8a^{-1}|\lambda_Z| \leq 8a^{-1}h^2(\log n)^{2/r}.$$

Now $p_\psi^\eta(y,\theta) = [p_\psi(\cdot,\theta) * G_\gamma](y)$, so that conclusion of the proposition follows from Young's inequality. ■

Proof of proposition 4.6. Using the expression of φ_{lm} of equation (4.6), we have

$$V_g\varphi_{lm} = c_l V_g(T_m M_l \varphi) + (-1)^{2m+l} c_l V_g(T_m M_{-l}\varphi).$$

Since $|V_g(T_m M_l\varphi)(x,\omega)| = |V_g(x-m,\omega-m)|$, it follows

$$|V_g\varphi_{lm}(x,\omega)| \leq c_l|V_g\varphi(x-m,\omega-l)| + c_l|V_g\varphi(x-m,\omega+l)|.$$

Now pick an arbitrary $\psi \in \mathcal{F}_n$. We have

$$\begin{aligned}\int_{\mathbb{R}^2} |V_g\psi(z)| \exp(\beta\|z\|^r)\, dz \\ &\leq \sum_{(l,m)\in\Lambda_Z} p_{lm} \int_{\mathbb{R}^2} |V_g\varphi_{lm}(z)| \exp(\beta\|z\|^r)\, dz \\ &\leq \sum_{(l,m)\in\Lambda_Z} p_{lm}c_l \int_{\mathbb{R}^2} |V_g\varphi(x-m,\omega-l)| \exp\left(\beta(x^2+\omega^2)^{r/2}\right) dxd\omega \\ &\quad + \sum_{(l,m)\in\Lambda_Z} p_{lm}c_l \int_{\mathbb{R}^2} |V_g\varphi(x-m,\omega+l)| \exp\left(\beta(x^2+\omega^2)^{r/2}\right) dxd\omega \\ &\leq 2\sum_{(l,m)\in\Lambda_Z} p_{lm} \exp\left(\beta(l^2+m^2)^{r/2}\right) \int_{\mathbb{R}^2} |V_g\varphi(z)| \exp(\beta\|z\|^r)\, dz \\ &\lesssim \sum_{(l,m)\in\Lambda_Z} p_{lm} \exp\left(\beta(l^2+m^2)^{r/2}\right) \lesssim (\log n)^{a_4},\end{aligned}$$

where the last line follows from Gröchenig and Zimmermann (2004, corollary 3.10), since both g and φ are in $\mathcal{S}_1^1(\mathbb{R})$ and $r < 1$ by assumption. The previous estimate show that $\mathcal{F}_n \subset \mathcal{C}_g(\beta, r, L_n)$ with $L_n \lesssim (\log n)^{a_4}$. Hence the conclusion follows from proposition 4.7. ■

4.C.2. Proofs of norm equivalence

Proof of proposition 4.7. Recall that $\mathscr{F}$ denote the L^1 Fourier transfom operator. By definition of W_ψ, it holds $W_\psi(u_1,u_2) = \mathscr{F}[\psi(u_1+\cdot/2)\overline{\psi(u_1-\cdot/2)}](u_2)$. Clearly if $\psi \in \mathcal{C}_g(\beta,r,L)$ then $W_\psi \in L^1(\mathbb{R}^2)$ by lemma 4.7. Moreover, for all $u_1 \in \mathbb{R}$ the mapping $t \mapsto \psi(u_1+t/2)\overline{\psi(u_1-t/2)}$ is in $L^1(\mathbb{R})$ because of Cauchy-Schwarz inequality and $\psi \in L^2(\mathbb{R})$. Then by Fourier inversion, we get

$$\int W_\psi(u_1,u_2)e^{-2\pi i u_2(-\xi_2)}\,du_2 = \psi(u_1+\xi_2/2)\overline{\psi(u_1-\xi_2/2)}.$$

Taking the Fourier transform with respect to u_1 yields

$$\begin{aligned}\iint W_\psi(u_1,u_2)e^{-2\pi i(u_1\xi_1+u_2\xi_2)}\,du_1du_2 &= \int \psi(u_1-\xi_2/2)\overline{\psi(u_1+\xi_2/2)}e^{-2\pi i u_1\xi_1}\,du_1 \\ &= e^{-\pi i\xi_1\xi_2}\int \psi(t)\overline{\psi(t+\xi_2)}e^{-2\pi i\xi_1 t}\,dt.\end{aligned}$$

Hence we proved that $\widehat{W}_\psi(\xi_1,\xi_2) = e^{-\pi i\xi_1\xi_2}V_\psi\psi(-\xi_2,\xi_1)$, at least when $\psi \in \mathcal{C}_g(\beta,r,L)$. By lemma 4.6, $|V_\psi\psi(-\xi_2,\xi_1)| \le (|V_g\psi| * |V_\psi g|)(-\xi_2,\xi_1)$ since $\|g\|_2 = 1$. Note that, by proposition 4.9 we have

$$\exp(\beta(\xi_1^2+\xi_2^2)^{r/2}) \le \exp(\beta((-\xi_2-u_1)^2+(\xi_1-u_2)^2)^{r/2})\exp(\beta(u_1^2+u_2^2)^{r/2}).$$

