Submitted to Bernoulli

A Bayesian nonparametric approach to log-concave density estimation

ESTER MARIUCCI[1], KOLYAN RAY[2] and BOTOND SZABÓ[2]

[1]*Institut für Mathematik, Universität Potsdam, Karl-Liebknecht-Str. 24-25, Potsdam 14476, Germany. E-mail:* mariucci@uni-potsdam.de

[2]*Mathematical Institute, Leiden University, Niels Bohrweg 1, Leiden 2333 CA, Netherlands. E-mail:* k.m.ray@math.leidenuniv.nl*;* b.t.szabo@math.leidenuniv.nl*,*

The estimation of a log-concave density on $\mathbb{R}$ is a canonical problem in the area of shape-constrained nonparametric inference. We present a Bayesian nonparametric approach to this problem based on an exponentiated Dirichlet process mixture prior and show that the posterior distribution converges to the log-concave truth at the (near-) minimax rate in Hellinger distance. Our proof proceeds by establishing a general contraction result based on the log-concave maximum likelihood estimator that prevents the need for further metric entropy calculations. We further present computationally more feasible approximations and both an empirical and hierarchical Bayes approach. All priors are illustrated numerically via simulations.

Keywords: Density estimation, log-concavity, Dirichlet mixture, posterior distribution, convergence rate, nonparametric hypothesis testing.

1. Introduction

Nonparametric shape constraints offer practitioners considerable modelling flexibility by providing infinite-dimensional families that cover a wide range of parameters whilst also including numerous common parametric families. Log-concave densities on $\mathbb{R}$, that is densities whose logarithm is a concave function taking values in $[-\infty, \infty)$, constitute a particularly important shape-constrained class. This class includes many well-known parametric densities that are frequently used in statistical modelling, including the Gaussian, uniform, Laplace, Gumbel, logistic, gamma distributions with shape parameter at least one, Beta(α, β) distributions with $\alpha, \beta \geq 1$ and Weibull distributions with parameter at least one.

One of the original statistical motivations for considering log-concave density estimation was the problem of estimating a unimodal density with unknown mode. While this is a natural constraint in many applications, the nonparametric MLE over this class does not exist [3]. Since the class of log-concave densities equals the class of strongly unimodal densities [19], Walther [36] argues that this class provides a natural alternative to the full set of all unimodal densities. The class of log-concave densities also preserves many of the attractive properties of Gaussian distributions, such as closure under convolution, marginalization, conditioning and taking products. One can therefore

view log-concave densities as a natural infinite-dimensional surrogate for Gaussians that retain many of their important features yet allow substantially more freedom, such as heavier tails. For these reasons, estimation of log-concave densities has received significant attention in recent years, particularly concerning the performance of the log-concave MLE [9, 30, 4, 5, 10, 8, 22, 21].

Outside density estimation, log-concavity as a modelling assumption has found applications in many statistical problems, such as mixture models [36, 1], tail index estimation [25], clustering [5], regression [10] and independent component analysis [29]. For general reviews of inference with log-concave distributions and estimation under shape constraints, see [37] and [15] respectively.

The Bayesian approach provides a natural way to encode shape constraints via the prior distribution, for instance under monotonicity [31, 20, 28, 27] or convexity contraints [32, 17, 16]. We present here a Bayesian nonparametric method for log-concave density estimation on $\mathbb{R}$ based on an exponentiated Dirichlet process mixture prior, which we show converges to a log-concave truth in Hellinger distance at the (near-)minimax rate. To the best of our knowledge, this is the first Bayesian nonparametric approach to this problem. We also study two computationally motivated approximations to the full Dirichlet process mixture based on standard Dirichlet process approximations, namely the Dirichlet multinomial distribution and truncating the stick-breaking representation (see Chapter 4.3.3 of [13]). We further propose both an empirical and hierarchical Bayes approach that have clear practical advantages, while behaving similarly to the above in simulations. All of these priors are easily implementable using a random walk Metropolis-Hastings within Gibbs sampling algorithm, which we illustrate in Section 3.

An advantage of the Bayesian method is that point estimates and credible sets can be approximately computed as soon as one is able to sample from the posterior distribution. In particular, the posterior yields easy access to statements on Bayesian uncertainty quantification as we show numerically in Section 3. Our numerical results suggest that pointwise credible sets have reasonable coverage at moderate sample sizes.

The Bayesian approach also permits inference about multiple quantities, such as functionals, in a unified way using the posterior distribution. A particular functional of interest is the mode of a log-concave density. While the pointwise limiting distribution of the log-concave MLE is known [2], it depends in a complicated way on the unknown density making it difficult to use to construct a confidence interval for the mode. An alternative approach to constructing a confidence interval based on comparing the log-concave MLE with the mode constrained MLE has recently been proposed [6]. For the Bayesian, the marginal posterior of the mode provides a natural approach to both estimation and uncertainty quantification. Indeed, it is easy to construct Bayesian credible intervals as we demonstrate numerically in Section 3. Whether such an approach is theoretically justified from a frequentist perspective is a subtle question related to the semiparametric Bernstein-von Mises phenomenon (Chapter 12 of [13]) that is, however, beyond the scope of this article. We also note that other constraints, such as a known mode [7], can similarly be enforced through suitable prior calibration.

Given the good performance of the log-concave MLE, one might expect that Bayesian procedures, being driven by the likelihood, behave similarly well. This is indeed the case,

as we show below. Our proof relies on the classic testing approach of Ghosal et al. [11] with interesting modifications in the log-concave setting. The existence and optimality of the MLE in Hellinger distance is closely linked to a uniform control of bracketing entropy [34]. In our setting, one can exploit the affine equivariance of the log-concave MLE (Remark 2.4 of [10]) to circumvent the need to control the metric entropy of the whole space by reducing the problem to studying a subset satisfying restrictions on the first two moments of the underlying density. This is a substantial reduction, since obtaining sharp entropy bounds in even this reduced case is highly technical, see Theorem 4 of Kim and Samworth [22]. One can then use the MLE to construct suitable plug-in tests with exponentially decaying type-II errors as in Giné and Nickl [14] that take full advantage of the extra structure of the problem compared to the standard Le Cam-Birgé testing theory for the Hellinger distance [23]. Indeed, a naive attempt to control the entropy directly, as is standard in the Bayesian nonparametrics literature (e.g. [11]), results in an overly small set on which the prior must place most of its mass. This leads to unnecessary restrictions on the prior, which in particular are not satisfied by the priors we consider in Section 2, see Remark 1. Beyond this, there remain significant technical hurdles to proving that the prior places sufficient mass in a Kullback-Leibler neighbourhood of the truth, in particular related to the approximation of log-concave densities using piecewise log-linear functions with suitably spaced knots.

The paper is structured as follows. In Section 2 we introduce our priors and present our main results, both on general contraction for log-concave densities and for the specific priors considered here. In Section 3 we present a simulation study, including a more practical empirical Bayes implementation, with some discussion in Section 4. In Section 5 we present the proofs of the main results with technical results placed in Section 6. The proofs of certain technical lemmas and some additional simulations can be found in the supplementary material [24].

Notation: For two probability densities p and q with respect to Lebesgue measure λ on $\mathbb{R}$, we write $h^2(p,q) = \int(\sqrt{p}-\sqrt{q})^2$ for the squared Hellinger distance, $K(p,q) = \int p \log \frac{p}{q}$ for the Kullback-Leibler divergence and $V = \int p(\log \frac{p}{q})^2$. We denote by $P_{f_0}^n$ the product probability measure corresponding to the joint distribution of i.i.d random variables $X_1, \ldots, X_n$ with density f_0 and write $P_{f_0} = P_{f_0}^1$. For a function w, we denote by w'_- and w'_+ its left and right derivatives respectively, that is

$$w'_-(x) = \lim_{s \nearrow x} w'(s) \quad \text{and} \quad w'_+(x) = \lim_{s \searrow x} w'(s).$$

Let $\mathbb{R}^+ = [0, \infty)$ and for two real numbers a, b, let $a \wedge b$ and $a \vee b$ denote the minimum and maximum of a and b respectively. Finally, the symbols $\lesssim$ and $\gtrsim$ stand for an inequality up to a constant multiple, where the constant is universal or (at least) unimportant for our purposes.

2. Main Results

Consider i.i.d. density estimation, where we observe $X_1, ..., X_n \sim f_0$ with $f_0 = e^{w_0}$ an unknown log-concave density to be estimated. Let $\mathcal{F}$ denote the class of upper semi-

continuous log-concave probability densities on $\mathbb{R}$. For $\alpha > 0$ and $\beta \in \mathbb{R}$, denote

$$\mathcal{F}_{\alpha,\beta} := \{f \in \mathcal{F} : f(x) \leq e^{\beta - \alpha|x|} \ \forall x \in \mathbb{R}\}.$$

By Lemma 1 of Cule and Samworth [4], for any log-concave density f_0 there exist constants $\alpha_{f_0} > 0$ and $\beta_{f_0} \in \mathbb{R}$ such that $f_0(x) \leq e^{\beta_{f_0} - \alpha_{f_0}|x|}$ for all $x \in \mathbb{R}$. Consequently, any upper semi-continuous log-concave density f_0 belongs to $\mathcal{F}_{\alpha,\beta}$ for $0 < \alpha \leq \alpha_{f_0}$ and $\beta \geq \beta_{f_0}$.

We establish a general posterior contraction theorem for priors on log-concave densities using the general testing approach introduced in [11], which requires the construction of suitable tests with exponentially decaying type-II errors. We construct plug-in tests based on the concentration properties of the log-concave MLE, similar to the linear estimators considered in [14, 26]. The MLE has been shown to converge to the truth at the minimax rate in Hellinger distance in Kim and Samworth [22] and the following theorem relies heavily on their result.

Theorem 1. *Let $\mathcal{F}$ denote the set of upper semi-continuous, log-concave probability densities on $\mathbb{R}$ and let Π_n be a sequence of priors supported on $\mathcal{F}$. Consider a sequence $\varepsilon_n \to 0$ such that $n^{-2/5} \lesssim \varepsilon_n \lesssim n^{-3/8-\rho}$ for some $\rho > 0$ and suppose there exists a constant $C > 0$ such that*

$$\Pi_n\Big(f \in \mathcal{F} : \int_{\mathbb{R}} f_0\Big(\log \frac{f_0}{f}\Big) \leq \varepsilon_n^2, \quad \int_{\mathbb{R}} f_0\Big(\log \frac{f_0}{f}\Big)^2 \leq \varepsilon_n^2\Big) \geq \exp(-Cn\varepsilon_n^2). \tag{1}$$

Then for sufficiently large M,

$$\Pi_n(f \in \mathcal{F} : h(f, f_0) \geq M\varepsilon_n | X_1, ..., X_n) \to 0$$

in $P_{f_0}^n$-probability as $n \to \infty$.

The upper bound $\varepsilon_n \lesssim n^{-3/8-\rho}$ is an artefact of the proof arising from the exponential inequality for the log-concave MLE that we use to construct our tests, see Lemma 1. Since our interest lies in obtaining the optimal rate $n^{-2/5}$, possibly up to logarithmic factors, it plays no further role in our results. It is typical in Bayesian nonparametrics to require metric entropy conditions, which come from piecing together tests for Hellinger balls into tests for the complements of balls, see for instance Theorem 7.1 of [11]. The lack of such a condition in Theorem 1 is tied to the optimality and specific structure of the log-concave MLE. Using the affine equivariance of the MLE (Remark 2.4 of [10]), one can reduce the testing problem to considering alternatives in the class $\mathcal{F}$ restricted to have zero mean and unit variance. Unlike the whole space $\mathcal{F}$, the bracketing Hellinger entropy of this latter set can be suitably controlled, thereby avoiding the need for additional entropy bounds.

Remark 1. *Obtaining sharp entropy bounds for log-concave function classes is a highly technical task and such bounds are only available for certain restricted subsets. Even in*

the case of mean and variance restrictions (Theorem 4 of [22]) and compactly supported and bounded densities (Proposition 14 of [21]), the proofs are lengthy and require substantial effort. To use such bounds for the classic entropy-based approach to prove posterior contraction would therefore require the prior to place most of its mass on the above types of restricted sets. For instance, the prior might be required to place all but exponentially small probability on $\mathcal{F}_{\alpha,\beta}$ for some given $\alpha > 0$, $\beta \in \mathbb{R}$. Such a prior construction is undesirable in practice and in fact none of our proposed priors satisfy such a restriction.

