Bayesian Analysis (2018) **13**, Number 3, pp. 703–720

Some Aspects of Symmetric Gamma Process Mixtures

Zacharie Naulet* and Éric Barat†

Abstract. In this article, we present some specific aspects of symmetric Gamma process mixtures for use in regression models. First we propose a new Gibbs sampler for simulating the posterior. The algorithm is tested on two examples, the mean regression problem with normal errors, and the reconstruction of two dimensional CT images. In a second time, we establish posterior rates of convergence related to the mean regression problem with normal errors. For location-scale and location-modulation mixtures the rates are adaptive over Hölder classes, and in the case of location-modulation mixtures are nearly optimal.

MSC 2010 subject classifications: Primary 62G08, 62G20; secondary 60G57.

Keywords: Bayesian nonparameterics, nonparametric regression, signed random measures.

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

$$f(\cdot) = \int_{\mathcal{X}} K(x;\cdot)Q(dx), \quad Q \sim \Pi_*, \tag{1}$$

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 (1) is popular in density estimation Escobar and West (1994); Müller et al. (1996); Ghosal and van der Vaart (2007); 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 (2004); 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.

*CEA, LIST, Laboratory of Modeling, Simulation and Systems, F-91191 Gif-sur-Yvette, France, zacharie.naulet@cea.fr

†CEA, LIST, Laboratory of Modeling, Simulation and Systems, F-91191 Gif-sur-Yvette, France, eric.barat@cea.fr

 https://doi.org/10.1214/17-BA1058

Indeed, both designing an algorithm or establishing posterior contraction results heavily depends on the choice of K and Π_* in (1); but above all also on the observation model we consider. This last point makes the study of mixtures in regression difficult 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.

The reason for the choice of symmetric Gamma process is two fold. From a practical point of view, we have $Q \overset{\text{a.s.}}{=} \sum_{i\in\mathbb{N}} q_i \delta_{x_i}$ for some random collection $(q_i, x_i)_{i\in\mathbb{N}}$, with *jump weights* q_i decaying very fast (almost-surely); hence (1) becomes $f(\cdot) \overset{\text{a.s.}}{=} \sum_{i\in\mathbb{N}} q_i K(x;\cdot)$, where the number of "large" q_i's is always small with high probability. In other words, symmetric Gamma process mixtures are *sparse* functions with high probability. The sparsity property is a great advantage, because the mixture f can be approximated well by a finite, relatively small, number of parameters, which allows for efficient posterior simulation. The second reason is theoretical, but somehow related to the previous. In nonparametric Bayes, we cannot be certain in general that posterior distributions contract at optimal rates at a given function f_0, though it is a desirable requirement. In general, the sparsity property has direct influence on the rates of contraction of the posterior distribution. If the prior mixture is too sparse with too much probability, then we are likely to achieve bad rates of contraction, and the same is true if we are not sparse enough. Symmetric Gamma process mixtures are among the priors with adequate sparsity property, making them appealing for theoretical study.

In the first part of the paper, we propose a Gibbs sampler to simulate 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 algorithm. Finally, we make an empirical study of the algorithm, with comparison with the *Reversible-Jump Monte Carlo Markov Chain* (RJMCMC) algorithm of Wolpert et al. (2011). Both our and Wolpert et al. (2011) algorithms are based on approximating the symmetric Gamma process random measure, but use different kind of approximations. Wolpert et al.'s algorithm applies to more general mixing measures, while we believe specializing to symmetric Gamma processes permits to benefit of finer approximation schemes.

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 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 Naulet and Barat (2017), Section S4). 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.

2.1 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

$$f(x;\omega) := \int_{\mathcal{E}\times\mathbb{R}^d} K_A(x-\mu)\, Q(dA d\mu;\omega), \qquad \forall x \in \mathbb{R}^d, \tag{2}$$

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 (2) is made clear in Rajput and Rosinski (1989).

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

$$f(x;\omega) := \int_{\mathbb{R}^d \times \mathbb{R}^d \times [0,\pi/2]} K_{\xi,\phi}(x-\mu)\, Q(d\xi d\mu d\phi;\omega), \qquad \forall x \in \mathbb{R}^d, \tag{3}$$

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 $\eta > 0$.

