Bayesian Analysis of Simple Random Densities

Paulo C. Marques F. and Carlos A. de B. Pereira
Instituto de Matemática e Estatística da Universidade de São Paulo

Abstract. A tractable nonparametric prior over densities is introduced which is closed under sampling and exhibits proper posterior asymptotics.

Keywords: Bayesian nonparametrics, Bayesian density estimation.

1 Introduction

The early 1970's witnessed Bayesian inference going nonparametric with the introduction of statistical models with infinite dimensional parameter spaces; the most conspicuous being the Dirichlet Process (Ferguson 1973), which is a prior on the class of all probability measures over a given sample space that trades great analytical tractability for a reduced support: as shown by Blackwell (1973), its realizations are, almost surely, discrete probability measures. The posterior expectation of a Dirichlet Process is a probability measure that gives positive mass to each observed value of the sample, making the plain Dirichlet Process unsuitable to handle inferential problems such as density estimation. Many extensions and alternatives to the Dirichlet Process have been proposed (Gosh and Ramamoorthi 2002).

In this paper we construct a prior distribution over the class of densities with respect to Lebesgue measure. Given a partition in subintervals of a bounded interval of the real line, we define a random density whose realizations have a constant positive value on each subinterval of the partition. The distribution of the values of the random density on each subinterval is specified by transforming and conditioning a multivariate normal distribution.

Our construction of the random density resembles the stochastic processes introduced by Thorburn (1986) and Lenk (1988), with the following differences. Since our definition relies on a finite dimensional random object, instead of a more general stochastic process, our proofs are simpler, we can represent the random density directly in our numerical computations, instead of keeping its values on a finite number of arbitrarily chosen points, and we do not need to interpolate our estimates. To make the distribution of his random density closed under sampling, Lenk (1988) was forced to introduce a parameter which does not have a natural interpretation, whereas in our case the desired closure follows more naturally, as does the proper asymptotical behavior of our posterior distribution.

An outline of the paper is as follows. In Section 2, we give the formal definition of a simple random density. In Section 3, we prove that the distribution of a simple random density is closed under sampling. The results of the simulations in Section 4 show the asymptotic behavior of the posterior distribution. We extend the model

hierarchically in Section 5 to deal with random partitions. Although the usual Bayes estimate of a simple random density is a discontinuous density, in Section 6 we compute smooth estimates solving a decision problem where the states of nature are realizations of the simple random density and the actions are smooth densities of a suitable class. Additional propositions and proofs of all the results in the paper are given in Section 7.

2 Simple Random Densities

Let $(\Omega, \mathscr{F}, P)$ be the probability space from which we induce the distributions of all random objects considered in the paper. For some integer $k \geq 1$, let $\mathbb{R}^k_+$ be the set of vectors of $\mathbb{R}^k$ with positive components. Write $\mathscr{R}^k$ for the Borel sigma-field of $\mathbb{R}^k$. Let λ_k denote Lebesgue measure over $(\mathbb{R}^k, \mathscr{R}^k)$. We omit the indexes when $k = 1$. The components of a vector $v \in \mathbb{R}^k$ are written as $v_1, \ldots, v_k$.

Suppose that we have been given an interval $[a, b] \subset \mathbb{R}$, and a set of real numbers $\Delta = \{t_0, t_1, \ldots, t_k\}$, such that $a = t_0 < t_1 < \cdots < t_k = b$, inducing a partition of $[a, b]$ into the $k \geq 1$ subintervals $[a, t_1), [t_1, t_2), \ldots, [t_{k-2}, t_{k-1}), [t_{k-1}, b]$. The class of simple densities with respect to this partition consists of the nonnegative simple functions that have a constant value on each subinterval and integrate to one. Let $d_i = t_i - t_{i-1}$, for $i = 1, \ldots, k$, and define $S_\Delta : \mathbb{R}^k \to \mathbb{R}$ by $S_\Delta(u) = \sum_{i=1}^k d_i u_i$. Each simple density $f : \mathbb{R} \to \mathbb{R}$ within this class can be represented as

$$f(x) = \sum_{i=1}^{k} h_i \, I_{[t_{i-1}, t_i)}(x) \,,$$

where $h = (h_1, \ldots, h_k) \in \mathbb{R}^k$ is such that each $h_i \geq 0$, and $S_\Delta(h) = 1$. The h_i's will be called heights of the steps of the simple density f.