We now introduce a prior on log-concave densities based on an exponentiated Dirichlet process mixture. For any measurable function $w : \mathbb{R} \to \mathbb{R}$, define the density

$$f_w(x) = \frac{e^{w(x)}}{\int_{\mathbb{R}} e^{w(y)} dy}, \tag{2}$$

which is well-defined if $\int_{\mathbb{R}} e^{w(y)} dy < \infty$. Recall that any monotone non-increasing probability density on $\mathbb{R}^+$ has a mixture representation [38]

$$f(x) = \int_x^\infty \frac{1}{u} dP(u),$$

where P is a probability measure on $\mathbb{R}^+$. Khazaei and Rousseau [20] and Salomond [28] used the above representation to obtain a Bayesian nonparametric prior for monotone non-increasing densities. Unfortunately, such a convenient mixture representation is unavailable for log-concave densities and so the prior construction is somewhat more involved. Integrating the right-hand side of the last display, we obtain a function $w : \mathbb{R}^+ \to \mathbb{R}$ as follows:

$$w(x) = \gamma_1 \int_0^\infty \frac{u \wedge x}{u} dP(u) - \gamma_2 x,$$

where $\gamma_1 > 0$, $\gamma_2 \in \mathbb{R}$ and P is a probability measure on $[0, \infty)$. Since its (left and right) derivative is monotone decreasing, w is concave. While not every concave function can be represented in this way, any log-concave density on $[0, \infty)$ can be approximated arbitrary well in Hellinger distance by a function of the form $e^w/(\int e^w)$, where w is as above with P a discrete probability measure, see Proposition 1. Translating the above thus gives a natural representation for a prior construction for log-concave densities on $\mathbb{R}$.

Consider therefore the following possibly n-dependent prior on the log-density $w : [a_n, b_n] \to \mathbb{R}$, where possibly $a_n \to -\infty$ and $b_n \to \infty$:

$$W(x) = \gamma_1 \int_0^{b_n - a_n} \frac{u \wedge (x - a_n)}{u} dP(u) - \gamma_2 (x - a_n), \tag{3}$$

where

- $P \sim DP(H \mathbb{1}_{[0, b_n - a_n]})$, the Dirichlet process with base measure $H \mathbb{1}_{[0, b_n - a_n]} = H(\mathbb{R}^+) \bar{H} \mathbb{1}_{[0, b_n - a_n]}$, where $0 < H(\mathbb{R}^+) < \infty$, $\bar{H}$ is a probability measure on $\mathbb{R}^+$ and every subset $U \subset [0, b_n - a_n]$ satisfies $H(U) \gtrsim \lambda(U)/(b_n - a_n)^\eta$ for some $\eta \geq 0$,

- $\gamma_i \sim p_{\gamma_i}$, $i = 1, 2$, where p_{γ_1}, p_{γ_2} are probability densities on $[0, \infty)$ and $\mathbb{R}$ respectively, satisfying $p_{\gamma_i}(|x|) \gtrsim e^{-c_i x^{1/4}}$, $c_i > 0$, for all $x \in [0, \infty)$ and $x \in \mathbb{R}$ respectively,
- γ_1, γ_2, and P are independent.

We denote by Π_n the full prior induced by f_W, where W is drawn as above. Some typical draws from the prior are plotted in Figure 1.

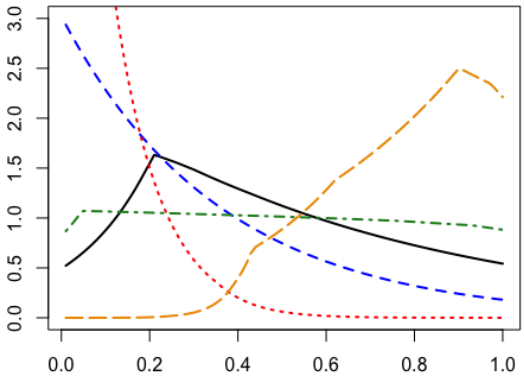

Figure 1. Prior draws with $[a_n, b_n] = [0, 1]$, $\gamma_1 \sim \text{Cauchy}_+(0, 1)$, $\gamma_2 \sim \text{Cauchy}(0, 1)$, $H = U(0, 1)$ and using the stick breaking construction.

Remark 2. *If $(b_n - a_n)$ grows polynomially in n, then H must have polynomial tails. On the other hand, if $(b_n - a_n)$ grows more slowly than any polynomial, one can relax this condition. For instance, if H has a density h with respect to the Lebesgue measure, then it is sufficient that $\min_{t \in [0, b_n - a_n]} h(t) \gtrsim n^{-\lambda}$ for some $\lambda > 0$. In particular, if $(b_n - a_n) \lesssim \log n$, then h may have exponential tails.*

We comment on several aspects of our prior. Firstly, since Dirichlet process draws are atomic with probability one, the prior draws (3) will be piecewise linear and concave. Moreover, we could add any concave function to (3), such as an $-\gamma_3 x^2$-type term, and still have a suitable concave prior. This permits greater modelling flexibility but complicates computation. In any case, the prior described above gives optimal contraction rates and can be computed in practice, so we restrict our attention to it. Another point to note is that if $(b_n - a_n) \to \infty$ and H is supported on the whole of $\mathbb{R}^+$, then the Dirichlet process base measure has total mass $H(\mathbb{R}^+)\bar{H}([a_n, b_n]) \leq H(\mathbb{R}^+)$ for fixed n. This has

the interpretation of assigning the prior more weight as $n \to \infty$, up to the full prior weight $H(\mathbb{R}^+)$. An alternative would be to re-weight the base measure to have full mass $H(\mathbb{R}^+)$ to give it equal weight for all n. This plays no role asymptotically and so we restrict to the first case for technical convenience.

A potentially more serious issue is that for fixed n, the support of the prior draws may not contain the support of the true density f_0, in which case observations outside $[a_n, b_n]$ cause the likelihood to be identically zero. While this is not a problem for n large enough if $-a_n, b_n \to \infty$ fast enough (see Theorem 3), it can be an issue for finite n. In practice, if one has an idea of the support of f_0, it is enough to select $[a_n, b_n]$ large enough to contain $\mathrm{supp}(f_0)$. A more pragmatic solution is to use an empirical Bayes approach and make the prior data-dependent by setting $a_n := X_{(1)}$, $b_n := X_{(n)}$ the first and last order statistics. This ensures that the likelihood is never zero and the posterior is always well-defined. Indeed, the MLE is supported on $[X_{(1)}, X_{(n)}]$ and so this can be thought of as plugging-in an estimate of the approximate support based on the likelihood. Moreover, since this approach yields the smallest support $[a_n, b_n]$ with non-zero likelihood, it also brings computational advantages. In particular, it can prevent the need to simulate the posterior distribution on potentially very large regions of $\mathbb{R}$ where the posterior draws are essentially indistinguishable from zero. The empirical Bayes method behaves very similarly to the prior (3) in simulations and we would advocate this approach in practice.

We first present a contraction result when the true density f_0 has known compact support.

Theorem 2. *Let $f_0 \in \mathcal{F}_{\alpha,\beta}$ for some $\alpha > 0$, $\beta \in \mathbb{R}$ and suppose further that f_0 is compactly supported. Let $a_n \equiv a$ and $b_n \equiv b$ for all n and denote by $\Pi_n = \Pi$ the prior described above. If $supp(f_0) \subset [a, b]$, then*

$$\Pi(f : h(f, f_0) \geq M(\log n)n^{-2/5} \mid X_1, ..., X_n) \to 0$$

in $P_{f_0}^n$-probability for some $M = M(\alpha, \beta) > 0$.

If $\mathrm{supp}(f_0)$ is not contained in a compact set or is unknown, it suffices to let $-a_n, b_n \to \infty$ fast enough. A slightly stronger lower bound on the tail of p_{γ_1} is consequently required, depending on the size of $(b_n - a_n)$.

Theorem 3. *Let $f_0 \in \mathcal{F}_{\alpha,\beta}$ for some $\alpha > 0$, $\beta \in \mathbb{R}$ and let Π_n denote the prior described above with $-a_n, b_n \gg \log n$. Assume further that $(b_n - a_n) \lesssim n^{\mu/5}$ and that the prior density p_{γ_1} for γ_1 satisfies the stronger lower bound $p_{\gamma_1}(x) \gtrsim e^{-c_1 x^{1/(4+\mu)}}$ for some $0 \leq \mu \leq 2$. Then*

$$\Pi_n(f : h(f, f_0) \geq M\varepsilon_n \mid X_1, ..., X_n) \to 0$$

in $P_{f_0}^n$-probability for some $M = M(\alpha, \beta) > 0$ and

$$\varepsilon_n = \max\left((\log n)n^{-2/5}, (b_n - a_n)n^{-4/5}\right). \tag{4}$$

Theorem 2 follows immediately from Theorem 3 and so its proof is omitted. If $(b_n - a_n) = O((\log n)n^{2/5})$, then we obtain the minimax rate for log-concave density estimation in Theorem 3, up to a logarithmic factor. Since the Hellinger distance dominates the total variation distance, the above also implies posterior convergence in total variation at the same rate ε_n given in (4). We also note that all the above statements are proved uniformly over $f_0 \in \mathcal{F}_{\alpha,\beta}$.

The posterior mean, also considered in the simulation study, is not necessarily log-concave. Nevertheless one can construct log-concave density estimators by separately computing the posterior mean for each parameter $\theta, p, \gamma_1, \gamma_2$ and then constructing the corresponding log-concave density according to (2) and (3). Another approach is to take the smallest Hellinger ball accumulating, say, 50% of the posterior mass and sample an arbitrary log-concave density from that ball. It is straightforward to verify that both of these estimators achieve the minimax concentration rate (up to the same logarithmic factor).

It is also of interest to obtain a fully Bayesian procedure that does not require the user to define the support of the prior draws. We therefore consider a hierarchical prior where one places a prior on the end points a and b, now not necessarily depending on n. This method has the advantage of employing a prior that does not depend on the data, but is slightly more involved computationally than the simple empirical Bayes approach. Assign to (a, b) a prior supported on the open half-space $\{(a, b) : a < b\}$ that has a Lebesgue density $\pi(a, b)$ satisfying

$$\pi(a, b) \geq Ce^{-c_1|a|^q - c_2|b-a|^r} \qquad \text{for all } a < b \tag{5}$$

and some $c_1, c_2, C, q, r > 0$. Such a distribution can be easily constructed by first drawing $a \sim \pi_1$, where the Lebesgue density $\pi_1(a) \geq Ce^{-c|a|^q}$, and then independently drawing $(b-a)|a \sim \pi_2$, where π_2 is a Lebesgue density on $(0, \infty)$ satisfying $\pi_2(b-a) \geq Ce^{-c|b-a|^r}$. Conditionally on $(a_n, b_n) = (a, b)$, the prior is then exactly as above. This hierarchical construction leads to a fully Bayesian procedure that again contracts at the (near-)minimax rate.

Theorem 4. *Let $f_0 \in \mathcal{F}_{\alpha,\beta}$ for some $\alpha > 0$, $\beta \in \mathbb{R}$ and let Π_n denote the prior described above with hyperprior $(a, b) \sim \pi(a, b)$ satisfying (5). Then*

$$\Pi_n(f : h(f, f_0) \geq M(\log n)n^{-2/5} \mid X_1, ..., X_n) \to 0$$

in $P^n_{f_0}$-probability for some $M = M(\alpha, \beta) > 0$.

Dirichlet process mixture priors are popular in density estimation due to the conjugacy of the posterior distribution, thereby providing methods that are highly efficient computationally. However, due to the exponentiation (2), this conjugacy property no longer holds, resulting in a less attractive prior choice that brings computational challenges. In practice, it is common to use approximations of the Dirichlet process to speed up computations, see for instance Chapter 4.3.3 of [13].

We firstly consider the Dirichlet multinomial distribution as a replacement for the Dirichlet process in our prior. By the proof of Theorem 3, the underlying true log-concave density can be well approximated by a piecewise log-linear density with at most $N = Cn^{1/5}\log n$ knots, for some large enough constant $C > 0$. In view of this, it is reasonable to take N atoms in the distribution. The corresponding prior on log-concave densities then takes the form

$$\begin{aligned} \theta_i &\overset{iid}{\sim} \bar{H}\mathbb{1}_{[0,b_n-a_n]}, \quad \text{for } i = 1, ..., N, \\ p = (p_1, ...p_N) &\sim Dir(\alpha_1, ..., \alpha_N), \\ \gamma_i &\overset{iid}{\sim} p_{\gamma_i}, \quad i = 1, 2, \\ f_{\theta,p,\gamma_1,\gamma_2}(x) &= \frac{\exp\{\gamma_1 \sum_{i=1}^N \frac{\theta_i \wedge (x-a_n)}{\theta_i} p_i - \gamma_2(x-a_n)\}\mathbb{1}_{[a_n,b_n]}(x)}{\int_{a_n}^{b_n} \exp\{\gamma_1 \sum_{i=1}^N \frac{\theta_i \wedge (u-a_n)}{\theta_i} p_i - \gamma_2(u-a_n)\}du}, \end{aligned} \tag{6}$$

where α_i, $i = 1, ..., N$, are chosen such that $\alpha_i = \alpha/N$ for some arbitrary $0 < \alpha \leq H(\mathbb{R}^+)$.

An alternative choice for the mixing prior is to truncate the stick-breaking representation of the Dirichlet process at a fixed level. Similarly to the Dirichlet multinomial distribution, we truncate the stick-breaking process at level $N = Cn^{1/5}\log n$, resulting in the same hierarchical prior as in (6) with the only difference being that the distribution of p in the N-simplex is given by

$$p_i \sim V_i \prod_{j=1}^{i-1}(1 - V_j), \text{ where } V_i \sim \text{Beta}(1, H(\mathbb{R}^+)), \; i = 1, ..., N-1. \tag{7}$$

Both of these computationally more efficient approximations have the same theoretical guarantees as the full exponentiated Dirichlet process prior Π_n or its hierarchical Bayes equivalent.