Also, by Cauchy-Schwarz inequality $|\widehat{W}_\psi(\xi_1,\xi_2)| \le \|\psi\|_2^2 = 1$. Therefore, by Young's inequality, and because $|V_\psi g| = |V_g\psi|$,

$$\begin{aligned}&\iint |\widehat{W}_\psi(\xi_1,\xi_2)|^2\exp(\beta(\xi_1^2+\xi_2^2)^{r/2})\,d\xi_1d\xi_2 \\ &\qquad\le \iint |\widehat{W}_\psi(\xi_1,\xi_2)|\exp(\beta(\xi_1^2+\xi_2^2)^{r/2})\,d\xi_1d\xi_2 \\ &\qquad\le \left(\iint |V_g\psi(\xi_1,\xi_2)|\exp(\beta(\xi_1^2+\xi_2^2)^{r/2})\,d\xi_1d\xi_2\right)^2,\end{aligned}$$

which concludes the proof. ■

Proof of lemma 4.5. The lower bound follows from lemma 4.11 in section 4.F. In the sequel we let $M_n := Mh^2(\log n)^{2/r}$ and $R_n := R(\log n)^{2a_4}\exp(-\beta u^r)$. To establish the upper bound, we first bound the L^2 distance between densities by the Hellinger distance. By triangular inequality and Young's inequality,

$$\begin{aligned}|p_\psi^\eta(y,\theta) - p_{\psi_0}^\eta(y,\theta)|^2 \le{}& 2\left|\sqrt{p_\psi^\eta(y,\theta)}\sqrt{p_{\psi_0}^\eta(y,\theta)} - \sqrt{p_\psi^\eta(y,\theta)}\sqrt{p_\psi^\eta(y,\theta)}\right|^2 \\ &+ 2\left|\sqrt{p_\psi^\eta(y,\theta)}\sqrt{p_{\psi_0}^\eta(y,\theta)} - \sqrt{p_{\psi_0}^\eta(y,\theta)}\sqrt{p_{\psi_0}^\eta(y,\theta)}\right|^2.\end{aligned}$$

Taking the integral both sides, under the assumptions of the lemma it comes

$$\iint |p_\psi^\eta(y,\theta) - p_{\psi_0}^\eta(y,\theta)|^2\,dyd\theta \le 2(C_0+M_n)H^2(P_\psi^\eta, P_{\psi_0}^\eta).$$

Recall that $\mathscr{F}$ denote the L^1-Fourier transform operator. Then by Parseval-Plancherel formula we can rewrite

$$\iint |\mathscr{F}[p_\psi^\eta(\cdot,\theta)](\xi) - \mathscr{F}[p_{\psi_0}^\eta(\cdot,\theta)](\xi)|^2 \, d\xi d\theta \leq 2(C_0 + M_n) H^2(P_\psi^\eta, P_{\psi_0}^\eta).$$

Recalling that $p_\psi^\eta(y,\theta) = [p_\psi(\cdot,\theta) * G_\gamma](y)$, where $\mathscr{F}[G_\gamma] = \widehat{G}_\gamma$, it follows $\mathscr{F}[p_\psi^\eta(\cdot,\theta)](\xi) = \mathscr{F}[p_\psi(\cdot,\theta)](\xi)\widehat{G}_\gamma(\xi)$. Therefore,

$$\iint |\mathscr{F}[p_\psi(\cdot,\theta)](\xi) - \mathscr{F}[p_{\psi_0}(\cdot,\theta)](\xi)|^2 |\widehat{G}_\gamma(\xi)|^2 \, d\xi d\theta \leq 2(C_0 + M_n) H^2(P_\psi^\eta, P_{\psi_0}^\eta).$$

Using that $\mathscr{F}[p_\psi(\cdot,\theta)](\xi) = \widehat{W}_\psi(\xi\cos\theta, \xi\sin\theta)$, and performing the suitable change of variables, we arrive at

$$\int_{\mathbb{R}^2} |\widehat{W}_\psi(z) - \widehat{W}_{\psi_0}(z)|^2 |\widehat{G}_\gamma(\|z\|)|^2 \, dz \leq 2(C_0 + M_n) H^2(P_\psi^\eta, P_{\psi_0}^\eta).$$

Now, using that the Fourier transform is isometric from $L^2(\mathbb{R})$ onto itself, and that the Wigner transform is isometric from $L^2(\mathbb{R})$ onto $L^2(\mathbb{R}^2)$, by Gröchenig (2001, proposition 4.3.2), we write

$$\begin{aligned}
\|\psi - \psi_0\|_2^2 &= \int_{\mathbb{R}^2} |\widehat{W}_\psi(z) - \widehat{W}_{\psi_0}(z)|^2 \, dz \\
&= \int_{\{\|z\| \leq u\}} |\widehat{W}_\psi(z) - \widehat{W}_{\psi_0}(z)|^2 \, dz + \int_{\{\|z\| > u\}} |\widehat{W}_\psi(z) - \widehat{W}_{\psi_0}(z)|^2 \, dz \\
&\leq \frac{1}{|\widehat{G}_\gamma(u)|^2} \int_{\mathbb{R}^2} |\widehat{W}_\psi(z) - \widehat{W}_{\psi_0}(z)|^2 |\widehat{G}_\gamma(\|z\|)|^2 \, dz \\
&\qquad + \int_{\{\|z\| > u\}} |\widehat{W}_\psi(z) - \widehat{W}_{\psi_0}(z)|^2 \, dz.
\end{aligned}$$