2.3 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, 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 (2) and (3) converge 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 $(\xi,\phi) \in \mathbb{R}^d \times [0,\pi/2]$.

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 section, inspired from a result about Dirichlet processes from Favaro et al. (2012), adapted to symmetric Gamma processes.

3.1 Finite support approximation to symmetric Gamma mixtures

In Theorem 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 to 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 $\{X_i \in \mathcal{X} : 1 \le i \le n\}$ 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 $A \in \mathcal{A}$,

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

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 S4.

Theorem 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 $\sqrt{\eta}$.*

In Theorem 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 establishes 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 S1.3. 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 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 $(x, y) \mapsto K(x; y)$. Then we have the following proposition, proved in Section S1.4.

Proposition 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 1. Then $\lim_{p \to \infty} \| f^{(Q_p)} - f^{(Q)} \|_\infty = 0$ almost-surely.*

3.2 Algorithm for posterior sampling

From Theorem 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 parametrized by the regression function $f : \mathbb{R}^d \to \mathbb{R}$ 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. $\mathrm{SGa}(1,1)$ 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. For $i = 1, \ldots, p$: sample $C_i \mid C_{-i}, Y, X^\star, J, T$. Let $n_{k,i} = \#_{\substack{1 \le l \le n \\ l \neq i}}\{C_l = k\}$, $\kappa^{(p)}$ the number of distinct X_k values and κ_0 a chosen natural,

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

   where $\mathcal{L}_{k,i}(X^\star, J, T \mid Y)$ 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. $X^\star \mid C, Y, J, T$. *Random Walk Metropolis Hastings* (RWMH) on parameters.

3. For $i = 1, \ldots, p$: sample $J_i \mid J_{-i}, C, Y, X^\star, T$ using independent Metropolis Hastings with prior $\mathrm{SGa}(1,1)$ taken as i.i.d. candidate distribution for J_i.

4. $T \mid C, Y, X^\star, J$. RWMH on scale parameter.

Of course, any other sampling method can be used in steps 2 to 4 if they allow to get samples from the right target distributions. Also, it often improves the performance of the algorithm to write $J_i = J_i^+ - J_i^-$ with $J_i^+, J_i^- \sim \mathrm{Ga}(1,1)$ and sample separately $J_i^+ \mid \text{rest}$ and $J_i^- \mid \text{rest}$.

3.3 Choosing the number of particles

It is not clear how to choose the number of particles in the algorithm. The next theorem shows that we can use the acceptance rate of the Metropolis updates in the step 3 of the algorithm to quantify the degree of approximation of Q_p. We write $\mathcal{L}(Q \mid y)$ for the likelihood function of the mixing measure Q given $Y = y$. The proof of Theorem 2 is to be found in Section S1.5.

Theorem 2. *Assume that for all observations y we have $\mathcal{L}(Q \mid y) > 0$ for all $Q \in \mathcal{M}(\mathcal{X})$ and $\sup_{Q \in \mathcal{M}(\mathcal{X})} \mathcal{L}(Q \mid y) < \infty$. Furthermore assume that $Q \mapsto \mathcal{L}(Q \mid y)$ is continuous (in the weak-* topology). Then for each $1 \le i \le p$ the law of $J_i \mid J_{-i}, C, Y, X^\star, T$ converges in distribution to* $\mathrm{SGa}(1,1)$ *as $p \to \infty$.*

The assumptions of the previous theorem are really mild and met by most of models encountered in practice. In particular, the continuity assumption is not that restrictive since the weak-* topology on $\mathcal{M}(\mathcal{X})$ is really weak. Henceforth, in general the acceptance rate for the J_i's moves in step 3 of the algorithm goes to one when $p \to \infty$, showing that it is a good indicator – a posteriori – of the degree of approximation of Q_p. In practice, we find that a level of acceptance around 30% is acceptable for most applications.