Corollary 1. *Let $f_0 \in \mathcal{F}_{\alpha,\beta}$ for some $\alpha > 0$, $\beta \in \mathbb{R}$ and let Π_n' denote either the prior (6) or (7). If $-a_n, b_n \gg \log n$, $(b_n - a_n) \lesssim n^{\mu/5}$ and the prior density p_{γ_1} for γ_1 satisfies the stronger lower bound $p_{\gamma_1}(x) \gtrsim e^{-c_1 x^{1/(4+\mu)}}$ for some $0 \leq \mu \leq 2$, then*

$$\Pi_n'(f : h(f, f_0) \geq M\varepsilon_n \mid X_1, ..., X_n) \to 0$$

in $P_{f_0}^n$-probability for some $M = M(\alpha, \beta) > 0$ and ε_n given by (4).

If we additionally assign a hyperprior $\pi(a, b)$ satisfying (5) to (a, b) and no longer require the strong lower bound for p_{γ_1}, then the above holds for the resulting posterior with $\varepsilon_n = (\log n)n^{-2/5}$.

The proofs of Theorem 3 and Corollary 1 establish the small-ball probability (1) by approximating a log-concave density in $\mathcal{F}_{\alpha,\beta}$ with a suitable piecewise log-linear density. This approximation requires several key properties, which make its construction non-standard and technically involved, and it may be of independent interest. The proof of Proposition 1 is deferred to Section 6.

Proposition 1. *Let $f_0 \in \mathcal{F}_{\alpha,\beta}$ and $([a_n, b_n])_n$ be a sequence of compact intervals such that $[-\frac{8}{5\alpha}\log n, \frac{8}{5\alpha}\log n] \subset [a_n, b_n]$ and $(b_n - a_n) = o(n^{4/5})$. For any $n \geq n_0$, where n_0 is an integer depending only on α and β, there exists a log-concave density $\bar{f}_n$ that is piecewise log-linear with $\bar{N} \leq C(\alpha,\beta)n^{1/5}\log n$ knots $z_1, ..., z_{\bar{N}} \in [0, b_n - a_n]$ satisfying the following properties:*

(i) $h^2(f_0, \bar{f}_n) \leq C(\alpha,\beta)[(\log n)^2 n^{-4/5} + (b_n - a_n)^2 n^{-8/5}]$,
(ii) $\{x \in \mathbb{R} : \bar{f}_n(x) > 0\} = [a_n, b_n]$,
(iii) *the knots are $cn^{-6/5}\log n$-separated for some universal constant $c > 0$,*
(iv) $f_0(x) \leq C(\alpha,\beta)\bar{f}_n(x)$ *for all* $x \in [a_n, b_n]$,
(v) *there exist* $\bar{\gamma}_1 \in [0, 2(b_n - a_n)n^{4/5}]$, $|\bar{\gamma}_2| \leq n^{4/5}$, $\bar{\gamma}_3 \in \mathbb{R}$ *and* $(\bar{p}_1, .., \bar{p}_{\bar{N}})$ *satisfying* $p_i \geq 0$ *and* $\sum_{i=1}^{\bar{N}} p_i = 1$, *such that*

$$\bar{f}_n(x) = \exp\left(\bar{\gamma}_1 \sum_{i=1}^{\bar{N}} \frac{z_i \wedge (x - a_n)}{z_i}\bar{p}_i - \bar{\gamma}_2(x - a_n) + \bar{\gamma}_3\right)\mathbb{1}_{[a_n,b_n]}(x).$$

It is relatively straightforward to establish an approximation of f_0 satisfying (i). However, approximating f_0 by $\bar{f}_n$ in a Kullback-Leibler type sense, as in (1), necessitates control of the support of $\bar{f}_n$ via (ii) and uniform control of the ratio $f_0/\bar{f}_n$ via (iv). The most difficult property to establish is the polynomial separation of the points in (iii). This is needed to ensure that the Dirichlet process prior simultaneously puts sufficient mass in a neighbourhood of each of the knots z_i, $i = 1, \ldots, \bar{N}$. Setting $[a_n, b_n] = [-\frac{8}{5\alpha}\log n, \frac{8}{5\alpha}\log n]$ yields the following corollary.

Corollary 2. *Let $f_0 \in \mathcal{F}_{\alpha,\beta}$. For any $n \geq n_0$, where n_0 is an integer depending only on α and β, there exists a log-concave density $\bar{f}_n$ supported on $[-\frac{8}{5\alpha}\log n, \frac{8}{5\alpha}\log n]$ that is piecewise log-linear with $O(n^{1/5}\log n)$ knots and satisfies $h^2(f_0, \bar{f}_n) \leq C(\alpha,\beta)(\log n)^2 n^{-4/5}$. Moreover, we may take the knots to be $cn^{-6/5}\log n$-separated for some universal constant $c > 0$.*

3. Simulation study

We present a simulation study to assess the performance of the proposed log-concave priors for density estimation. In particular, we investigate the prior based on the truncated stick breaking representation (7), firstly with deterministically chosen support $[a_n, b_n]$, secondly its empirical Bayes counterpart with support $[X_{(1)}, X_{(n)}]$, where $X_{(1)}$ and $X_{(n)}$ denote the smallest and largest observations, respectively, and thirdly the hierarchical Bayes version where the parameters a and $b - a$ are endowed with independent Cauchy and half-Cauchy distributions, respectively. In all cases we plot the posterior mean and 95% pointwise credible sets, and compare them with the log-concave maximum likelihood estimator (computed using the R function "mlelcd").

Consider first the posterior distribution arising from the prior with deterministic support $[a_n, b_n]$. We have drawn random samples of size $n = 50, 200, 500$ and 2500 from a

gamma distribution with shape and rate parameters 2 and 1, respectively. We took the number of linear pieces in the exponent of the prior to be $N = Cn^{1/5}\log n$, with $C = 1$, set $[a_n, b_n] = [-2.3\log n, 2.3\log n]$, endowed the break-point parameters $\theta = (\theta_1, ..., \theta_m)$ with independent uniform priors on $[0, b_n - a_n]$, assigned the weight parameters p a stick-breaking distribution truncated at level m, and endowed γ_1 and γ_2 a half Cauchy and a Cauchy distribution, respectively, with location parameter 0 and scale parameter 1. Since the posterior distribution does not have a closed-form expression, we drew approximate samples from the posterior using a random walk Metropolis-Hastings within Gibbs sampling algorithm for 10000 iterations out of which the first 5000 are discarded as burn-in period. In Figure 2, we have plotted the true distribution (solid red), posterior mean (solid blue), 95% pointwise credible band (dashed blue) and the maximum likelihood estimator (solid green). The data is represented by a histogram on the figures.

We see that the posterior mean gives an adequate estimator for the true log-concave density with similar, if not superior, performance compared to the more jagged maximum likelihood estimator, and the 95% pointwise credible bands mostly contain the true function except for points close to zero. We further investigate the frequentist coverage properties of the pointwise Bayesian credible sets. We repeat the above experiment for the empirical Bayes procedure 100 times (each with 2000 iterations out of which half were discarded as burn in) and report the frequencies where the density $f(x)$ at given points $x \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ is inside of the corresponding credible interval. We consider sample sizes $n = 50, 200$ and 500 and report the empirical coverage probabilities in Table 1. One can see that in this particular example we get quite reliable uncertainty quantification, especially for larger sample sizes. It should be noted, however, that the frequentist coverage of Bayesian credible sets is a delicate subject in nonparametric statistics, see for instance Szabó et al. [33], and is beyond the scope of this article.

Remark 3. *The constant C in the number $N = Cn^{1/5}\log n$ of knots can be chosen relatively freely from a theoretical point of view without affecting the convergence rate. In practice, however, larger C results in smaller bias, larger variance and increased computational cost. For the relatively large sample sizes we consider here (except perhaps $n = 50$), taking $C = 1$ already gives reasonable estimators, see Figure 2. The optimal choice of C depends on the unknown underlying density and one could experiment with selecting C in a data-driven manner, for example by estimating it empirically or endowing C with a prior. We think that $C = 1$ works sufficiently well for moderate sample sizes, while for small samples sizes one can take C slightly larger, say 2 or 3, to have enough knots.*

We next investigate the behaviour of the empirical and hierarchical Bayes versions of the proposed prior. We again simulate $n = 50, 200, 500$ and 2500 independent draws from a Gamma(2,1) distribution and set the compact support of the prior densities to be $[a_n, b_n] = [X_{(1)}, X_{(n)}]$, that is the smallest and largest observations, for the empirical Bayes procedure and $[a, b]$, with $a \sim Cauchy(0, 1)$ and $b - a \sim Cauchy_+(0, 1)$ independent, for the hierarchical Bayes method. As before we set $m = n^{1/5}\log n$ and endowed the parameters θ, p, γ_1 and γ_2 with the same priors as above. We ran the algorithm again

n \ x	0.5	1	1.5	2	2.5	3
50	0.51	0.81	0.88	0.85	0.87	0.93
200	0.68	0.97	0.94	0.87	0.94	0.97
500	0.83	0.96	0.88	0.87	0.86	0.91

Table 1. Frequencies out of 100 experiments when the empirical Bayes credible set contained the true function values at points $x = 0.5, 1, ..., 3$. From top to bottom the sample size increases from $n = 50$ until $n = 500$.

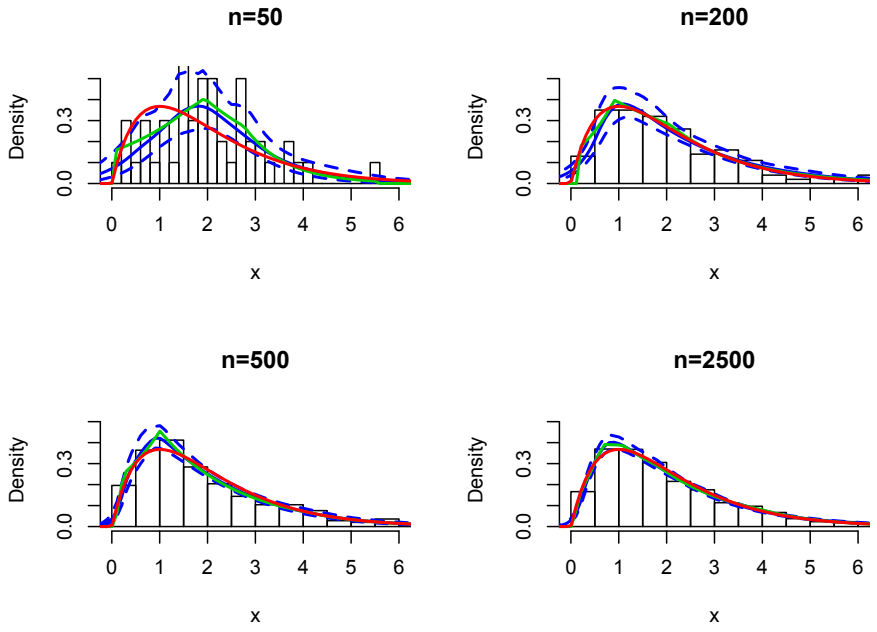

Figure 2. Prior with $[a_n, b_n]$ selected deterministically: the underlying Gamma(2,1) density function (red), posterior mean (solid blue), pointwise credible bands (dashed blue), maximum likelihood estimator (solid green) and data is represented with a histogram. We have increasing sample size from left to right and top to bottom $n = 50, 200, 500$ and 2500.

for 10000 iterations, taking the first half of the chain as a burn-in period. We plot the outcomes in Figures 3 and 4 for the empirical and hierarchical Bayes procedures, respectively. One can see that for $n \geq 500$ observations, the posterior mean (solid blue) closely resembles the underlying gamma density (solid red), while the fit is already reasonable for $n = 200$. The pointwise 95%-credible bands contain the true density, even near zero, which was problematic in case of the prior with support selected deterministically. Comparing Figures 2, 3 and 4, we see that the empirical and hierarchical Bayes approaches of selecting the support $[a_n, b_n]$ in a data-driven way outperform a deterministic selection.

We also note that the algorithm for the empirical Bayes method was considerably faster than the others due to the smaller support, which reduces the computation time of the normalizing constants $\int e^{w(y)}dy$ of the densities.