Under the hypothesis of the lemma, the second term in the rhs of the last equation is bounded by $4R_n$ when n is large, because by proposition 4.7 we have

$$\begin{aligned}
\int_{\{\|z\| > u\}} |\widehat{W}_{\psi_0}(z)|^2 \, dz &= \int_{\{\|z\| > u\}} |\widehat{W}_{\psi_0}(z)|^2 \, e^{\beta\|z\|^r} e^{-\nu\|z\|^r} dz \\
&\leq e^{-\beta u^r} \int_{\mathbb{R}^2} |\widehat{W}_{\psi_0}(z)|^2 \, e^{\beta\|z\|^r} dz \leq L^2 e^{-\beta u^r}.
\end{aligned}$$

Since $\widehat{G}_\gamma(\xi) = \exp(-\gamma\xi^2)$, it follows,

$$\begin{aligned}
\|\psi - \psi_0\|_2^2 &\leq \frac{1}{|\widehat{G}_\gamma(u)|^2} \int_{\mathbb{R}^2} |\widehat{W}_\psi(z) - \widehat{W}_{\psi_0}(z)|^2 |\widehat{G}_\gamma(\|z\|)|^2 \, dz + 4R_n \\
&\leq 2(C_0 + M_n) e^{2\gamma u^2} H^2(P_\psi^\eta, P_{\psi_0}^\eta) + 4R_n.
\end{aligned}$$

Consequently, when $\|\psi - \psi_0\|_2^2 \geq 8R_n$ we have

$$\|\psi - \psi_0\|_2^2 \leq 4(C_0 + M_n) e^{2\gamma u^2} H^2(P_\psi^\eta, P_{\psi_0}^\eta). \qquad \blacksquare$$

4.C.3. Construction of global test functions

Proof of theorem 4.4. Let $N \equiv N(\sqrt{2}\delta_n^2, \mathcal{F}_n, \|\cdot\|_2)$ denote the number of balls of radius $\sqrt{2}\delta_n^2$ and centers in $\mathcal{F}_n$, needed to cover $\mathcal{F}_n$. Let $(B_1, \dots, B_N)$ denote the corresponding covering with centers $(\psi_1, \dots, \psi_N)$. Now let J be the index set of balls B_j with $\|\psi_j - \psi_0\|_2 \geq \epsilon_n$. Using proposition 4.8, for each of these balls B_j with $j \in J$, we can build a test function $\phi_{n,j}$ satisfying

$$P_{\psi_0}^{\eta,n}\phi_{n,j} \leq \exp(-6n\delta_n^2), \qquad \sup_{\psi \in B_j} P_{\psi}^{\eta,n}(1 - \phi_{n,j}) \leq \exp(-6n\delta_n^2).$$

Define the test function $\phi_n := \max_{j\in J} \phi_{n,j}$. Then it holds $P_{\psi_0}^{\eta,n}\phi_n \leq \sum_{j\in J} P_{\psi_0}^{\eta,n}\phi_{n,j} \leq N\exp(-6n\delta_n^2)$ and $P_{\psi}^{\eta,n}(1-\phi_n) \leq \min_{j\in J} \sup_{\psi' \in B_j} P_{\psi'}^{\eta,n}(1-\phi_{n,j}) \leq \exp(-6n\delta_n^2)$ for any $\psi \in \mathcal{F}_n$ with $\|\psi - \psi_0\|_2 \geq \epsilon_n - \sqrt{2}\delta_n^2$ (recall that $\delta_n \ll \epsilon_n$), and hence for any $\psi \in \mathcal{F}_n$ with $\|\psi - \psi_0\|_2 \geq \epsilon_n$. ■

4.D. Proofs for uniform series prior on simplex

Proof of proposition 4.1. From the definition of of G and Hölder's inequality, for $K \geq 0$ integer, $Z = KM$ and $(p_{lm})_{(l,m)\in\Lambda_Z}$ in the support of $G(\cdot \mid Z)$, we get estimate

$$\begin{aligned}\sum_{(l,m)\in\Lambda_Z} p_{lm}\exp(\beta(l^2+m^2)^{r/2}) &\leq \sum_{k=1}^{K} \theta_k \sum_{(l,m)\in\mathcal{I}_k} \eta_{lm}\exp(\beta(l^2+m^2)^{r/2}) \\ &\leq \sum_{k=1}^{K} \theta_k \sqrt{|\mathcal{I}_k|}\exp(\beta k^r M^r),\end{aligned}$$

because $\sum_{k=1}^{K}\theta_k^2 \geq \theta_1^2 = 1$. The conclusion is direct because $\theta_1 = 1$ and θ_k is upper bounded by $\sqrt{2}L\exp(-\beta(k^r-1)M^r)$ for any $k = 2, \dots, K$. ■