4 Univariate simulation study

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; see for instance Section 5 where we present simulation results for the multivariate inverse problem of CT imaging.

4.1 Models used for simulation study

We present results of our algorithm on several standard test functions from the wavelet regression literature (see Marron et al., 1998, and Figures 2 and 3), 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 more general.

For each test function, the noise variance is chosen so that the root signal-to-noise ratio is equal to 3 (a high noise level). We ran the algorithm for location-scale mixtures either of Gaussians or 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 3.2, we also added

- a Gibbs step estimation of the noise variance, with Inverse-Gamma prior distribution,
- 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.

4.2 Assessing the convergence of the Markov Chain

We propose to assess the convergence of the Markov Chain and ensuring it is well mixing using a combination of Geweke's convergence diagnostic (Geweke, 1992) and *Effective Sample Size* (ESS). The Markov Chain is initialized at random from prior distribution and we apply the tests to the log-likelihood function samples produced by the algorithm.

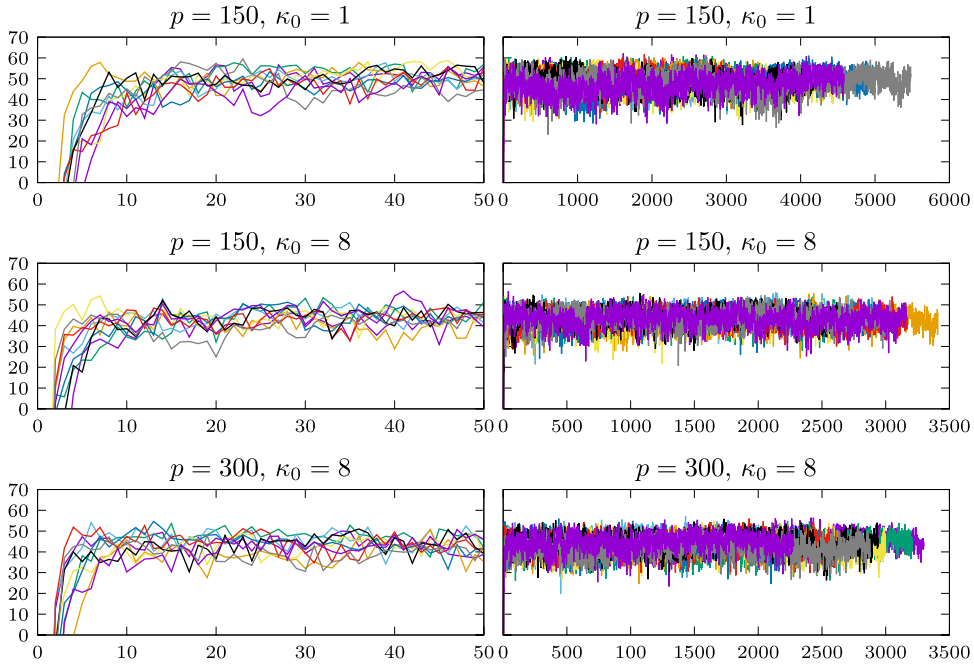

Figure 1: Time evolution of the log-likelihood for random starting point of the Markov Chain for various parameters of the algorithm. The figures are taken from the test function *angles* of Marron et al. (1998) with $n = 512$ equispaced observations and a root signal-to-noise ratio of 3 and a location-scale mixture of Gaussian prior. The left column shows the first samples of the chains, while the right column displays the whole chains. In every cases, the algorithm was run until getting an ESS of 1000 after removing the burn-in samples according to Geweke's test.

Here and after, we always choose step sizes in RWMH steps of the algorithm to achieve approximately 30% acceptance rates for each class of updates. Similarly, the number of particles is chosen to achieve around 30% of acceptance of the moves in the update of J_i's.