From now on, let $\mathbb{H}_r = \{v \in \mathbb{R}^k_+ : d_1 v_1 + \cdots + d_k v_k = r\}$, for $r \in \mathbb{R}$. Note that, by the definition of the d_i's given above, it follows that $\mathbb{H}_r = \emptyset$ if $r \leq 0$. Also, define the projection on the first $k - 1$ coordinates $\pi : \mathbb{R}^k \to \mathbb{R}^{k-1}$ by $\pi(v_1, \ldots, v_{k-1}, v_k) = (v_1, \ldots, v_{k-1})$. For a normal random vector $Z = (Z_1, \ldots, Z_k)$ with mean $m \in \mathbb{R}^k$ and $k \times k$ covariance matrix Σ, denote by $U \sim L_k(m, \Sigma)$ the distribution of the lognormal random vector $U = (e^{Z_1}, \ldots, e^{Z_k})$. If Σ is nonsingular, it is easy to show that U has a density

$$f_U(u) = (2\pi)^{-k/2} \, |\Sigma|^{-1/2} \left(\prod_{i=1}^{k} u_i^{-1} \right) \exp\left(-\frac{1}{2} (\log u - m)^\top \Sigma^{-1} \, (\log u - m) \right) I_{\mathbb{R}^k_+}(u) \,,$$

where $|\Sigma|$ is the determinant of Σ, $\log u = (\log u_1, \ldots, \log u_k)^\top$ and $m = (m_1, \ldots, m_k)^\top$.

We define a random density whose realizations are simple densities with respect to the partition induced by Δ by specifying the distribution of the random vector of its steps heights. Informally, the steps heights will have the distribution of a lognormal random vector U given that $S_\Delta(U) = 1$. The formal definition of the random density is given in terms of a version of the conditional distribution of U given $S_\Delta(U)$ and the

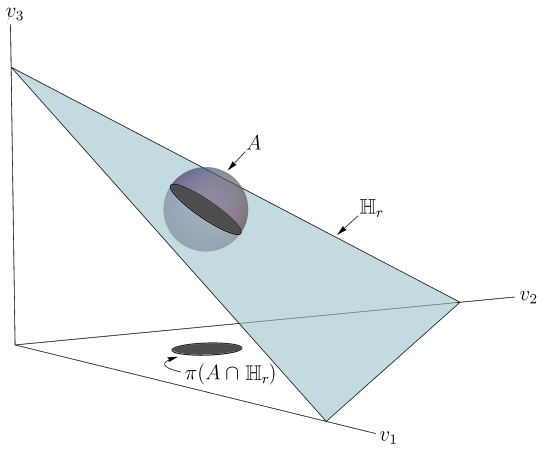


Figure 1: Geometrical interpretation of the measures τ_r of Lemma 2.1, for $r > 0$, in the particular case when $k = 3$. The value of $\tau_r(A)$ is the area of the projection $\pi(A \cap \mathbb{H}_r)$ multiplied by d_3^{-1}.

expression of its conditional density with respect to a dominating measure. However, we are outside the elementary case where the joint distribution is dominated by a product measure. In fact, we have in Proposition 7.1 a simple proof that Lebesgue measure λ_{k+1} and the joint distribution of U and $S_\Delta(U)$ are mutually singular.

A suitable family of measures that dominate the conditional distribution of U given $S_\Delta(U)$, for each value of $S_\Delta(U)$, is described in the following lemma.

Lemma 2.1. *Let $\tau_r : \mathscr{R}^k \to \mathbb{R}$ be defined by $\tau_r(A) = d_k^{-1}\lambda_{k-1}(\pi(A \cap \mathbb{H}_r))$, for $r \in \mathbb{R}$. Then, each τ_r is a measure over $(\mathbb{R}^k, \mathscr{R}^k)$.*

A proof of Lemma 2.1 is given in section 7. Figure 1 gives a simple geometric interpretation of the measures τ_r when the underlying partition is formed by three subintervals.

The following result is the basis for the formal definition of the random density.

Theorem 2.2. *Let $U \sim L_k(m, \Sigma)$, with nonsingular Σ, and let $\{\tau_r\}_{r\in\mathbb{R}}$ be the family of measures over $(\mathbb{R}^k, \mathscr{R}^k)$ defined on Lemma 2.1. Then, $\mu_{U|S_\Delta(U)} : \mathscr{R}^k \times \mathbb{R}_+ \to \mathbb{R}$ defined by*

$$\mu_{U|S_\Delta(U)}(A \mid r) = \int_A \frac{f_U(u)}{f_{S_\Delta(U)}(r)}\, I_{\mathbb{H}_r}(u)\, d\tau_r(u)\,,$$

is a regular version of the conditional distribution of U given $S_\Delta(U)$, where

$$f_{S_\Delta(U)}(r) = \int_{\mathbb{R}^k} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u)\,.$$

Moreover, $\mu_{U|S_\Delta(U)}(\mathbb{H}_r \mid r) = 1$, for each $r > 0$.

The necessary lemmata and a proof of Theorem 2.2 are given in Section 7. The following definition of the random density uses the specific version of the conditional distribution constructed in Theorem 2.2.