Proof of proposition 4.2. Let $Z = KM$ for $K > 0$ integer, and $(q_{lm})_{(l,m)\in\Lambda_Z} \in \Delta_Z^w(\beta, r, L)$ be arbitrary. For any $(l,m) \in \Lambda_Z$, and any sequence $(p_{lm})_{(l,m)\in\Lambda_Z} \in \Delta_Z$, let define the unnormalized coefficients

$$\widetilde{q}_{lm} := \frac{q_{lm}}{\sqrt{\sum_{(n,p)\in\mathcal{I}_1} q_{np}^2}}, \qquad \widetilde{p}_{lm} := \frac{p_{lm}}{\sqrt{\sum_{(n,p)\in\mathcal{I}_1} p_{np}^2}},$$

Note that $\sum_{(l,m)\in\mathcal{I}_1} \widetilde{q}_{lm}^2 = \sum_{(l,m)\in\mathcal{I}_1} \widetilde{p}_{lm}^2 = 1$. Moreover, we also have

$$\sum_{(l,m)\in\Lambda_Z} p_{lm}^2 = \sum_{(l,m)\in\Lambda_Z} q_{lm}^2 = 1;$$

it turns out that

$$q_{lm} = \frac{\widetilde{q}_{lm}}{\sqrt{\sum_{(n,p)\in\Lambda_Z} \widetilde{q}_{np}^2}}, \qquad p_{lm} = \frac{\widetilde{p}_{lm}}{\sqrt{\sum_{(n,p)\in\Lambda_Z} \widetilde{p}_{np}^2}}.$$

By the triangle inequality, the two previous expressions of q_{lm}, p_{lm} yield the bound,

$$\sqrt{\sum_{(l,m)\in\Lambda_Z} |q_{lm}-p_{lm}|^2} \le \frac{2\sqrt{\sum_{(l,m)\in\Lambda_Z} |\widetilde{q}_{lm}-\widetilde{p}_{lm}|^2}}{\sqrt{\sum_{(l,m)\in\Lambda_Z} \widetilde{q}_{lm}^2}} \le 2\sqrt{\sum_{(l,m)\in\Lambda_Z} |\widetilde{q}_{lm}-\widetilde{p}_{lm}|^2}.$$

For any $k=1,\dots,K$, define $t_k := \sum_{(l,m)\in\mathcal{I}_k} \widetilde{q}_{lm}^2$ and $e_{lm} := \widetilde{q}_{lm} t_k^{-1} \mathbb{1}((l,m)\in\mathcal{I}_k)$. Note that by construction we have $t_1 = 1$. With obvious definition for θ_k and η_{lm}, we have

$$\begin{aligned}\sum_{(l,m)\in\Lambda_Z} |p_{lm}-q_{lm}|^2 &\le 2\sum_{k=1}^{K} \sum_{(l,m)\in\mathcal{I}_k} |\theta_k\eta_{lm} - t_k e_{lm}|^2 \\ &\le 4\sum_{k=1}^{K} t_k^2 \sum_{(l,m)\in\mathcal{I}_k} |\eta_{lm}-e_{lm}|^2 + 4\sum_{k=2}^{Z} |\theta_k - t_k|^2.\end{aligned}$$

We can choose $M > 0$ large enough to have $\sum_{(l,m)\in\mathcal{I}_1} q_{lm}^2 \ge 1/2$; it turns out that $\sum_{k=1}^K t_k^2 \le 2$. Moreover, with $M > 0$ chosen as previously we have

$$\begin{aligned}t_k \exp(\beta k^r M^r) &= \sqrt{2} e^{\beta M^r} \sum_{(l,m)\in\mathcal{I}_k} q_{lm} \exp(\beta(k-1)^r M^r) \\ &\le \sqrt{2} e^{\beta M^r} \sum_{(l,m)\in\mathcal{I}_k} q_{lm} \exp(\beta(l^2+m^2)^{r/2}) \le \sqrt{2} L e^{\beta M^r},\end{aligned}$$

thus the coefficients $(t_k)_{k=1}^K$ are in the support of $G(\cdot \mid Z)$. By independence structure of the prior, and since $\sum_{k=1}^K t_k^2 \le 2$, it suffices to prove that for any $t > 0$,

$$\prod_{k=1}^K F_k\left(\sum_{(l,m)\in\mathcal{I}_k} |\eta_{lm}-e_{lm}|^2 \le t\right) \gtrsim \exp(-cK^{b_1-r}\log t^{-1}), \tag{4.20}$$

$$\mathbb{P}\left(\sum_{k=2}^K |\theta_k - t_k|^2 \le t\right) \gtrsim \exp(-c' K^{b_1-r}\log t^{-1}), \tag{4.21}$$

for some constants $c, c' > 0$. Equation (4.20) is automatically satisfied by the assumptions on $F_1, F_2, \dots$ in the proposition. Equation (4.21) is straightforward from the definition of $G(\cdot \mid Z)$. ■