We discard the first n_0 samples of the chain according to the result of Geweke's test with significance level of 5%, and then we run the chain long enough to get an ESS of 1000 samples. In Figure 1, we draw some examples of temporal evolution of the log-likelihood for different runs of the algorithm. Each subfigure represent 10 simulations with random starting point of the Markov Chain, distributed according to the prior distribution. We drew each subfigure varying the parameters liable to influence the mixing time of the chain, notably κ_0 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. Clearly increasing p or κ_0 improve the mixing of the chain and reduce the number of samples needed to get an ESS of 1000. This last remark have to be

balanced with the complexity in time of the algorithm which is $O(np\kappa_0)$ for a naive implementation.

4.3 Error analysis

We ran the algorithm for $n = 128$ and $n = 1024$ data with rsnr $= 3$, 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 the algorithm with the specifications of Sections 4.1 and 4.2 about the prior and the assessment of the chain convergence. For each test function of Marron et al. (1998), a simulation run was repeated 100 times with all simulation parameters constant, excepting the noise which was regenerated.

In Wolpert et al. (2011), authors develop a RJMCMC 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 ran their algorithm on the same datasets 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 Section 4.1. We use the criteria of Section 4.2 to stop the running of the chain.

Tables 1 and 2 summarize the results for location-scale mixtures of Gaussians and Symmlet8 produced by our algorithm and by the RJMCMC algorithm of Wolpert et al. (2011). We give the average error of the TI-H with Symmlet8 method as reference, which is one the best performing algorithm on this collection of test functions (see Antoniadis et al., 2001). We used $p = 150$ particles and $\kappa_0 = 1$ for both the datasets with $n = 128$ covariates and $n = 1024$ covariates, which is a nice compromise in terms

	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 1: Summary of root mean squared errors averaged over 100 runs of different algorithms for $n = 128$ covariates and a root signal to noise ratio of 3.

	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: Summary of root mean squared errors averaged over 100 runs of different algorithms for $n = 1024$ covariates and a root signal to noise ratio of 3.

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.

The Gibbs algorithm allows for sampling the *full* posterior distribution, permitting estimation of posterior credible bands, as illustrated in Figures 2 and 3, where 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 examples of Figures 2 and 3. Despite the algorithm efficiency, future work should be done to develop new sampling techniques for regression with mixture models, mainly to improve computation cost.

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 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.