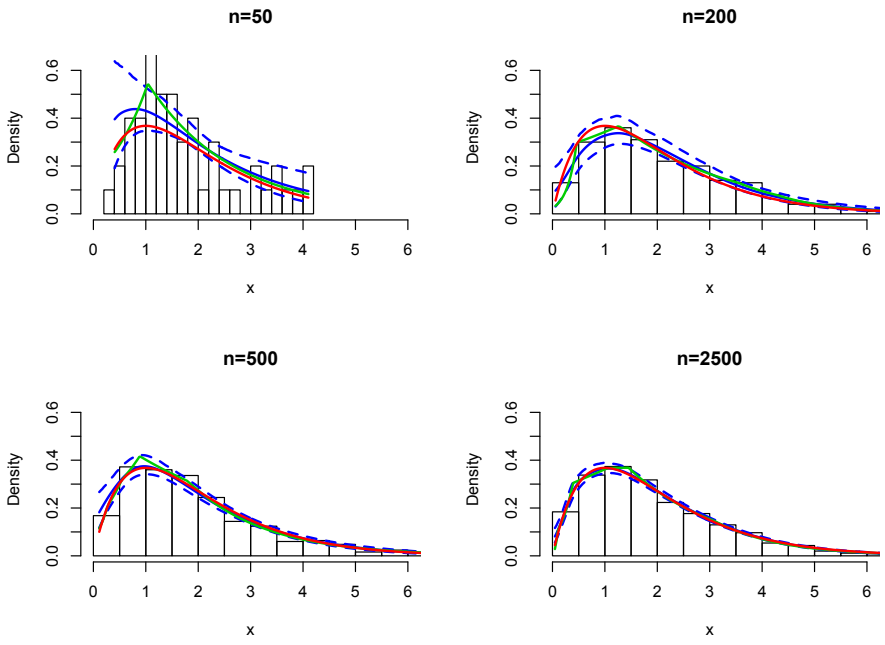

Figure 3. Empirical Bayes prior with data-driven support: the underlying Gamma(2,1) density function (red), posterior mean (solid blue), pointwise credible bands (dashed blue) and data is represented with a histogram. We have increasing sample size from left to right and top to bottom $n = 50, 200, 500$ and 2500.

We then investigate the performance of the posterior distribution corresponding to the empirical and hierarchical Bayes methods for recovering different log-concave densities and again compare them with the MLE. We have considered a standard normal distribution, a gamma distribution with shape parameter 2 and rate parameter 1, a beta distribution with shape parameters 2 and 3, and a Laplace distribution with location parameter 0 and dispersion parameter 1. In all four examples we have taken sample size $n = 1500$. The posterior mean (solid blue), the 95% pointwise credible bands (dashed blue) and the MLE (green) are plotted in Figures 5 and 6 for the empirical and hierarchical Bayes procedures, respectively. All four subpictures for both data-driven methods show satisfactory results, both for recovery using the posterior mean and for uncertainty quantification using the pointwise credible bands. We note that the displayed plots convey typical behaviour and are representative of multiple simulations. We hence draw the conclusion that the proposed method seems to work well in practice for various choices

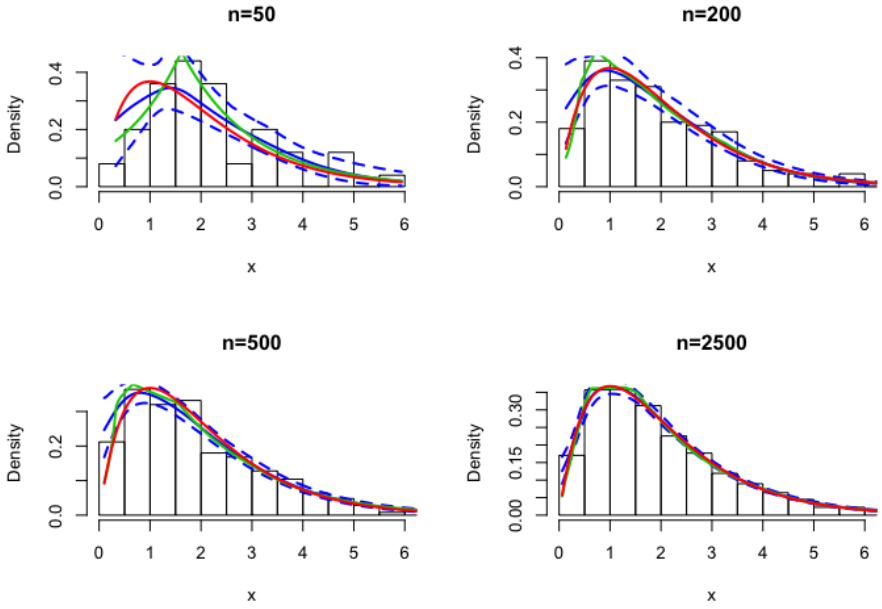

Figure 4. Hierarchical Bayes prior with data-driven support: the underlying Gamma(2,1) density function (red), posterior mean (solid blue), pointwise credible bands (dashed blue) and data is represented with a histogram. We have increasing sample size from left to right and top to bottom $n = 50, 200, 500$ and 2500.

of common log-concave densities.

Lastly, we investigate the performance of the proposed Bayesian procedures for estimating the mode of the underlying log-concave density. We consider the standard normal distribution and take i.i.d. random samples of size ranging from 50 to 20000. We run the Gibbs sampler for 20000 iterations and take the first half of the iterations as burn-in period. For each posterior draw we compute the mode and use the resulting histogram to approximate the one-dimensional marginal posterior. The histograms from the empirical Bayes procedure are displayed in Figure 7. One can see that the posterior concentrates around the true mode (i.e. 0) as the sample size increases.

The marginal posterior concentrates substantially slower than $n^{-1/2}$-rate. This is as expected, since the best possible minimax rate for estimating the mode m_0 of a unimodal or log-concave density f_0 satisfying $f_0''(m_0) < 0$ is $n^{-1/5}$, see [18, 2]. Indeed, the mode of the log-concave MLE attains this rate [2]. Interestingly, the marginal posterior does not seem to be Gaussian, which may be linked to the irregular asymptotic distribution of the mode of the log-concave MLE. This rather complicated distribution equals the mode of the second derivative of the lower invelope of a certain Gaussian process, see [2] for full

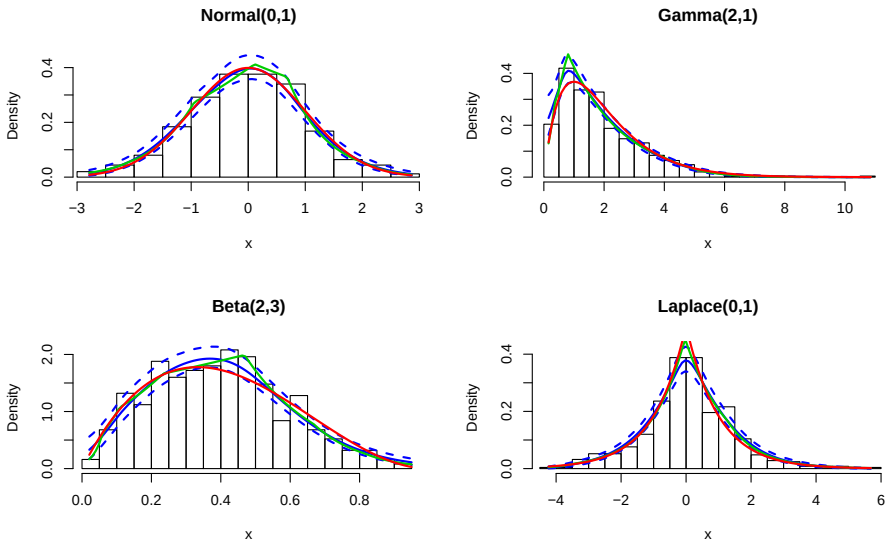

Figure 5. The underlying density function (red), empirical Bayes posterior mean (solid blue) and pointwise credible bands (dashed blue). The data is represented with a histogram. The true density functions are from left to right and top to bottom: standard Gaussian, Gamma(2,1), Beta(2,3) and Laplace(0,1).

details. A better understanding of the limiting shape of the marginal posterior would be interesting, but is beyond the scope of this article.

In the supplementary material we provide additional simulations for the marginal posterior for the mode from the empirical Bayes posterior for different underlying log-concave densities, namely the beta and gamma distributions. We also numerically investigate the applicability of our log-concave Bayesian prior for estimating mixtures of log-concave densities.

4. Discussion

We have proposed a novel Bayesian procedure for log-concave density estimation. The prior is defined on compactly supported densities, where the support can be chosen either deterministically, empirically or using a fully Bayesian hierarchical procedure. We have shown theoretically that both the deterministic and fully Bayesian choices of the support give (near-)optimal posterior contraction rates, and have demonstrated the good small sample performance of the posterior for all three methods in a simulation study. We have also plotted the 95% pointwise credible bands, which in our simulation study provide reliable frequentist uncertainty quantification. However, this might depend heavily on the choice of the underlying true density and it is unclear at present whether our methods

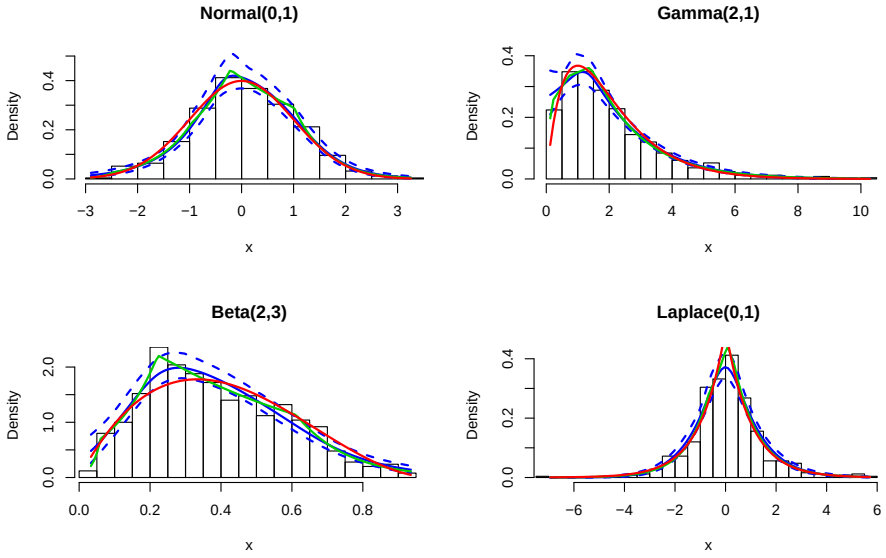

Figure 6. The underlying density function (red), hierarchical Bayes posterior mean (solid blue) and pointwise credible bands (dashed blue). The data is represented with a histogram. The true density functions are from left to right and top to bottom: standard Gaussian, Gamma(2,1), Beta(2,3) and Laplace(0,1).

generally provide trustworthy frequentist uncertainty quantification. The rigourous study of this question is beyond the scope of the present paper.

In our simulation study, we further investigated the behaviour of the marginal posterior for the mode functional. A natural next question is whether one can obtain semiparametric Bernstein-von Mises type results for the mode. In view of the irregular behaviour of the log-concave MLE, this is an interesting problem as it is unclear whether the limiting distribution of the posterior is indeed Gaussian.

A possible application of our proposed approach is clustering based on mixture models. Assuming that clusters have log-concave densities instead of (say) Gaussians broadens their modelling flexibility. We have executed a small simulation study to explore this direction. For simplicity we have considered a mixture of only two log-concave densities and modified our prior accordingly. In the considered examples (see the Supplementary material) our procedure performs reasonably well. However, we should note that the computational time is much worse than using simple Gaussian kernels. Extending this to mixtures with more than two (possibly unknown number of) components seems to be possible, but requires optimization of the Gibbs sampler and perhaps introducing other approximation steps, which are beyond the scope of the present paper.

Another natural question if whether one can extend these results to multivariate den-

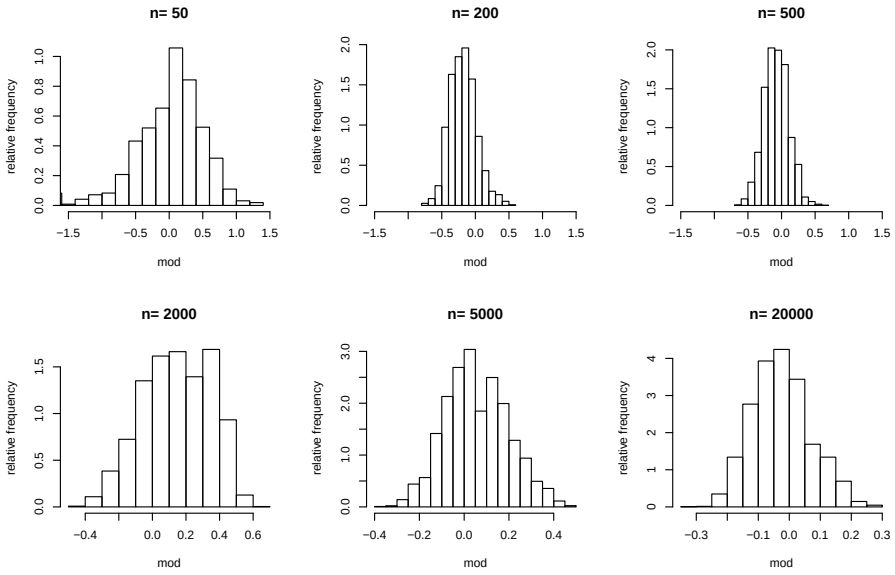

Figure 7. The empirical Bayes posterior distribution of the mode for standard normal distribution with increasing sample size from left to right and top to bottom, ranging between $n = 50$ and $n = 20000$. .

sity estimation, especially in view of the difficulty of computing the log-concave MLE in higher dimensions. Since our present prior construction is based on using a mixture representation to model a decreasing function, which corresponds to the derivative of the concave exponent, this will require new ideas. A possible approach is presented in Hannah and Dunson [17], who place a prior over all functions that are the maximum of a set of hyperplanes. This yields a prior on the set of convex functions that could potentially be adapted to the multivariate log-concave setting.