4.E. Bounding the posterior distribution

We bound the posterior distribution as follows. Let Ω_n be the event of equation (4.11). Then, with the notation $Z_i := (Y_i, \theta_i)$ and $Z^n = (Z_1, \dots, Z_n)$, for any measurable set U_n,

$$P_{\psi_0}^{\eta,n}\Pi(U_n \mid Z^n) = P_{\psi_0}^{\eta,n}(\Omega_n)\,[I_1^n + I_2^n + I_3^n] + P_{\psi_0}^{\eta,n}(\Omega_n^c) I_4^n, \tag{4.22}$$

where

$$\begin{aligned}I_1^n &:= \int_{\Omega_n} \Pi(U_n \cap \mathcal{F}_n^c \mid z^n)\, dP_{\psi_0}^{\eta,n}(z^n \mid \Omega_n), \\ I_2^n &:= \int_{\Omega_n} \phi_n(z^n)\Pi(U_n \cap \mathcal{F}_n \mid z^n)\, dP_{\psi_0}^{\eta,n}(z^n \mid \Omega_n), \\ I_3^n &:= \int_{\Omega_n} (1-\phi_n(z^n))\Pi(U_n \cap \mathcal{F}_n \mid z^n)\, dP_{\psi_0}^{\eta,n}(z^n \mid \Omega_n), \\ I_4^n &:= \int_{\Omega_n^c} \Pi(U_n \mid z^n)\, dP_{\psi_0}^{\eta,n}(z^n \mid \Omega_n^c).\end{aligned}$$

This decomposition of the expectation for the posterior distribution serves as a basis for the proof of the next lemma.

Lemma 4.10 – *Let $\delta_n \to 0$ with $n\delta_n^2 \to \infty$. Assume that there are sets $\mathcal{F}_n \subset \mathcal{F}$ with $\Pi(\mathcal{F}_n^c) \leq e^{-6n\delta_n^2}$ and a sequence of test functions $(\phi_n)_{n\geq 1}$, $\phi_n : (\mathbb{R}^+ \times [0, 2\pi])^n \to [0, 1]$, such that $P_{\psi_0}^{\eta,n}\phi_n \to 0$ and $\sup_{\psi\in U_n\cap\mathcal{F}_n} P_\psi^{\eta,n}(1-\phi_n) \leq e^{-6n\delta_n^2}$. Also assume that $\Pi(B_n(\delta_n)) \gtrsim e^{-n\delta_n^2}$, where $B_n(\delta_n)$ are the sets defined in equation (4.12). Then $P_{\psi_0}^{\eta,n}\Pi(U_n \mid Z^n) \to 0$ as $n \to \infty$.*

Proof. The proof looks like Ghosal et al. (2000), with careful adaptions. It is obvious that $I_4^n \leq 1$ so that $P_{\psi_0}^{\eta,n}(\Omega_n^c)I_4^n \to 0$ by lemma 4.1. With the same argument we have that $I_2^n \leq P_{\psi_0}^{\eta,n}(\Omega_n)^{-1}P_{\psi_0}^{\eta,n}\phi_n$. Now we bound I_3^n. As usual, recalling that the observations are i.i.d we rewrite

$$\Pi(U_n \cap \mathcal{F}_n \mid z^n) = \frac{\int_{U_n\cap\mathcal{F}_n} \prod_{i=1}^n p_\psi^\eta(y_i,\theta_i)/p_{\psi_0}^\eta(y_i,\theta_i)\, d\Pi(\psi)}{\int \prod_{i=1}^n p_\psi^\eta(y_i,\theta_i)/p_{\psi_0}^\eta(y_i,\theta_i)\, d\Pi(\psi)}. \tag{4.23}$$

We lower bound the integral in the denominator of equation (4.23) by integrating on the smaller set B_n. Consider the events

$$A_n := \left\{((y_1,\theta_1),\dots,(y_n,\theta_n)) \,:\, \int_{B_n} \prod_{i=1}^n \frac{p_\psi^\eta(y_i,\theta_i)}{p_{\psi_0}^\eta(y_i,\theta_i)} \frac{d\Pi(\psi)}{\Pi(B_n)} \leq \exp(-4n\delta_n^2)\right\}$$
$$C_n := \left\{((y_1,\theta_1),\dots,(y_n,\theta_n)) \,:\, \sum_{i=1}^n \int_{B_n} \log\frac{p_{\psi_0}^\eta(y_i,\theta_i)}{p_\psi^\eta(y_i,\theta_i)} \frac{d\Pi(\psi)}{\Pi(B_n)} \geq 4n\delta_n^2\right\}.$$