5 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

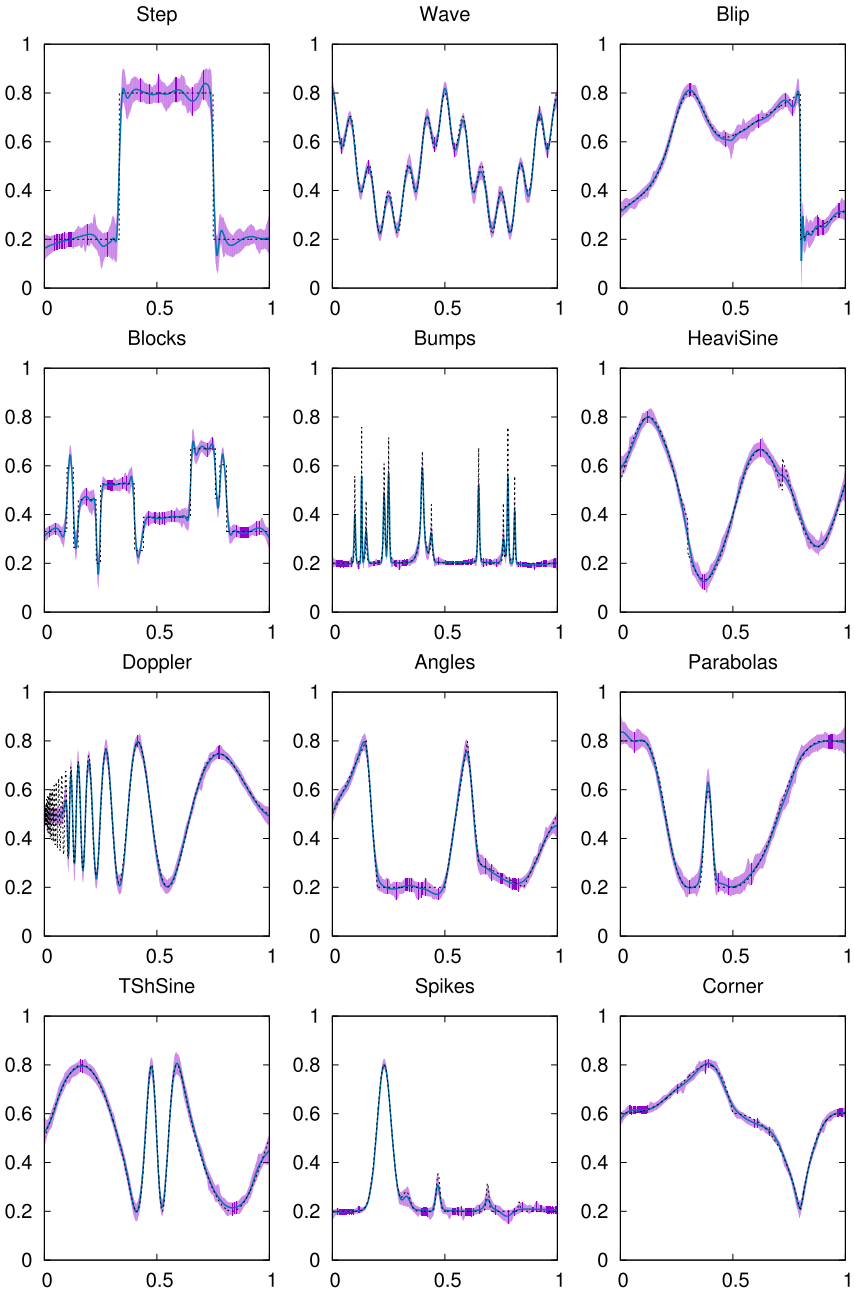

Figure 2: Example of simulation results using location-scale mixtures of Gaussians. The root signal-to-noise ratio is equal to 3 for sample size of 1024 design points. The true regression function is represented with dashes, the mean of the sampled posterior distribution in blue and sampled 95% credible bands in pink.