5. Proofs

Define the following classes of log-concave densities with mean and variance restrictions:

$$\overline{\mathcal{F}}^{\xi,\eta} = \left\{ f \in \mathcal{F} : \mu_f := \int x f(x) dx = \xi, \quad \sigma_f^2 := \int (x - \mu_f)^2 f(x) dx = \eta \right\}$$

and

$$\widetilde{\mathcal{F}}^{\xi,\eta} = \{ f \in \mathcal{F} : |\mu_f| \leq \xi, \quad |\sigma_f^2 - 1| \leq \eta \}.$$

Let $\hat{f}_n$ denote the log-concave MLE based on i.i.d. random variables $X_1, \ldots, X_n$ arising from a density $f_0 \in \mathcal{F}$.

The proof of Theorem 1 relies on a concentration inequality for the log-concave MLE based on data from moment-restricted densities. This is the content of the following lemma, whose proof is essentially contained in Kim and Samworth [22] for the more difficult case of general $d \geq 1$. However, we require a sharper probability bound than they provide and so make some minor modifications to their argument. The proof can be found in the supplementary material [24].

Lemma 1. *For every $\varepsilon > 0$, there exist positive constants L_0, C, c, n_0, depending only on ε, and positive universal constants $D, d > 0$, such that for all $L \geq L_0$ and $n \geq n_0$,*

$$\sup_{g_0 \in \overline{\mathcal{F}}^{0,1}} P_{g_0}^n \left(h(\hat{g}_n, g_0) \geq L n^{-2/5} \right) \leq C \exp\left(-c n^{1/(4+2\varepsilon)} \right) + D \exp\left(-d L^2 n^{1/5} \right),$$

where $\hat{g}_n$ denotes the log-concave maximum likelihood estimator based on an i.i.d. sample $Z_1, ..., Z_n$ from g_0.

Proof of Theorem 1. As in the proof of Theorem 2.1 of [11], using the lower bound on the small ball probability from (1), it suffices to construct tests $\phi_n = \phi_n(X_1, ..., X_n; f_0)$ such that

$$P_{f_0}^n \phi_n \to 0, \qquad \text{and} \qquad \sup_{f \in \mathcal{F}: h(f, f_0) \geq M\varepsilon_n} P_f^n (1 - \phi_n) \leq e^{-(C+4)n\varepsilon_n^2}$$

for n large enough, where the constant $C > 0$ matches that in (1).

For M_0 a constant to be chosen below, set $\phi_n = \mathbb{1}\{h(\hat{f}_n, f_0) \geq M_0 \varepsilon_n\}$, where $\hat{f}_n$ is the log-concave MLE based on i.i.d. observations $X_1, ..., X_n$ from a density $f_0 \in \mathcal{F}$. Let $\mu_{f_0} = \mathbb{E}X_i$, $\sigma_{f_0}^2 = \mathrm{Var}(X_i)$ and define $Z_i = (X_i - \mu_{f_0})/\sigma_{f_0}$, so that $\mathbb{E}Z_i = 0$ and $\mathrm{Var}(Z_i) = 1$. Further set $g_0(z) = \sigma_{f_0} f_0(\sigma_{f_0} z + \mu_{f_0})$ and $\hat{g}_n(z) = \sigma_{f_0} \hat{f}_n(\sigma_{f_0} z + \mu_{f_0})$, so $g_0 \in \overline{\mathcal{F}}^{0,1}$. By affine equivariance (Remark 2.4 of [10]), $\hat{g}_n$ is the log-concave maximum likelihood estimator of g_0 based on $Z_1, ..., Z_n$.

Using the invariance of the Hellinger distance under affine transformations and Lemma 1 with $\varepsilon = 1/2$, the type-I error satisfies

$$P_{f_0}^n \phi_n = P_{g_0}^n (h(\hat{g}_n, g_0) \geq M_0 \varepsilon_n) \leq P_{g_0}^n (h(\hat{g}_n, g_0) \geq L_0 n^{-2/5}) \leq C e^{-c n^{1/5}} \to 0$$

as $n \to \infty$ for M_0 large enough since $\varepsilon_n \gtrsim n^{-2/5}$. For $f \in \mathcal{F}$ such that $h(f, f_0) \geq M\varepsilon_n$,

$$\begin{aligned} P_f^n (1 - \phi_n) &= P_f^n (h(f_0, \hat{f}_n) < M_0 \varepsilon_n) \\ &\leq P_f^n (h(f_0, f) - h(f, \hat{f}_n) < M_0 \varepsilon_n) \\ &\leq P_f^n ((M - M_0)\varepsilon_n < h(f, \hat{f}_n)). \end{aligned}$$

Since $\varepsilon_n \lesssim n^{-3/8-\rho}$ implies $\varepsilon_n \lesssim n^{-3/8-\rho'}$ for any $0 < \rho' \leq \rho$, we may take $\rho > 0$ arbitrarily small. Applying Lemma 1 with $\varepsilon(\rho) > 0$ to be chosen below and $L_n = (M -$

$M_0)\varepsilon_n n^{2/5}$, which satisfies $L_n \geq L_0$ for $M > 0$ large enough since $\varepsilon_n \gtrsim n^{-2/5}$, yields

$$\begin{aligned}\sup_{f\in\mathcal{F}:\, h(f,f_0)\geq M\varepsilon_n} P_f^n(1-\phi_n) &\leq \sup_{g\in\overline{\mathcal{F}}^{0,1}} P_g^n(h(\hat{g}_n,g) > (M-M_0)\varepsilon_n)\\ &= \sup_{g\in\overline{\mathcal{F}}^{0,1}} P_g^n(h(\hat{g}_n,g) > L_n n^{-2/5})\\ &\leq C(\rho)\exp\left(-c(\rho)n^{\frac{1}{4+2\varepsilon(\rho)}}\right) + D\exp\left(-dL_n^2 n^{1/5}\right)\end{aligned}$$

for all $n \geq n_0(\rho)$. Since $\varepsilon_n \lesssim n^{-3/8-\rho}$ by assumption, it follows that $n\varepsilon_n^2 \lesssim n^{1/4-2\rho} = o(n^{1/(4+2\varepsilon(\rho))})$ for $\varepsilon(\rho) > 0$ small enough. Therefore,

$$\sup_{f\in\mathcal{F}:\, h(f,f_0)\geq M\varepsilon_n} P_f^n(1-\phi_n) \leq (1+o(1))\, De^{-d(M-M_0)^2 n\varepsilon_n^2}.$$

Since we can make the constant in the exponent arbitrarily large by taking $M > 0$ large enough, this completes the proof. □

Proof of Theorem 3. Let $f_0 \in \mathcal{F}_{\alpha,\beta}$ for some $\alpha > 0$ and $\beta \in \mathbb{R}$. We may restrict to a suitable compactly supported density approximating f_0 using the first paragraph of the proof of Theorem 2 of Ghosal and van der Vaart [12]. For completeness we reproduce their argument in this paragraph. Let $\psi_n(x) = \mathbb{1}_{[-t_n,t_n]}(x)$ for $t_n = a' \log n$ for some $a' > \alpha^{-1}$. Define new observations $\bar{X}_1, ..., \bar{X}_{\bar{n}}$ from the original observations $X_1, ..., X_n$ by rejecting each X_i independently with probability $1 - \psi_n(X_i)$. Since $P_{f_0}[-t_n, t_n]^c \leq 2e^{\beta}\alpha^{-1}e^{-\alpha t_n} = o(n^{-1})$, the probability that at least one of the X_i's is rejected is $o(1)$ and so the posterior based on the original and modified observations are the same with $P_{f_0}^n$-probability tending to one. Since posterior contraction is defined via convergence in $P_{f_0}^n$-probability, this implies that the posterior contraction rates are the same. The new observations come from a density $f_{0,n}$ that is proportional to $f_0\psi_n$, which is log-concave and upper semi-continuous. Since $|1 - \int f_0\psi_n| \leq P_{f_0}[-t_n, t_n]^c = o(n^{-1})$,

$$\begin{aligned}h^2(f_0, f_{0,n}) &\leq 2\int_{\mathbb{R}} f_0\left(1 - \frac{1}{\sqrt{\int f_0\psi_n}}\right)^2 dx + \frac{2}{\int f_0\psi_n}\int_{\mathbb{R}} f_0(1-\sqrt{\psi_n})^2 dx\\ &\leq 2\frac{(\int f_0\psi_n - 1)^2}{\int f_0\psi_n} + \frac{2}{\int f_0\psi_n}\int_{\mathbb{R}\setminus[-t_n,t_n]} f_0 dx = o(n^{-1}).\end{aligned} \tag{8}$$

It therefore suffices to establish contraction for the posterior based on the new observations about the density $f_{0,n} = f_0\psi_n / \int f_0\psi_n$.

Under the assumed conditions on $(b_n - a_n)$, ε_n given by (4) satisfies $n^{-2/5} \lesssim \varepsilon_n \lesssim n^{-3/8-\rho}$ for some $\rho > 0$ small enough. We thus apply Theorem 1 so that we need only show the small-ball probability (1). Note that $f_{0,n}(x) \leq e^{\beta-\alpha|x|}(1 + o(n^{-1}))$, so that $f_{0,n} \in \mathcal{F}_{\alpha,2\beta}$ for n large enough. Since $-a_n, b_n \gg \log n$ and $b_n - a_n = o(n^{4/5})$, we may construct an approximation $\bar{f}_n$ of $f_{0,n}$ based on the interval $[a_n, b_n]$ for n large enough

using Proposition 1. By Lemma 8 of [12],

$$\int_{\mathbb{R}} f_{0,n}\Big(\log\frac{f_{0,n}}{f_W}\Big)^k \lesssim \big(h^2(f_{0,n},\bar f_n)+h^2(\bar f_n,f_W)\big) \times\left(1+\log\left\|\frac{f_{0,n}}{\bar f_n}\right\|_{L^\infty([a_n,b_n])}+\log\left\|\frac{\bar f_n}{f_W}\right\|_{L^\infty([a_n,b_n])}\right)^k$$

for $k=1,2$. By Proposition 1(i) and (iv), the first term in the first bracket and the second term in the second bracket are $O((\log n)^2 n^{-4/5}+(b_n-a_n)^2 n^{-8/5})$ and $O(1)$ respectively.

By Proposition 1(v), $\bar f_n$ has representation

$$\bar f_n(x)=\exp\left(\bar\gamma_1\sum_{i=1}^{\bar N}\frac{z_i\wedge(x-a_n)}{z_i}\bar p_i-\bar\gamma_2(x-a_n)+\bar\gamma_3\right)\mathbb{1}_{[a_n,b_n]}(x), \tag{9}$$

where $(z_i)_{i=1}^{\bar N}\subset[0,b_n-a_n]$ are the knots written in increasing order, $\bar N=\bar N_n=O(n^{1/5}\log n)$ and $\sum_{i=1}^{\bar N}\bar p_i=1$. Let $\bar w_n(x)=(\log\bar f_n(x)-\bar\gamma_3)\mathbb{1}_{[a_n,b_n]}(x)-\infty\mathbb{1}_{\mathbb{R}\setminus[a_n,b_n]}(x)$ so that $\bar f_n=f_{\bar w_n}$ using the transformation (2). We may thus without loss of generality take $\bar\gamma_3=0$ since it is contained in the normalization (2).

Suppose that f_w is a (log-concave) density with support equal to $[a_n,b_n]$ and such that $\|\bar w_n-w\|_{L^\infty([a_n,b_n])}\le c\varepsilon_n$. Since $\int e^w=e^{O(\varepsilon_n)}\int e^{\bar w_n}$, it follows that for $x\in[a_n,b_n]$, $\bar f_n(x)/f_w(x)\le e^{O(\varepsilon_n)}=e^{o(1)}$. Since $h^2(\bar f_n,f_w)\lesssim\varepsilon_n^2$ by Lemma 3.1 of [35], we can conclude that

$$\{w:\|\bar w_n-w\|_{L^\infty([a_n,b_n])}\le c\varepsilon_n\}\subset\{w:K(f_{0,n},f_w)\le\varepsilon_n^2,\ V(f_{0,n},f_w)\le\varepsilon_n^2\} \tag{10}$$

for some $c>0$. It therefore suffices to lower bound the prior probability of the left-hand set.

Fix $\delta>0$ to be chosen sufficiently large below. Since the z_i are $n^{-6/5}$-separated by Proposition 1(iii), we can find a collection of disjoint intervals $(U_i)_{i=1}^{\bar N}$ in $[a_n,b_n]$ with Lebesgue measure $\lambda(U_i)=\varepsilon_n^\delta$ and such that $z_i\in U_i$ for $i=1,\dots,\bar N$. Further denote $U_0:=\mathbb{R}\setminus\cup_{i=1}^{M_n}U_i$. Let W be a prior draw of the form (3) with parameters γ_1,γ_2 and P.