By Jensen's inequality, we have the inclusion $C_n \subseteq A_n$, thus $P_{\psi_0}^{\eta,n}(A_n \mid \Omega_n) \leq P_{\psi_0}^{\eta,n}(C_n \mid \Omega_n)$. Moreover, using that the observations are independent, and Fubini's theorem, we have

$$\begin{aligned}
&P_{\psi_0}^{\eta,n}\left[\sum_{i=1}^n \int_{B_n} \log\frac{p_{\psi_0}^\eta(y_i,\theta_i)}{p_\psi^\eta(y_i,\theta_i)} \frac{d\Pi(\psi)}{\Pi(B_n)} \mid \Omega_n\right] \\
&\qquad = \frac{1}{P_{\psi_0}^{\eta,n}(\Omega_n)} \int_{\Omega_n} \sum_{i=1}^n \int_{B_n} \log\frac{p_{\psi_0}^\eta(y_i,\theta_i)}{p_\psi^\eta(y_i,\theta_i)} \frac{d\Pi(\psi)}{\Pi(B_n)}\, dP_{\psi_0}^{\eta,n}(\textstyle\prod_{j=1}^n dy_j d\theta_j \cap \Omega_n) \\
&\qquad = \frac{nP_{\psi_0}^\eta(E_n)^{n-1}}{P_{\psi_0}^{\eta,n}(\Omega_n)} \int_{B_n} \left[\int_{E_n} \log\frac{p_{\psi_0}(y,\theta)}{p_\psi(y,\theta)}\, dP_{\psi_0}^\eta(dyd\theta)\right] \frac{d\Pi(\psi)}{\Pi(B_n)} \\
&\qquad = \frac{n}{P_{\psi_0}^\eta(E_n)} \int_{B_n} \left[\int_{E_n} \log\frac{p_{\psi_0}(y,\theta)}{p_\psi(y,\theta)}\, dP_{\psi_0}^\eta(dyd\theta)\right] \frac{d\Pi(\psi)}{\Pi(B_n)}.
\end{aligned}$$

Likewise, we can bound the variance with respect to $P_{\psi_0}^{\eta,n}(\cdot \mid \Omega_n)$, denoted var for the

sake of simplicity; with the same arguments as previously,

$$\begin{aligned}\operatorname{var}\left[\sum_{i=1}^{n}\int_{B_n}\log\frac{p^{\eta}_{\psi_0}(y_i,\theta_i)}{p^{\eta}_{\psi}(y_i,\theta_i)}\frac{d\Pi(\psi)}{\Pi(B_n)}\right]&\\ \leq \frac{n}{P^{\eta}_{\psi_0}(E_n)}\int_{E_n}\left(\int_{B_n}\log\frac{p^{\eta}_{\psi_0}(y,\theta)}{p^{\eta}_{\psi}(y,\theta)}\frac{d\Pi(\psi)}{\Pi(B_n)}\right)^2 dP^{\eta}_{\psi_0}(y,\theta)&\\ \leq \frac{n}{P^{\eta}_{\psi_0}(E_n)}\int_{B_n}\left[\int_{E_n}\left(\log\frac{p^{\eta}_{\psi_0}(y,\theta)}{p^{\eta}_{\psi}(y,\theta)}\right)^2 dP^{\eta}_{\psi_0}(y,\theta)\right]\frac{d\Pi(\psi)}{\Pi(B_n)},&\end{aligned}$$

From the definition of B_n and because $P^{\eta}_{\psi_0}(E_n)\geq 1/2$ for n large enough, we get from Chebychev inequality that for those n,

$$P^{\eta,n}_{\psi_0}(A_n\mid\Omega_n)\leq P^{\eta,n}_{\psi_0}(C_n\mid\Omega_n)\leq\frac{1}{8n\delta_n^2}.$$

Hence,

$$\int_{\Omega_n\cap A_n}(1-\phi_n(z^n))\Pi(U_n\cap\mathcal{F}_n\mid z^n)\,dP^{\eta,n}_{\psi_0}(z^n\mid\Omega_n)\lesssim\frac{P^{\eta,n}_{\psi_0}(A_n)}{P^{\eta,n}_{\psi_0}(\Omega_n)}\leq\frac{(n\delta_n^2)^{-1}}{P^{\eta,n}_{\psi_0}(\Omega_n)},$$

and,

$$\begin{aligned}&\int_{\Omega_n\cap A_n^c}(1-\phi_n(z^n))\Pi(U_n\cap\mathcal{F}_n\mid z^n)\,dP^{\eta,n}_{\psi_0}(z^n\mid\Omega_n)\\ &\quad\leq\frac{e^{4n\delta_n^2}}{\Pi(B_n)}\int_{\Omega_n\cap A_n^c}(1-\phi_n(z^n))\int_{U_n\cap\mathcal{F}_n}\prod_{i=1}^{n}\frac{p^{\eta}_{\psi}(y_i,\theta_i)}{p^{\eta}_{\psi_0}(y_i,\theta_i)}\,d\Pi(\psi)dP^{\eta,n}_{\psi_0}(z^n\mid\Omega_n)\\ &\quad=\frac{e^{4n\delta_n^2}}{\Pi(B_n)}\int_{U_n\cap\mathcal{F}_n}\int_{\Omega_n\cap A_n^c}(1-\phi_n(z^n))\prod_{i=1}^{n}\frac{p^{\eta}_{\psi}(y_i,\theta_i)}{p^{\eta}_{\psi_0}(y_i,\theta_i)}\,dP^{\eta,n}_{\psi_0}(z^n\mid\Omega_n)d\Pi(\psi)\\ &\quad\leq\frac{e^{4n\delta_n^2}\Pi(U_n\cap\mathcal{F}_n)}{\Pi(B_n)}\frac{\sup_{\psi\in U_n\cap\mathcal{F}_n}P^{\eta,n}_{\psi}(1-\phi_n)}{P^{\eta,n}_{\psi_0}(\Omega_n)}.\end{aligned}$$

where the third line follows from Fubini's theorem. Combining the last two results yields $P^{\eta,n}_{\psi_0}(\Omega_n)I_3^n\to 0$. The bound on I_1^n follows exactly the same lines as the bound on I_3^n (see also Ghosal et al., 2000). ■