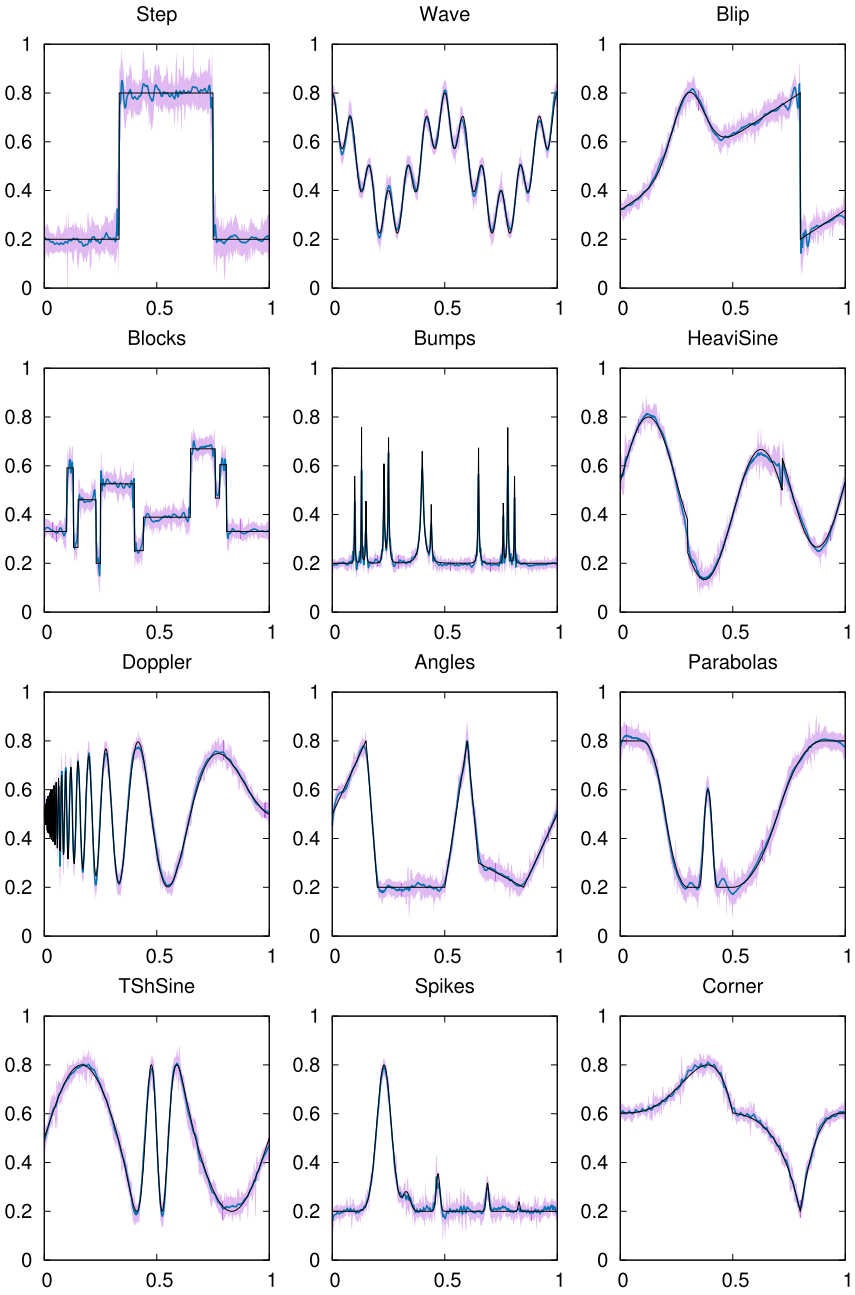

Figure 3: Example of simulation results using location-scale mixtures of Symmlet8. The root signal-to-noise ratio is equal to 3 for sample size of 1024 design points. The true regression function is represented with dashes, the mean of the sampled posterior distribution in blue and sampled 95% credible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

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,

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

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 (Easley et al., 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 Figure 4.

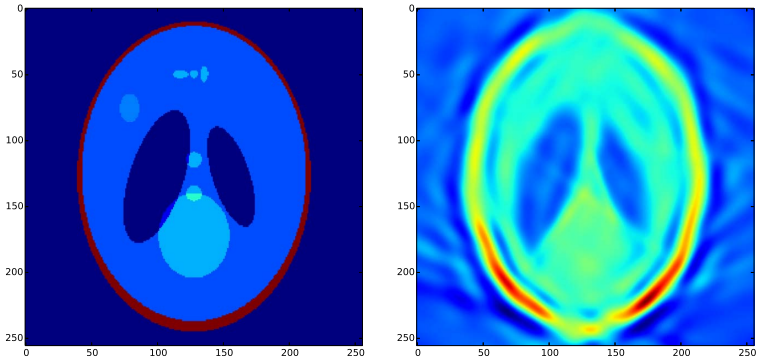

Figure 4: Simulation of X-ray computed tomography imaging using symmetric Gamma process location-scale mixture of Gaussians. On the left: the original image. On the right: the reconstructed image from 32768 observations of the Radon transform of the original image in a Gaussian noise.

The choice of a Gaussian kernel for the mixture is convenient theoretically since it allows to compute the likelihood analytically. In this example, however, the original image is rough, discontinuous and almost everywhere flat. Thus we cannot expect to represent it efficiently with a sparse numbers of Gaussians kernels. Since the sparsity of the mixture is governed a priori by the parameter α, using Gaussian kernels requires to increase α to a high value to get a reasonable posterior estimate, but, it automatically increases the computation cost and tends to reduce the sampling efficiency. Hence we believe mixtures should be used when there is a strong a priori on the regression function that can guide the choice of the kernel. In other words, mixtures are useful when we know a priori that the regression function has sparse representation in term of linear combinations of the kernels. This is often the case for many kernels and both location-scale or location-modulation mixtures when the regression function has Hölder smoothness, see for instance the results in Sections S2.6 and S3.3.

6 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.