Definition 2.3. Let $U \sim L_k(m, \Sigma)$, with nonsingular Σ. We say that $\varphi : \mathbb{R} \times \Omega \to \mathbb{R}$ defined by

$$\varphi(x, \omega) = \sum_{i=1}^{k} H_i(\omega)\, I_{[t_{i-1}, t_i)}(x)$$

is a *simple random density*, where $H = (H_1, \ldots, H_k)$ are the *random heights of the steps* of φ, with distribution given by $\mu_H(A) = \mu_{U|S_\Delta(U)}(A \mid 1)$, for $A \in \mathscr{R}^k$, where $\mu_{U|S_\Delta(U)}$ is the regular version of the conditional distribution of U given $S_\Delta(U)$ obtained in Theorem 2.2. Hence, for every $A \in \mathscr{R}^k$, we have

$$\mu_H(A) = \int_A \frac{f_U(h)}{f_{S_\Delta(U)}(1)}\, I_{\mathbb{H}_1}(h)\, d\tau_1(h)\,,$$

where $\tau_1(A) = d_k^{-1} \lambda_{k-1}(\pi(A \cap \mathbb{H}_1))$ and it holds that $\mu_H(\mathbb{H}_1) = 1$. We use the notation $\varphi \sim \Delta(m, \Sigma)$.

3 Conditional Model

Now we model a set of absolutely continuous observables conditionally, given the value of a simple random density φ. The following lemma, proved in Section 7, describes the conditional model and determines the form of the likelihood.

Lemma 3.1. *Let $\varphi \sim \Delta(m, \Sigma)$ with representation $\varphi(x, \omega) = \sum_{i=1}^{k} H_i(\omega)\, I_{[t_{i-1}, t_i)}(x)$. Suppose that the random variables $X_1, \ldots, X_n$ are conditionally independent and identically distributed, given that $H = h$, with distribution $\mu_{X_1|H}(A \mid h) = \int_A f(y)\, d\lambda(y)$, where we have defined $f(y) = \sum_{i=1}^{k} h_i\, I_{[t_{i-1}, t_i)}(y)$. Define $X = (X_1, \ldots, X_n)$ and let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Then, $\mu_{X|H}(\,\cdot \mid h) \ll \lambda_n$, almost surely $[\mu_H]$, with Radon-Nikodym derivative*

$$\frac{d\mu_{X|H}}{d\lambda_n}(x \mid h) = f_{X|H}(x \mid h) = \prod_{i=1}^{k} h_i^{c_i}\,,$$

where $c_i = \sum_{j=1}^{n} I_{[t_{i-1}, t_i)}(x_j)$, for $i = 1, \ldots, k$.

The factorization criterion yields that $c = (c_1, \ldots, c_n)$ is a sufficient statistic for φ. That is, in this conditional model, as one should expect, all the sample information is contained in the countings of how many sample points belong to each subinterval of the partition induced by Δ.

Using the notation of Lemma 3.1, and defining $c = (c_1, \ldots, c_k)^\top$, we can prove that the prior distribution of φ is closed under sampling.

Theorem 3.2. *if $\varphi \sim \Delta(m, \Sigma)$, then $\varphi \mid X = x \sim \Delta(m^*, \Sigma)$, where $m^* = m + \Sigma c$.*

This result, proved in Section 7, has practical consequences, as it makes the simulations of prior and posterior distributions essentially the same, the only difference being the computation of m^*.

4 Stochastic Simulations

We summarize the distribution of a simple random density $\varphi \sim \Delta(m, \Sigma)$, represented as $\varphi(x, \omega) = \sum_{i=1}^{k} H_i(\omega)\, I_{[t_{i-1}, t_i)}(x)$, in two ways. First, motivated by the fact, proved in Proposition 7.5, that the prior and posterior expectations are predictive densities, we take as an estimate the expectation of the steps heights $\hat{h} = (\mathrm{E}[H_1], \ldots, \mathrm{E}[H_k])$. Second, the uncertainty of this estimate is assessed defining

$$B(\hat{h}, \epsilon) = \left\{ h \in \mathbb{H}_1 : d(\hat{h}, h) < \epsilon \right\} ,$$

for $\epsilon > 0$, and taking as a credible set the $B(\hat{h}, \epsilon)$ with the smallest positive ϵ such that $P\{\omega : H(\omega) \in B(\hat{h}, \epsilon)\} = \gamma$, where $\gamma \in (0, 1)$ is the credibility level.