4.F. Remaining proofs and auxiliary results

Lemma 4.11 – *Let $\psi,\psi_0\in\mathbb{S}^2(\mathbb{R})$. Then, $H^2(P^{\eta}_{\psi},P^{\eta}_{\psi_0})\leq\sqrt{2}H(P_{\psi},P_{\psi_0})\leq\sqrt{2}\|\psi-\psi_0\|_2$. Moreover, we also have that $H(P_{\psi}(\cdot\mid\theta),P_{\psi_0}(\cdot\mid\theta))\leq\|\psi-\psi_0\|_2$ for all $\theta\in[0,\pi]$.*

Proof. First, we recall that $p^{\eta}_{\psi}(y,\theta)=[p_{\psi}(\cdot,\theta)*G_{\gamma}](y)$. The same holds for $p^{\eta}_{\psi_0}$. Then using that the square Hellinger distance is bounded by the total variation distance, which is in turn bounded by the Hellinger distance,

$$\begin{aligned}H^2(P^{\eta}_{\psi},P^{\eta}_{\psi_0})&\leq\iint|[p_{\psi}(\cdot,\theta)*G_{\gamma}](y)-[p_{\psi_0}(\cdot,\theta)*G_{\gamma}](y)|\,dyd\theta\\ &\leq\|G_{\gamma}\|_1\iint|p_{\psi}(x,\theta)-p_{\psi_0}(x,\theta)|\,dxd\theta\leq\sqrt{2}H(P_{\psi},P_{\psi_0}),\end{aligned}$$

where the second line follows from Young's inequality. Now let $\theta \neq 0$ and $\theta \neq \pi/2$. Using that $|x| - |y| = |x - y + y| - |y| \leq |x - y|$ for all $x, y \in \mathbb{C}$, it holds from equation (4.3) that,

$$\sqrt{p_\psi(x,\theta)} - \sqrt{p_{\psi_0}(x,\theta)} \leq \frac{1}{2\pi\sqrt{|\sin\theta|}} \left| \int_{-\infty}^{+\infty} (\psi(z) - \psi_0(z)) \exp\left(i\frac{\cot\theta}{2} z^2 - i\frac{x}{\sin\theta} z \right) dz \right|.$$

On almost recognize the expression of the square-root of a density in the rhs of the last equation. Indeed, it is not because $\psi - \psi_0$ is not normalized in L^2. But, letting $\psi_v := (\psi - \psi_0)/\|\psi - \psi_0\|_2$,

$$\left(\sqrt{p_\psi(x,\theta)} - \sqrt{p_{\psi_0}(x,\theta)} \right)^2 \leq p_v(x,\theta) \|\psi - \psi_0\|_2^2. \tag{4.24}$$

One can show easily that the same bound holds when $\theta = 0$ or $\theta = \pi/2$ (although it is even not necessary). The conclusion of the lemma then follows from the definition of the Hellinger distance and the fact that p_v is a probability density. The results for conditional densities is immediate from equation (4.24) since $p_\psi(x \mid \theta) = \pi p_\psi(x,\theta)$ for any $\psi \in \mathbb{S}^2(\mathbb{R})$. ■

Proposition 4.10 – *There exists n_0 such that for all $n \geq n_0$ and all $\psi \in \mathcal{M}_n(Z,U)$ it holds $P_\psi(|X| \leq D_n^{\beta,r} \mid \theta) \geq 1/2$ for all $\theta \in [0,\pi]$.*

Proof. It suffices to write that,

$$\begin{aligned} P_{\psi_0}(|X| \leq D_n^{\beta,r} \mid \theta) &\leq \int_{[-D_n^{\beta,r}, +D_n^{\beta,r}]} p_\psi(x \mid \theta)\, dx + \int_{\mathbb{R}} |p_\psi(x \mid \theta) - p_{\psi_0}(x \mid \theta)|\, dx \\ &\leq P_\psi(|X| \leq D_n^{\beta,r} \mid \theta) + \sqrt{2} H(P_\psi(\cdot \mid \theta), P_{\psi_0}(\cdot \mid \theta)). \end{aligned}$$

By lemma 4.11, $\sqrt{2}H(P_\psi(\cdot \mid \theta), P_{\psi_0}(\cdot \mid \theta)) \leq 1/4$ for all $\psi \in \mathcal{M}_n$ if n is large enough. Moreover, is n is sufficiently large, we also have $P_{\psi_0}(|X| \leq D_n^{\beta,r} \mid \theta) \geq 3/4$, concluding the proof. ■

Bibliography

P. Alquier, K. Meziani, and G. Peyré. Adaptive estimation of the density matrix in quantum homodyne tomography with noisy data. *Inverse Problems*, 29(7):075017, 2013.