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,

$$\begin{aligned} Y_i \mid \epsilon_i &= f(x_i)+\epsilon_i, \quad i=1,\dots,n,\\ \epsilon_1,\dots,\epsilon_n \mid \sigma^2 &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0,\sigma^2), \quad \text{independently of } (f,\sigma),\\ (f,\sigma) &\sim \Pi, \end{aligned} \tag{4}$$

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.

Let $P_{\theta,i}$ denote the distribution of Y_i under the parameter $\theta=(f,\sigma)$, P_θ^n denote 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

$$\lim_{n\to\infty}\Pi\left(\left\{\theta\in\Theta:\rho_n(\theta,\theta_0)>M\epsilon_n\right\}\mid Y_1,\dots,Y_n\right)=0 \quad P_{\theta_0}^\infty\text{-a.s.} \tag{5}$$

Regarding the model of (4), with deterministic covariates $x_1,\dots,x_n$ arbitrary spread in $[-S,S]^d$, we have the following theorem for location-scale mixtures. 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 3. *Suppose that $f_0 \in \mathcal{C}^\beta[-S,S]^d$ for some $S>0$. Under assumptions on the base measure of the Gamma process and on the mother function g, there exist constants $\kappa, t > 0$ such that (5) holds for the location-scale prior with $\epsilon_n^2 = n^{-2\beta/(2\beta+d+\kappa/2)}(\log n)^t$.*

The hypotheses on the Gamma process and the mother function in the Theorem 3 are deliberately passed over in silence in order to keep the statement of the result short. Nevertheless, the mentioned assumptions are fairly standard and given in details in the proof of the theorem, in Section S2.

Theorem 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, known to be $n^{-2\beta/(2\beta+d)}$ (up to a power of $\log n$ term). Indeed, if one takes $\alpha F = \alpha F_A \times F_\mu$ with F_A the Inverse-Wishart distribution, then $\kappa = 2$ in the theorem. We can achieve $\kappa = 1$ with F_A taken as 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 for location-modulation mixtures relies 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. Proof and details of the assumptions can be found in Section S3.

Theorem 4. *Suppose that $f_0 \in \mathcal{C}^\beta[-S,S]^d$ for some $S>0$. Under assumptions on the base measure of the Gamma process and on the mother function g, there exists a constant $t > 0$ such that (5) holds for the location-modulation prior with $\epsilon_n^2 = n^{-2\beta/(2\beta+d)}(\log n)^t$.*

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 $\beta > 0$.

Supplementary Material

Some aspects of symmetric Gamma process mixtures: Supplementary material (DOI: 10.1214/17-BA1058SUPP; .pdf).

References

Abramovich, F., Sapatinas, T., and Silverman, B. (2000). "Stochastic expansions in an overcomplete wavelet dictionary." *Probability Theory and Related Fields*, 117(1): 133–144. MR1759511. doi: https://doi.org/10.1007/s004400050268. 703

Antoniadis, A., Bigot, J., and Sapatinas, T. (2001). "Wavelet estimators in nonparametric regression: a comparative simulation study." *Journal of Statistical Software*, 6: 1–83. URL http://hal.archives-ouvertes.fr/hal-00823485/. 709, 711, 712

Barndorff-Nielsen, O. E. and Schmiegel, J. (2004). "Lévy-based spatial-temporal modelling, with applications to turbulence." *Russian Mathematical Surveys*, 59(1): 65. URL http://stacks.iop.org/0036-0279/59/i=1/a=R06. 705

Canale, A. and De Blasi, P. (2017). "Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation." *Bernoulli*, 23(1): 379–404. MR3556776. doi: https://doi.org/10.3150/15-BEJ746. 703, 717

De Blasi, P., Peccati, G., and Prünster, I. (2009). "Asymptotics for posterior hazards." *Annals of Statistics*, 37(4): 1906–1945. MR2533475. doi: https://doi.org/10.1214/08-AOS631. 703

de Jonge, R. and van Zanten, J. H. (2010). "Adaptive nonparametric Bayesian inference using location-scale mixture priors." *Annals of Statistics*, 38(6): 3300–3320. MR2766853. doi: https://doi.org/10.1214/10-AOS811. 703, 716