Writing $p_i = P(U_i)$, $\bar{p}_0 = 0$ and using the triangle inequality, for any $x \in [a_n, b_n]$,

$$\begin{aligned}
|\bar{w}_n(x) - W(x)| &= \left| \bar{\gamma}_1 \sum_{i=1}^{\bar{N}} \frac{z_i \wedge (x-a_n)}{z_i} \bar{p}_i - \bar{\gamma}_2 (x-a_n) - \gamma_1 \int_0^\infty \frac{\theta \wedge (x-a_n)}{\theta} dP(\theta) - \gamma_2 (x-a_n) \right| \\
&\leq |\bar{\gamma}_1 - \gamma_1| \int_0^\infty \frac{\theta \wedge (x-a_n)}{\theta} dP(\theta) + \bar{\gamma}_1 \left| \int_{U_0} \frac{\theta \wedge (x-a_n)}{\theta} dP(\theta) \right| \\
&\quad + \bar{\gamma}_1 \left| \sum_{i=1}^{\bar{N}} \int_{U_i} \frac{\theta \wedge (x-a_n)}{\theta} dP(\theta) - \sum_{i=1}^{\bar{N}} \frac{z_i \wedge (x-a_n)}{z_i} p_i \right| \\
&\quad + \bar{\gamma}_1 \left| \sum_{i=1}^{\bar{N}} \frac{z_i \wedge (x-a_n)}{z_i} (p_i - \bar{p}_i) \right| + (b_n - a_n) |\bar{\gamma}_2 - \gamma_2| \\
&\leq |\bar{\gamma}_1 - \gamma_1| + \bar{\gamma}_1 \sum_{i=1}^{\bar{N}} \sup_{\theta \in U_i} \frac{|\theta - z_i|}{\theta \wedge z_i} p_i + \bar{\gamma}_1 p_0 + \bar{\gamma}_1 \sum_{i=1}^{\bar{N}} |p_i - \bar{p}_i| + (b_n - a_n) |\bar{\gamma}_2 - \gamma_2| \\
&\leq |\bar{\gamma}_1 - \gamma_1| + 2\bar{\gamma}_1 \sum_{i=1}^{\bar{N}} \frac{\lambda(U_i)}{z_i} p_i + \bar{\gamma}_1 \sum_{i=0}^{\bar{N}} |p_i - \bar{p}_i| + (b_n - a_n) |\bar{\gamma}_2 - \gamma_2|,
\end{aligned} \tag{11}$$

where we have used in the second to last line that the maximal distance between the (piecewise) lines $(y \wedge a)/a$ and $(y \wedge b)/b$ occurs at $y = a \wedge b$ and in the last line that $\theta > z_i/2$ for all $\theta \in U_i$ for a sufficiently large choice of the parameter $\delta > 0$. By Proposition 1(v), we have $\bar{\gamma}_1 \leq 2n^{4/5}(b_n - a_n) \lesssim n^{(4+\mu)/5}$. Furthermore, by the separation of the knots, $z_i \geq z_1 \geq cn^{-6/5}$, $i = 1, ..., \bar{N}$, and so by the assumptions on the (U_i), the second term is bounded by $2c^{-1}\bar{\gamma}_1 n^{6/5} \varepsilon_n^\delta \leq \tilde{c}\varepsilon_n$ for some $\delta, \tilde{c} > 0$ large enough.

The remaining three terms are independent under the prior and so can be dealt with separately. By the assumptions on the base measure of the Dirichlet process, we have that $\sum_{i=0}^{\bar{N}} H(U_i) \leq H(\mathbb{R}^+)$ and $H(U_i) = H(\mathbb{R}^+)\bar{H}(U_i) \gtrsim \lambda(U_i)/(b_n - a_n)^\eta \geq \varepsilon_n^\delta/(b_n - a_n)^\eta \geq \varepsilon_n^{\delta'}$ for $i = 1, ..., \bar{N}$ and some $\delta' > \delta$. For $i = 0$, note that $\lambda(U_0) \geq (b_n - a_n) - \bar{N}\varepsilon_n^\delta \gtrsim 1$. Using the lower the bounds for the $\lambda(U_i)$, which come from the polynomial separation of the knots in Proposition 1(iii), we can apply Lemma 10 of [12] to get

$$\Pi_n \left(\bar{\gamma}_1 \sum_{i=0}^{\bar{N}} |p_i - \bar{p}_i| \leq \varepsilon_n \right) \gtrsim e^{-c\bar{N} \log(2\bar{\gamma}_1/\varepsilon_n)} \gtrsim e^{-c' n \varepsilon_n^2}. \tag{12}$$

From the tail assumption on the density of γ_1 and the upper bound on $\bar{\gamma}_1$, we have

$$\Pi_n\big(|\gamma_1 - \bar{\gamma}_1| \leq \varepsilon_n\big) \gtrsim \varepsilon_n e^{-c(\bar{\gamma}_1 + \varepsilon_n)^{1/(4+\mu)}} \geq e^{-c' n^{1/5}} \geq e^{-n\varepsilon_n^2}. \tag{13}$$

By Proposition 1(v), $|\bar{\gamma}_2| \lesssim n^{4/5}$, which, combined with the tail bound on the density of γ_2, yields

$$\Pi_n((b_n - a_n)|\gamma_2 - \bar{\gamma}_2| \leq \varepsilon_n) \gtrsim \tfrac{\varepsilon_n}{b_n - a_n} e^{-c(|\bar{\gamma}_2| + \varepsilon_n/(b_n - a_n))^{1/4}} \geq e^{-c' n^{1/5}} \geq e^{-n\varepsilon_n^2} \tag{14}$$

since $\varepsilon_n/(b_n - a_n) \to 0$ no faster than polynomially in n. Combining the above, we have that $\Pi_n(\|\bar{w}_n - W\|_{L^\infty([a_n,b_n])} \leq (3+\tilde{c})\varepsilon_n) \geq e^{-(2+c')n\varepsilon_n^2}$. □

Proof of Theorem 4. Since the proof follows that of Theorem 3, we only specify the details where the present proof differs. Using the same arguments, we restrict to studying posterior contraction based on observations arising from the log-concave density $f_{0,n} = f_0\psi_n / \int f_0\psi_n$ for $\psi_n(x) = \mathbb{1}_{[-t_n,t_n]}(x)$ with $t_n = a' \log n$ for some $a' > \alpha^{-1}$. We again apply Theorem 1 so that we only need to show the small-ball probability (1) for $\varepsilon_n = (\log n)n^{-2/5}$.

Writing $\Pi_{a,b}$ for the prior conditional on (a,b) and setting $\Delta_n = \{(a,b) : -3t_n \leq a \leq -2t_n, 2t_n \leq b \leq 3t_n\}$, the small ball probability in (1) is lower bounded by

$$\begin{aligned} &\int_{\Delta_n} \Pi_{a,b}\left(K(f_{0,n}, f_W) \leq \varepsilon_n^2,\ V(f_{0,n}, f_W) \leq \varepsilon_n^2\right) \pi(a,b) da\, db \\ &\quad \geq \inf_{(a,b)\in\Delta_n} \Pi_{a,b}\left(K(f_{0,n}, f_W) \leq \varepsilon_n^2,\ V(f_{0,n}, f_W) \leq \varepsilon_n^2\right) \times \int_{\Delta_n} \pi(a,b) da\, db. \end{aligned} \tag{15}$$

Using the lower bound assumption (5) on $\pi(a,b)$, the last integral is lower bounded by $Ce^{-c_1(3t_n)^q - c_2(6t_n)^r} \int_{\Delta_n} da\, db \geq Ct_n^2 e^{-c_3(\log n)^{q\vee r}} \gtrsim e^{-C'n\varepsilon_n^2}$. It thus suffices to lower bound the infimum in the last display.

Since $[-\frac{8}{5\alpha}\log n, \frac{8}{5\alpha}\log n] \subset [a,b]$ and $(b-a) = O(\log n) = o(n^{4/5})$ for all $(a,b) \in \Delta_n$, we may apply Proposition 1 to construct an approximation $\bar{f}_n$ of $f_{0,n}$ based on the interval $[a,b]$ for any $(a,b) \in \Delta_n$. One can then proceed exactly as in the proof of Theorem 3 to lower bound the prior small-ball probability by $e^{-Cn\varepsilon_n^2}$ for fixed $(a,b) \in \Delta_n$. Since all constants in that argument depend only on α, β and the prior hyperparameters, the lower bound is uniform over all (a,b) with $(b-a) = O(n^{2/5})$ (to ensure ε_n in (4) takes the value $(\log n)n^{-2/5}$) and $-a, b \geq \frac{8}{5\alpha}\log n$. In particular, the lower bound is uniform over Δ_n. □

Proof of Corollary 1. We use the notation employed in the proof of Theorem 3. By (10), it suffices to lower bound the prior probability of an L^∞-small ball about $\bar{w}_n$, where $\bar{f}_n = f_{\bar{w}_n}$ is the approximation (9). Since $\bar{N} \leq N$ (at least for n large enough), we can add additional breakpoints to the piecewise linear function $\bar{w}_n$ with weights $\bar{p}_i = 0$, $i = \bar{N}+1, ..., N$, without changing $\bar{w}_n$. Without loss of generality, pick any such additional breakpoints to be no smaller than $cn^{-6/5}$. Using similar computations to (11), for any

$x \in [a_n, b_n]$,

$$\begin{aligned}
|\bar{w}_n(x) - w(x)| &= \left| \bar{\gamma}_1 \sum_{i=1}^N \frac{z_i \wedge (x-a_n)}{z_i} \bar{p}_i - \bar{\gamma}_2 (x-a_n) - \gamma_1 \sum_{i=1}^N \frac{\theta_i \wedge (x-a_n)}{\theta_i} p_i - \gamma_2 (x-a_n) \right| \\
&\leq |\bar{\gamma}_1 - \gamma_1| \sum_{i=1}^N \frac{\theta_i \wedge (x-a_n)}{\theta_i} p_i + \bar{\gamma}_1 \left| \sum_{i=1}^N \frac{\theta_i \wedge (x-a_n)}{\theta_i} p_i - \sum_{i=1}^N \frac{z_i \wedge (x-a_n)}{z_i} p_i \right| \\
&\quad + \bar{\gamma}_1 \sum_{i=1}^N \frac{z_i \wedge (x-a_n)}{z_i} |p_i - \bar{p}_i| + (b_n - a_n) |\bar{\gamma}_2 - \gamma_2| \\
&\leq |\bar{\gamma}_1 - \gamma_1| + \bar{\gamma}_1 \sum_{i=1}^N \frac{|\theta_i - z_i|}{\theta_i \wedge z_i} p_i + \bar{\gamma}_1 \sum_{i=1}^N |p_i - \bar{p}_i| + (b_n - a_n) |\bar{\gamma}_2 - \gamma_2|.
\end{aligned}$$

The first and fourth terms are bounded from above by ε_n with prior probability at least $e^{-n\varepsilon_n^2}$ by (13) and (14), respectively, for both priors. For the second term note, similarly to the proof of Theorem 3, that $z_1 \geq cn^{-6/5}$. Taking $\theta_i \in [z_i, z_i + cn^{-6/5}\varepsilon_n/\bar{\gamma}_1]$, the second term is bounded by $\bar{\gamma}_1 \sum_{i=1}^N p_i \varepsilon_n/\bar{\gamma}_1 = \varepsilon_n$. The probability of this set under the base measure is $H([z_i, z_i + cn^{-6/5}\varepsilon_n/\bar{\gamma}_1]) \gtrsim cn^{-6/5}\varepsilon_n/\big(\bar{\gamma}_1 (b_n - a_n)^\eta\big)$ by the assumptions on $\bar{H}$. The joint probability that $\theta_i \in [z_i, z_i + cn^{-6/5}\varepsilon_n/\bar{\gamma}_1]$ for every $i = 1, ..., N$ is therefore bounded from below by a multiple of $(cn^{-6/5}\varepsilon_n/\bar{\gamma}_1)^N/(b_n - a_n)^{\eta N} \gtrsim e^{-c_1 N \log n} \geq e^{-c_2 n\varepsilon_n^2}$, for sufficiently large constants $c_1, c_2 > 0$.