L. Artiles, R. Gill, et al. An invitation to quantum tomography. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):109–134, 2005.

J.-M. Aubry, C. Butucea, and K. Méziani. State estimation in quantum homodyne tomography with noisy data. *Inverse Problems*, 25(1):015003, 2008.

C. Butucea, M. Guţă, and L. Artiles. Minimax and adaptive estimation of the wigner function in quantum homodyne tomography with noisy data. *The Annals of Statistics*, 35(2):465–494, 2007.

E. Cordero. Gelfand–shilov window classes for weighted modulation spaces. *Integral Transforms and Special Functions*, 18(11):809–817, 2007.

E. Cordero, S. Pilipović, L. Rodino, and N. Teofanov. Localization operators and exponential weights for modulation spaces. *Mediterranean Journal of Mathematics*, 2(4): 381–394, 2005.

I. Daubechies, S. Jaffard, and J.-L. Journé. A simple wilson orthonormal basis with exponential decay. *SIAM Journal on Mathematical Analysis*, 22(2):554–573, 1991.

S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 02 2007. doi: 10.1214/009053606000001172. URL http://dx.doi.org/10.1214/009053606000001172.

S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

K. Gröchenig. Foundations of time-frequency analysis, 2001.

K. Gröchenig. Weight functions in time-frequency analysis. *arXiv preprint math/0611174*, 2006.

K. Gröchenig and G. Zimmermann. Hardy's theorem and the short-time fourier transform of schwartz functions. *Journal of the London Mathematical Society*, 63(1):205–214, 2001.

K. Gröchenig and G. Zimmermann. Spaces of test functions via the stft. *Journal of Function Spaces and Applications*, 2(1):25–53, 2004.

M. Guţă and L. Artiles. Minimax estimation of the wigner function in quantum homodyne tomography with ideal detectors. *Mathematical Methods of Statistics*, 16(1):1–15, 2007.

B. C. Hall. *Quantum theory for mathematicians*. Springer, 2013.

S. Helgason. *The Radon Transform on R n*. Springer, 2011.

B. Knapik and J.-B. Salomond. A general approach to posterior contraction in nonparametric inverse problems. *arXiv preprint arXiv:1407.0335*, 2014.

K. Lounici, K. Meziani, and G. Peyré. Minimax and adaptive estimation of the wigner function in quantum homodyne tomography with noisy data. *arXiv preprint arXiv:1506.06941*, 2015.

T. Markus, H. Bryan, and O. Jorge. Phase-space optics fundamentals and applications, 2010.

K. Méziani. Nonparametric estimation of the purity of a quantum state in quantum homodyne tomography with noisy data. *Mathematical Methods of Statistics*, 16(4): 354–368, 2007.

K. Méziani. Nonparametric goodness-of fit testing in quantum homodyne tomography with noisy data. *Electronic journal of statistics*, 2:1195–1223, 2008.

Z. Naulet and E. Barat. Some aspects of symmetric gamma process mixtures. *arXiv preprint arXiv:1504.00476*, 2015.

W. Shen, S. T. Tokdar, and S. Ghosal. Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

Résumé

Cette thèse porte sur les statistiques bayésiennes non paramétriques. La thèse est divisée en une introduction générale et trois parties traitant des aspects relativement différents des approches par mélanges (échantillonage, asymptotique, problème inverse).

Dans les modèles de mélanges, le paramètre à inférer depuis les données est une fonction. On définit une distribution a priori sur un espace fonctionnel abstrait au travers d'une intégrale stochastique d'un noyau par rapport à une mesure aléatoire.

Habituellement, les modèles de mélanges sont surtout utilisés dans les problèmes d'estimation de densités de probabilité. Une des contributions de ce manuscrit est d'élargir leur usage aux problèmes de régressions.

Dans ce contexte, on est essentiellement concernés par les problèmes suivants:
- Echantillonage de la distribution a posteriori
- Propriétés asymptotiques de la distribution a posteriori
- Problèmes inverses, et particulièrement l'estimation de la distribution de Wigner à partir de mesurements de Tomographie Quantique Homodyne.

Abstract

This dissertation deals with Bayesian nonparametric statistics, in particular nonparametric mixture models. The manuscript is divided into a general introduction and three parts on rather different aspects of mixtures approaches (sampling, asymptotic, inverse problem).

In mixture models, the parameter to infer from the data is a function. We set a prior distribution on an abstract space of functions through a stochastic integral of a kernel with respect to a random measure.

Usually, mixture models were used primilary in probability density function estimation problems. One of the contributions of the present manuscript is to use them in regression problems.

In this context, we are essentially concerned with the following problems :
- Sampling of the posterior distribution
- Asymptotic properties of the posterior distribution
- Inverse problems, in particular the estimation of the Wigner distribution from Quantum Homodyne Tomography measurements.

Mots Clés

Statistiques bayésiennes non paramétriques, mélanges, vitesses de convergence, échantillonnage, tomographie quantique homodyne

Keywords

Bayesian nonparametric statistics, mixtures, rates of convergence, sampling, quantum homodyne tomography