Easley, G. R., Colonna, F., and Labate, D. (2009). "Improved radon based imaging using the shearlet transform." http://dx.doi.org/10.1117/12.820066 715

Escobar, M. D. and West, M. (1994). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 703

Favaro, S., Guglielmi, A., and Walker, S. G. (2012). "A class of measure-valued Markov chains and Bayesian nonparametrics." *Bernoulli*, 18(3): 1002–1030. MR2948910. doi: https://doi.org/10.3150/11-BEJ356. 706, 707

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to calculating posterior moments." In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, J. F. M. (eds.), *Bayesian Statistics 4*, 169–193. Oxford University Press. MR1380276. 709

Ghosal, S. and van der Vaart, A. (2007). "Convergence rates of posterior distributions for noniid observations." *Annals of Statistics*, 35(1): 192–223. MR2332274. doi: https://doi.org/10.1214/009053606000001172. 703

Ishwaran, H. and James, L. F. (2004). "Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data." *Journal of the American Statistical Association*, 99(465): 175–190. MR2054297. doi: https://doi.org/10.1198/016214504000000179. 703

Kingman, J. F. C. (1992). *Poisson processes*, volume 3. Oxford university press. MR1207584. 706

Lijoi, A. and Nipoti, B. (2014). "A Class of Hazard Rate Mixtures for Combining Survival Data From Different Experiments." *Journal of the American Statistical Association*, 109(506): 802–814. MR3223751. doi: https://doi.org/10.1080/01621459.2013.869499. 703

Lo, A. Y. and Weng, C.-S. (1989). "On a class of Bayesian nonparametric estimates: II. Hazard rate estimates." *Annals of the Institute of Statistical Mathematics*, 41(2): 227–245. MR1006487. doi: https://doi.org/10.1007/BF00049393. 703

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H., and Patil, P. (1998). "Exact Risk Analysis of Wavelet Regression." *Journal of Computational and Graphical Statistics*, 7(3): 278–309. URL http://amstat.tandfonline.com/doi/abs/10.1080/10618600.1998.10474777. 709, 710, 711

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures." *Biometrika*, 83(1): 67–79. URL http://biomet.oxfordjournals.org/content/83/1/67.abstract. MR1399156. doi: https://doi.org/10.1093/biomet/83.1.67. 703

Naulet Z. and Barat É. (2017). "Some aspects of symmetric Gamma process mixtures: Supplementary material." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1058SUPP. 705

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. URL http://amstat.tandfonline.com/doi/abs/10.1080/10618600.2000.10474879. MR1823804. doi: https://doi.org/10.2307/1390653. 707

Peccati, G. and Prünster, I. (2008). "Linear and quadratic functionals of random hazard rates: An asymptotic analysis." *Annals of Applied Probability*, 18(5): 1910–1943. MR2462554. doi: https://doi.org/10.1214/07-AAP509. 703

Rajput, B. S. and Rosinski, J. (1989). "Spectral representations of infinitely divisible processes." *Probability Theory and Related Fields*, 82(3): 451–487. MR1001524. doi: https://doi.org/10.1007/BF00339998. 705, 706

Rudin, W. (1974). *Real and Complex Analysis, 1966*. McGraw-Hill, New York. MR0210528. 706

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures." *Biometrika*, 100(3): 623–640. URL http://biomet.oxfordjournals.org/content/100/3/623.abstract. MR3094441. doi: https://doi.org/10.1093/biomet/ast015. 703, 717

Shepp, L. A. and Logan, B. F. (1974). "The Fourier reconstruction of a head section." *IEEE Transactions on Nuclear Science*, 21(3): 21–43. 715

West, M. (1992). *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper #92-A03. 709

Wolpert, R. L., Clyde, M. A., and Tu, C. (2011). "Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels." *Annals of Statistics*, 39(4): 1916–1962. MR2893857. doi: https://doi.org/10.1214/11-AOS889. 703, 704, 705, 706, 711

Acknowledgments

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. Part of this work has been supported by the BNPSI ANR project no ANR-13-BS-03-0006-01.