It remains to show that the third term is bounded from above by ε_n with probability at least $e^{-cn\varepsilon_n^2}$ for some $c > 0$. In the case where $(p_1, ..., p_N)$ is endowed with a Dirichlet distribution, this statement follows from (12). In the case of the truncated stick-breaking prior, writing $(\bar{p}_1, \ldots, \bar{p}_N)$ in decreasing order, we note that there exist $0 \leq \bar{v}_1, ..., \bar{v}_{N-1} \leq 1$ such that $\bar{p}_i = \prod_{j=1}^{i-1}(1 - \bar{v}_j)\bar{v}_i$, $i = 1, ..., N-1$, and $\bar{p}_N = \prod_{j=1}^{N-1}(1 - \bar{v}_j)$. Define for $i = 1, 2, ..., N-1$ the intervals

$$I_i = \big[(\bar{v}_i - \varepsilon_n n^{-4/5} N^{-2}) \vee \varepsilon_n n^{-4/5} N^{-2}/2, (\bar{v}_i + \varepsilon_n n^{-4/5} N^{-2}) \wedge (1 - \varepsilon_n n^{-4/5} N^{-2}/2)\big].$$

For $v_i \in I_i \subset [0,1]$, $i = 1, .., N-1$, we have

$$\begin{aligned}
\Big|(1-v_1)...(1-v_i)v_{i+1} - (1-\bar{v}_1)...(1-\bar{v}_i)\bar{v}_{i+1}\Big| &\leq |v_1 - \bar{v}_1| + |v_2 - \bar{v}_2| + ... + |v_{i+1} - \bar{v}_{i+1}| \\
&\leq (i+1)\varepsilon_n n^{-4/5} N^{-2} \leq \varepsilon_n n^{-4/5} N^{-1}.
\end{aligned}$$

Hence for $p_i := (1-v_1)(1-v_2)...(1-v_{i-1})v_i$, $i = 1, ..., N-1$, $(p_1, .., p_N)$ is in the N-dimensional simplex and $\bar{\gamma}_1 \sum_{i=1}^N |\bar{p}_i - p_i| \leq \bar{\gamma}_1 N \varepsilon_n n^{-4/5} N^{-1} \lesssim \varepsilon_n$. Finally, we note that for $v_i \sim \text{Beta}(a, b)$, we have $P(v_i \in I_i) \gtrsim (\varepsilon_n n^{-4/5} N^{-2})^{a \vee b}$ and we can therefore conclude

$$P\left(\bar{\gamma}_1 \sum_{i=1}^N |p_i - \bar{p}_i| \leq c\varepsilon_n\right) \geq \prod_{i=1}^{N-1} P(v_i \in I_i) \geq e^{N(a \vee b)\log(\varepsilon_n N^{-2} n^{-4/5})} \geq e^{-c_1 N \log n} \geq e^{-c_2 n \varepsilon_n^2},$$

for some large enough constants $c_1, c_2 > 0$, thereby completing the proof.

For the hierarchical case where we assign a prior to (a, b), the proof follows as in that of Theorem 4 using (15) and the lower bound for the small-ball probability just derived. □

6. Proof of Proposition 1

In this section, we construct the piecewise log-linear approximation for an upper semi-continuous log-concave density given in Proposition 1. In particular, we require that the number of knots in the approximating function does not grow too quickly and that the knots are polynomially separated, thereby rendering the construction somewhat involved. The proof relies on firstly approximating any continuous concave function on a given compact interval using a piecewise linear function. One then splits $\mathrm{supp}(f_0)$ into sets, depending on the size of both $\log f_0$ and $|(\log f_0)'|$, and obtains suitable piecewise linear approximations defined locally on each of these sets. Piecing together these local functions gives the desired global approximation.

We now construct a piecewise linear approximation of a continuous concave function w on a compact interval $[a, b]$. For any partition $a = x_0 < x_1 < \cdots < x_m = b$ of $[a, b]$, let $\widetilde{w}_m$ denote the piecewise linear approximation of w given by

$$\widetilde{w}_m(x) := \sum_{i=2}^{m} \left(\frac{x - x_{i-1}^*}{x_i^* - x_{i-1}^*} \frac{1}{x_i - x_{i-1}} \theta_i + \frac{x_i^* - x}{x_i^* - x_{i-1}^*} \frac{1}{x_{i-1} - x_{i-2}} \theta_{i-1} \right) \mathbb{1}_{(x_{i-1}^*, x_i^*]}(x), \quad (16)$$

where $\theta_i := \int_{x_{i-1}}^{x_i} w(s)ds$ and $x_i^* := \frac{x_i + x_{i-1}}{2}$. On $[a, x_1^*]$ and $(x_m^*, b]$, the function is defined by linearly extending the piecewise linear function defined above, that is

$$\begin{aligned} \widetilde{w}_m(a) &:= \frac{1}{x_2^* - x_1^*} \Big(\frac{x_2^* - a}{x_1 - a} \theta_1 - \frac{x_1^* - a}{x_2 - x_1} \theta_2 \Big), \\ \widetilde{w}_m(b) &:= \frac{1}{x_m^* - x_{m-1}^*} \Big(\frac{b - x_{m-1}^*}{b - x_{m-1}} \theta_m - \frac{b - x_m^*}{x_{m-1} - x_{m-2}} \theta_{m-1} \Big). \end{aligned} \quad (17)$$

The function $\widetilde{w}_m$ takes value $\widetilde{w}_m(x_i^*) = \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} w(s)ds$ at the midpoint $x_i^* = \frac{x_i + x_{i-1}}{2}$ of the interval $[x_{i-1}, x_i]$ and interpolates linearly in between. We next state several technical lemmas whose proofs can be found in the supplementary material [24].

Lemma 2. *Let $w : [a, b] \to \mathbb{R}$ be a continuous concave function, where $-\infty < a < b < \infty$. For any partition $a = x_0 < x_1 < \cdots < x_m = b$ of $[a, b]$, let $\widetilde{w}_m$ denote the piecewise linear approximation of w defined in (16) and (17). Then $\widetilde{w}_m$ is a concave function.*

Lemma 3. *Let $w : [a, b] \to \mathbb{R}$ be a continuous concave function with $w'_+(a) - w'_-(b) \leq M$ and where $-\infty < a < b < \infty$. Then there exists a partition $a = x_0 < x_1 < \cdots < x_m = b$ of $[a, b]$ with $\min_{i=1,\ldots,m}(x_i - x_{i-1}) \geq (b - a)(2m)^{-2}$ and such that*

$$\sup_{x \in [a,b]} |w(x) - \widetilde{w}_m(x)| \leq C \frac{M(b - a)}{m^2},$$

where $\widetilde{w}_m$ is the piecewise linear approximation of w defined in (16) and (17) and $C > 0$ is a universal constant (i.e. not depending on a, b, m).

Lemma 4. *Any piecewise linear concave function $w : [a, b] \to \mathbb{R}$ with N knots $\{z_1, ..., z_N\}$ can be written in the form*

$$w(x) = \gamma_1 \sum_{i=1}^{N} \frac{z_i \wedge (x-a)}{z_i} p_i - \gamma_2(x-a) + \gamma_3,$$

with parameters $0 \le \gamma_1 \le (w'_+(a) - w'_-(b))(b-a)$, $|\gamma_2| \le |w'_-(b)|$, $\gamma_3 \in \mathbb{R}$, $\sum_{i=1}^N p_i = 1$ and $p_i \ge 0$ for $i = 1, \dots, N$.

Proof of Proposition 1. Let $\psi_n(x) = \mathbb{1}_{[-s_n, s_n]}(x)$ for $s_n = \frac{4}{5\alpha} \log n$. The log-concave density function $f_1 = f_{1,n} = f_0\psi_n / \int f_0\psi_n$ supported on $[-s_n, s_n]$ satisfies $|1 - \int f_0\psi_n| \le P_{f_0}[-s_n, s_n]^c \le 2e^\beta \alpha^{-1} n^{-4/5}$. Arguing as in (8), one has $h^2(f_0, f_{1,n}) \le 12 e^\beta \alpha^{-1} n^{-4/5}$ for $n \ge (4e^\beta/\alpha)^{5/4}$.

We write $f_1 = e^{w_1}$ and construct the approximating function $\bar{f}_n$ according to the value of w_1 and its left and right derivatives $w'_{1,-}$ and $w'_{1,+}$. Let

$$A_0^n = \{x \in [a_n, b_n] : w_1(x) < -\tfrac{4}{5}\log n\},$$
$$A_1^n = \{x \in [a_n, b_n] : w_1(x) \ge -\tfrac{4}{5}\log n, |w'_{1,\pm}(x)| > n^{4/5}\},$$
$$A_{2,j}^n = \{x \in [a_n, b_n] : w_1(x) \ge -\tfrac{4}{5}\log n, 2^{-j-1} n^{4/5} < |w'_{1,\pm}(x)| \le 2^{-j} n^{4/5}\}, \quad j = 0, \dots, j_n,$$
$$A_3^n = \{x \in [a_n, b_n] : w_1(x) \ge -\tfrac{4}{5}\log n, |w'_{1,\pm}(x)| \le D\},$$

where $D > 0$ is some fixed constant, $|w'_{1,\pm}(x)| = \max(|w'_{1,+}(x)|, |w'_{1,-}(x)|)$ and $j_n = \lceil \log_2(n^{4/5}/D) \rceil - 1$. In fact the set where the left and right derivatives of the concave function w_1 do not agree has measure zero. Note that the above sets are all disjoint except A_{2,j_n}^n and A_3^n: since j_n is the smallest integer such that $2^{-j_n-1} n^{4/5} \le D$, these last two sets may overlap. In particular, we can express $[a_n, b_n]$ as the almost disjoint union of the above sets. Write $B_n = (\cup_{j=0}^{j_n} A_{2,j}^n) \cup A_3^n \subset [a_n, b_n]$ and note that by the concavity of w_1, this is an interval. Since $\|f_1\|_\infty \le 2e^\beta$ for $n \ge (4e^\beta/\alpha)^{5/4}$, the set A_1^n consists of at most two intervals, each of width $O(n^{-4/5}\log n)$. Using again the boundedness of f_1, the definition of A_0^n and that $|\mathrm{supp}(f_1)| \lesssim \log n$,

$$\int_{B_n} f_1 dx = 1 - O(n^{-4/5}\log n), \tag{18}$$

so that in particular, $B_n \neq \emptyset$ for $n \ge n_0(\alpha, \beta)$ large enough.

We now construct a partition $\mathcal{P}_n$ of B_n based on which we take the piecewise linear approximation (16)-(17) of the function w_1. Note that $A_{2,j}^n$ consists of at most two disjoint intervals, $A_{2,j,+}^n$ and $A_{2,j,-}^n$, which by the boundedness of f_1 are each of length $O(2^{j+1} n^{-4/5} \log n)$. Let $\mathcal{P}_{n, A_{2,j,+}^n}$ denote the partition of the interval $A_{2,j,+}^n$ given by Lemma 3 with partition size $m = \lceil 2^{-j/2} n^{3/5} |A_{2,j,+}^n|^{1/2} / \sqrt{\log n} \rceil = O(n^{1/5})$ and let

$\mathcal{P}_{n,A^n_{2,j,-}}$ be the analogous partition constructed on $A^n_{2,j,-}$. Similarly, $A^n_3 \cap (A^n_{2,j_n})^c$ consists of a single interval of length $O(\log n)$. Let $\mathcal{P}_{n,A^n_3\cap(A^n_{2,j_n})^c}$ denote the corresponding partition of $A^n_3 \cap (A^n_{2,j_n})^n$ given by Lemma 3 with partition size $m = \lceil n^{1/5}(D|A^n_3 \cap (A^n_{2,j_n})^c|/\log n)^{1/2}\rceil = O(n^{1/5})$. Define the overall partition

$$\mathcal{P}_n = \mathcal{P}_{n,A^n_3\cap(A^n_{2,j_n})^c} \cup \bigcup_{j=0}^{j_n} (\mathcal{P}_{n,A^n_{2,j,+}} \cup \mathcal{P}_{n,A^n_{2,j,-}})$$

of B_n, which has $O(j_n n^{1/5}) = O(n^{1/5}\log n)$ points. The associated piecewise linear function $\widetilde{w}_n$ defined in (16)-(17) based on $\mathcal{P}_n$ is concave by Lemma 2 and by construction corresponds to the partition given in Lemma 3 for each of the sets comprising B_n. It therefore satisfies the conclusions of Lemma 3 on each such set (with the appropriate m), so that in particular,

- $\|w_1 - \widetilde{w}_n\|_{L^\infty(A^n_{2,j})} \le Cn^{-2/5}\log n$ for some universal constant $C > 0$ independent of j,
- the partition points in $A^n_{2,j}$ are distance at least $c2^j n^{-6/5}\log n \ge cn^{-6/5}\log n$ apart for some universal constant $c > 0$ independent of j,
- on $A^n_3 \cap (A^n_{2,j_n})^c$, we have the same L^∞-bound with the partition points being $cn^{-2/5}\log n$-separated.

Moreover, since these intervals meet only at their boundaries, and the boundary points of the intervals are contained in the partition presented in Lemma 3, the interval boundaries will be contained in $\mathcal{P}_n$. Consequently, the separation property continues to hold even across the different subpartitions. In conclusion, we have shown that $\widetilde{w}_n$ is concave and piecewise linear with $O(n^{1/5}\log n)$ knots, which are $cn^{-6/5}\log n$-separated, and satisfies

$$\sup_{x\in B_n} |\widetilde{w}_n(x) - w_1(x)| = O(n^{-2/5}\log n). \tag{19}$$

We now extend the approximating function to $[a_n, b_n] \supset B_n$. Write $\mathcal{P}_n = (x_i)_{i=0}^M$, where $\min(B_n) = x_0 < x_1 < \ldots < x_M = \max(B_n)$ and $M = O(n^{1/5}\log n)$. Define $\bar{w}_n : [a_n, b_n] \to \mathbb{R}$ as

$$\bar{w}_n(x) = \begin{cases} \widetilde{w}_n(x_0) + \big(w'_{1,-}(x_0) \wedge n^{4/5} \vee (-n^{4/5})\big)(x - x_0) & x \in [a_n, x_0], \\ \widetilde{w}_n(x) & x \in B_n, \\ \widetilde{w}_n(x_M) + \big(w'_{1,+}(x_M) \wedge n^{4/5} \vee (-n^{4/5})\big)(x - x_M) & x \in [x_M, b_n]. \end{cases} \tag{20}$$

This is simply the function $\widetilde{w}_n$ extended linearly from the boundary points of B_n with slope $w'_{1,-}(x_0) \wedge n^{4/5} \vee (-n^{4/5})$ and $w'_{1,+}(x_M) \wedge n^{4/5} \vee (-n^{4/5})$ on $[a_n, x_0]$ and $[x_M, b_n]$ respectively. We now verify that $\bar{w}_n$ is concave, for which it is enough to show that $\bar{w}'_{n,+}(x_0) \le \bar{w}'_{n,-}(x_0)$ and $\bar{w}'_{n,+}(x_M) \le \bar{w}'_{n,-}(x_M)$. For the first inequality, using equation (8) in the supplement [24], the concavity of w_1 and the boundary construction of $\widetilde{w}_n$ given by (16), $\bar{w}'_{n,+}(x_0) = \widetilde{w}'_{n,+}(x_0) = \widetilde{w}'_{n,+}(x_1^*) \le w'_{1,+}(x_0) \le w'_{1,-}(x_0)$. Since $x_0 \in B_n$, it also holds that $|\bar{w}'_{n,+}(x_0)| \le n^{4/5}$. The second inequality can be proved analogously.

Since $\log(1+z) = O(z)$ as $|z| \to 0$, it follows that $\log \int f_0 \psi_n = O(n^{-4/5})$. Using this and (19),

$$|\bar{w}_n(x_0) - \log f_0(x_0)| \le \left| \log \int f_0 \psi_n \right| + O(n^{-2/5} \log n) = O(n^{-2/5} \log n).$$

By concavity, the slope of the linear extension on $[a_n, x_0]$ satisfies $w'_{1,-}(x_0) = (\log f_0)'_-(x_0) \le (\log f_0)'_+(x)$ for all $x < x_0$ such that $f_0(x) > 0$. Combining the above yields $\bar{w}_n(x) \ge \log f_0(x) - O(n^{-2/5} \log n)$ for all $x \in [a_n, x_0]$. The same computation also gives the result for $x \in [x_M, b_n]$, so that for some $C > 0$,

$$\sup_{x \in [a_n, b_n] \setminus B_n} (\log f_0(x) - \bar{w}_n(x)) \le C n^{-2/5} \log n. \tag{21}$$

Define the log-concave density

$$\bar{f}_n(x) = \begin{cases} e^{\bar{w}_n(x)} / \int_{a_n}^{b_n} e^{\bar{w}_n} & x \in [a_n, b_n], \\ 0 & x \notin [a_n, b_n]. \end{cases}$$

This function is piecewise log-linear, has $O(n^{1/5} \log n)$ knots and satisfies (ii) and (iii) by construction. We have

$$h^2(f_1, \bar{f}_n) \le 2 \int_{B_n^c} f_1 + 2 \int_{B_n^c} \bar{f}_n + \int_{B_n} (f_1^{1/2} - \bar{f}_n^{1/2})^2. \tag{22}$$

The first integral is $O(n^{-4/5} \log n)$ by (18). Using (19), $\int_{B_n} e^{\bar{w}_n} = e^{o(1)} \int_{B_n} f_1$. Write the second integral as $\int_{B_n^c} \bar{f}_n = \int_{A_0^n} \bar{f}_n + \int_{A_1^n} \bar{f}_n$. By the definition of A_0^n and (21), $\int_{A_0^n} e^{\bar{w}_n} \le \int_{A_0^n} e^{-(4/5)\log n + n^{-2/5} \log n} \le (x_0 - a_n + b_n - x_M) e^{o(1)} n^{-4/5} = O((b_n - a_n) n^{-4/5})$. For the integral over A_1^n we simply observe that by (21), $\bar{w}_n \le \beta + n^{-2/5} \log n$, and recall that the measure of A_1^n is at most $2n^{-4/5}(\beta + \frac{4}{5} \log n)$. Since $b_n - a_n \ge \frac{16}{5\alpha} \log n$, then $\int_{A_1^n} \bar{f}_n$ is also $O((b_n - a_n) n^{-4/5})$. This implies that the second integral in (22) is $O((b_n - a_n) n^{-4/5})$.

Using (18), (19), Lemma 3.1 of [35] and the above, the third term of (22) is bounded by a multiple of

$$\begin{aligned} &\int_{B_n} e^{w_1} \left(1 - \frac{1}{\sqrt{\int_{B_n} e^{w_1}}} \right)^2 + \int_{B_n} \left(\frac{e^{w_1/2}}{\sqrt{\int_{B_n} e^{w_1}}} - \frac{e^{\bar{w}_n/2}}{\sqrt{\int_{B_n} e^{\bar{w}_n}}} \right)^2 \\ &\qquad + \int_{B_n} e^{\bar{w}_n} \left(\frac{1}{\sqrt{\int_{B_n} e^{\bar{w}_n}}} - \frac{1}{\sqrt{\int_{a_n}^{b_n} e^{\bar{w}_n}}} \right)^2 \\ &\lesssim \left(\int_{B_n} f_1 - 1 \right)^2 + \|\bar{w}_n - w_1\|^2_{L^\infty(B_n)} e^{\|\bar{w}_n - w_1\|_{L^\infty(B_n)}} + \left(\int_{B_n} e^{\bar{w}_n} - \int_{a_n}^{b_n} e^{\bar{w}_n} \right)^2 \\ &= O((\log n)^2 n^{-4/5} + (b_n - a_n)^2 n^{-8/5}), \end{aligned}$$

which establishes (i).

Consider (iv). Note that this is trivial if $f_0(x) = 0$, so assume $f_0(x) \neq 0$. If $x \in B_n$, then by (19),

$$f_0(x)/\bar{f}_n(x) = e^{w_1(x)-\bar{w}_n(x)} \int f_0\psi_n \int_{a_n}^{b_n} e^{\bar{w}_n} = e^{O(n^{-2/5}\log n)}(1+o(1)) = 1+o(1).$$

If $x \in [a_n, b_n]\backslash B_n$, then the result follows from (21).

Consider lastly (v). Since $\bar{w}_n$ defined in (20) is piecewise linear with $|w'_+(a_n)| \vee |w'_-(b_n)| \leq n^{4/5}$, in view of Lemma 4 it takes the form

$$\bar{w}_n(x) = \gamma_1 \sum_{i=1}^{M} \frac{z_i \wedge (x-a_n)}{z_i} - \gamma_2(x-a_n) + \gamma_3, \quad x \in [a_n, b_n],$$

with $M = O(n^{1/5}\log n)$, $\gamma_1 \leq |w'_+(a_n) - w'_-(b_n)|(b_n - a_n) \leq 2n^{4/5}(b_n - a_n)$ and $|\gamma_2| \leq |w'_-(b_n)| \leq n^{4/5}$. This completes the proof. □

Acknowledgements: The authors would like to thank Richard Samworth and Arlene Kim for helpful discussions. The authors would further like to thank the AE and two referees for their helpful suggestions which lead to an improved version of the manuscript.

All three authors received funding from the European Research Council under ERC Grant Agreement 320637 for this research. Ester Mariucci was further supported by the Federal Ministry for Education and Research through the Sponsorship provided by the Alexander von Humboldt Foundation, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 314838170, GRK 2297 MathCoRe, and by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 'Data Assimilation'. Botond Szabó also received funding from the Netherlands Organization for Scientific Research (NWO) under Project number: 639.031.654.

References

[1] Balabdaoui, F., and Doss, C. R. Inference for a two-component mixture of symmetric distributions under log-concavity. *Bernoulli 24*, 2 (2018), 1053–1071.

[2] Balabdaoui, F., Rufibach, K., and Wellner, J. A. Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist. 37*, 3 (2009), 1299–1331.

[3] Birgé, L. Estimation of unimodal densities without smoothness assumptions. *Ann. Statist. 25*, 3 (1997), 970–981.

[4] Cule, M., and Samworth, R. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat. 4* (2010), 254–270.

[5] Cule, M., Samworth, R., and Stewart, M. Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol. 72*, 5 (2010), 545–607.

[6] Doss, C. R., and Wellner, J. A. Inference for the mode of a log-concave density. *Ann. Statist.*, to appear (arXiv:1611.10348).

[7] Doss, C. R., and Wellner, J. A. Univariate log-concave density estimation with symmetry or modal constraints. *Electron. J. Stat.*, to appear (arXiv:1611.10335).

[8] Doss, C. R., and Wellner, J. A. Global rates of convergence of the MLEs of log-concave and s-concave densities. *Ann. Statist. 44*, 3 (2016), 954–981.

[9] Dümbgen, L., and Rufibach, K. Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli 15*, 1 (2009), 40–68.

[10] Dümbgen, L., Samworth, R., and Schuhmacher, D. Approximation by log-concave distributions, with applications to regression. *Ann. Statist. 39*, 2 (2011), 702–730.

[11] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. Convergence rates of posterior distributions. *Ann. Statist. 28*, 2 (2000), 500–531.

[12] Ghosal, S., and van der Vaart, A. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist. 35*, 2 (2007), 697–723.

[13] Ghosal, S., and van der Vaart, A. *Fundamentals of nonparametric Bayesian inference*, vol. 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2017.

[14] Giné, E., and Nickl, R. Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist. 39*, 6 (2011), 2883–2911.

[15] Groeneboom, P., and Jongbloed, G. *Nonparametric estimation under shape constraints*, vol. 38 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, 2014.

[16] Han, Q. Bayes model selection. *ArXiv e-prints* (Apr. 2017).

[17] Hannah, L. A., and Dunson, D. B. Bayesian nonparametric multivariate convex regression. *ArXiv e-prints* (Sept. 2011).

[18] Has′minskiĭ, R. Z. Lower bound for the risks of nonparametric estimates of the mode. In *Contributions to statistics*. Reidel, Dordrecht-Boston, Mass.-London, 1979, pp. 91–97.

[19] Ibragimov, I. On the composition of unimodal distributions. *Theory of Probability & Its Applications 1*, 2 (1956), 255–260.

[20] Khazaei, S., and Rousseau, J. Bayesian Nonparametric Inference of decreasing densities. In *42èmes Journées de Statistique* (Marseille, France, 2010).

[21] Kim, A. K. H., Guntuboyina, A., and Samworth, R. J. Adaptation in log-concave density estimation. *Ann. Statist. 46*, 5 (2018), 2279–2306.

[22] Kim, A. K. H., and Samworth, R. J. Global rates of convergence in log-concave density estimation. *Ann. Statist. 44*, 6 (2016), 2756–2779.

[23] Le Cam, L. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986.

[24] Mariucci, E., Ray, K., and Szabó, B. Supplement to "A Bayesian nonparametric approach to log-concave density estimation". 2018.

[25] Müller, S., and Rufibach, K. Smooth tail-index estimation. *J. Stat. Comput. Simul. 79*, 9-10 (2009), 1155–1167.

[26] Ray, K. Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat. 7* (2013), 2516–2549.

[27] Reiss, M., and Schmidt-Hieber, J. Nonparametric Bayesian analysis of the compound Poisson prior for support boundary recovery. *Ann. Statist.*, to appear (arXiv:1809.04140).

[28] Salomond, J.-B. Concentration rate and consistency of the posterior distribution for selected priors under monotonicity constraints. *Electron. J. Stat. 8*, 1 (2014), 1380–1404.

[29] Samworth, R. J., and Yuan, M. Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist. 40*, 6 (2012), 2973–3002.

[30] Seregin, A., and Wellner, J. A. Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist. 38*, 6 (2010), 3751–3781. With supplementary material available online.

[31] Shively, T. S., Sager, T. W., and Walker, S. G. A Bayesian approach to nonparametric monotone function estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol. 71*, 1 (2009), 159–175.

[32] Shively, T. S., Walker, S. G., and Damien, P. Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *J. Econometrics 161*, 2 (2011), 166–181.

[33] Szabó, B., van der Vaart, A. W., and van Zanten, J. H. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist. 43*, 4 (2015), 1391–1428.

[34] van de Geer, S. A. *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[35] van der Vaart, A. W., and van Zanten, J. H. Rates of contraction of posterior distributions based on gaussian process priors. *Ann. Statist. 36*, 3 (2008), 1435–1463.

[36] Walther, G. Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc. 97*, 458 (2002), 508–513.

[37] Walther, G. Inference and modeling with log-concave distributions. *Statist. Sci. 24*, 3 (2009), 319–327.

[38] Williamson, R. E. Multiply monotone functions and their Laplace transforms. *Duke Math. J. 23* (1956), 189–207.