The Random Walk Metropolis algorithm (Robert and Casella 2004) is used to draw dependent realizations of the steps of φ as values of a Markov chain $\{H^{(i)}\}_{i \geq 0}$. The two summaries are computed through ergodic means of this chain. For example, the credible set is determined with the help of the almost sure convergence of

$$\frac{1}{N} \sum_{i=0}^{N} I_{B(\hat{h}, \epsilon)} \left(H^{(i)} \right) \xrightarrow[N \to \infty]{} \mathrm{E} \left[I_{B(\hat{h}, \epsilon)}(H) \right] = P \left\{ \omega : H(\omega) \in B(\hat{h}, \epsilon) \right\} .$$

As for the parameters appearing in Definition 2.3, we take in our experiments all the m_i's equal to one, and the covariance matrix $\Sigma = (\sigma_{ij})$ is chosen in the following way. Given some positive definite covariance function $C : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we induce Σ from C defining

$$\sigma_{ij} = C\left(\frac{t_{i-1} + t_i}{2}, \frac{t_{j-1} + t_j}{2} \right) ,$$

for $i, j = 1, \ldots, k$. In our examples we study the family of Gaussian covariance functions defined by $C_{\rho,\theta}(x, y) = \rho\, e^{-\theta\,(x-y)^2}$, with dispersion parameter $\rho > 0$ and scale parameter $\theta > 0$.

Example 4.1. Let $\varphi \sim \Delta(m, \Sigma)$ and consider the sample space $[0, 1]$ with $\Delta = \{0, 0.01, 0.02, \ldots, 0.98, 0.99, 1\}$. For the sake of generality, we induce Σ from the family of Gaussian covariance functions with fixed dispersion parameter ρ_0 but with random scale parameter $\Theta = Y + 20\,000$, where $Y \sim Gamma(2, 0.001)$. These choices guarantee that computations with Σ are numerically stable. In Figure 2, the summaries of the prior distribution of φ show that the value of ρ_0 controls the concentration of the prior. Fixing $\rho_0 = 0.05$ and generating data from a mixture

$$\frac{1}{3} \cdot Beta(1, 10) + \frac{1}{3} \cdot Beta(10, 10) + \frac{1}{3} \cdot Beta(30, 5) ,$$

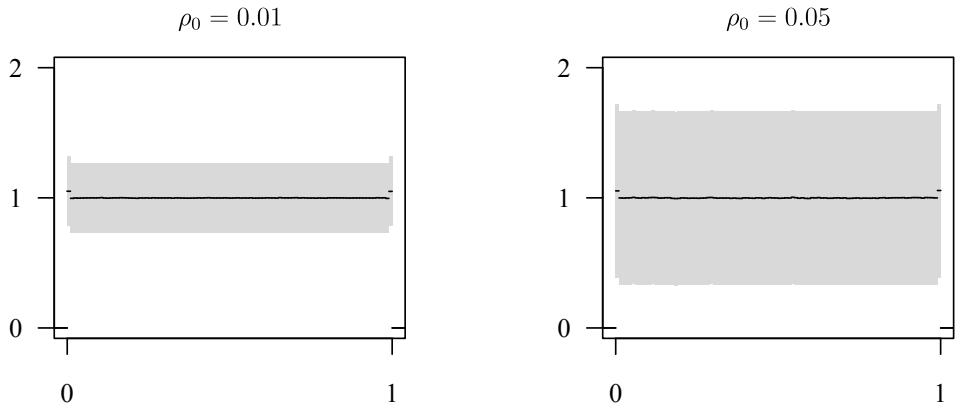


Figure 2: Effect of the value of ρ_0 on the concentration of the prior. The curves in black are prior expectations and the gray regions are credible sets with credibility level of 95%.

we have in Figure 3 the posterior summaries for different sample sizes. Note the concentration of the posterior as we increase the size of the samples. ∎

We observe the same asymptotic behavior of the posterior distribution with data coming from a triangular distribution and a mixture of normals, where in the second case we truncate the sample space appropriately.

5 Random Partitions

Inferentially, we have a richer construction when the definition of the simple random density involves a random partition. Informally, we want a model for the random density where the underlying partition adapts itself according to the information contained in the data.

We consider a family of uniform partitions of a given interval $[a, b]$. Each partition of this family will be described by a positive integer random variable K, which determines the number of subintervals in the partition. Since the parameter ρ of the family of Gaussian covariance functions used to induce Σ may have different meanings for different partitions, we treat it as a positive random variable R.

Explicitly, we are considering the following hierarchical model: K and R are independent. Given that $K = k$ e $R = \rho$, we choose the uniform partition of the interval $[a, b]$ induced by

$$\Delta = \left\{ a, a + \frac{b-a}{k}, a + \frac{2(b-a)}{k}, \ldots, a + \frac{(k-1)(b-a)}{k}, b \right\},$$

induce $\Sigma_{\rho,\theta}$ from the family of Gaussian covariance functions, and make $\varphi \sim \Delta(m, \Sigma_{\rho,\theta})$. Finally, the observables are modeled as in Lemma 3.1. This hierarchy is described in

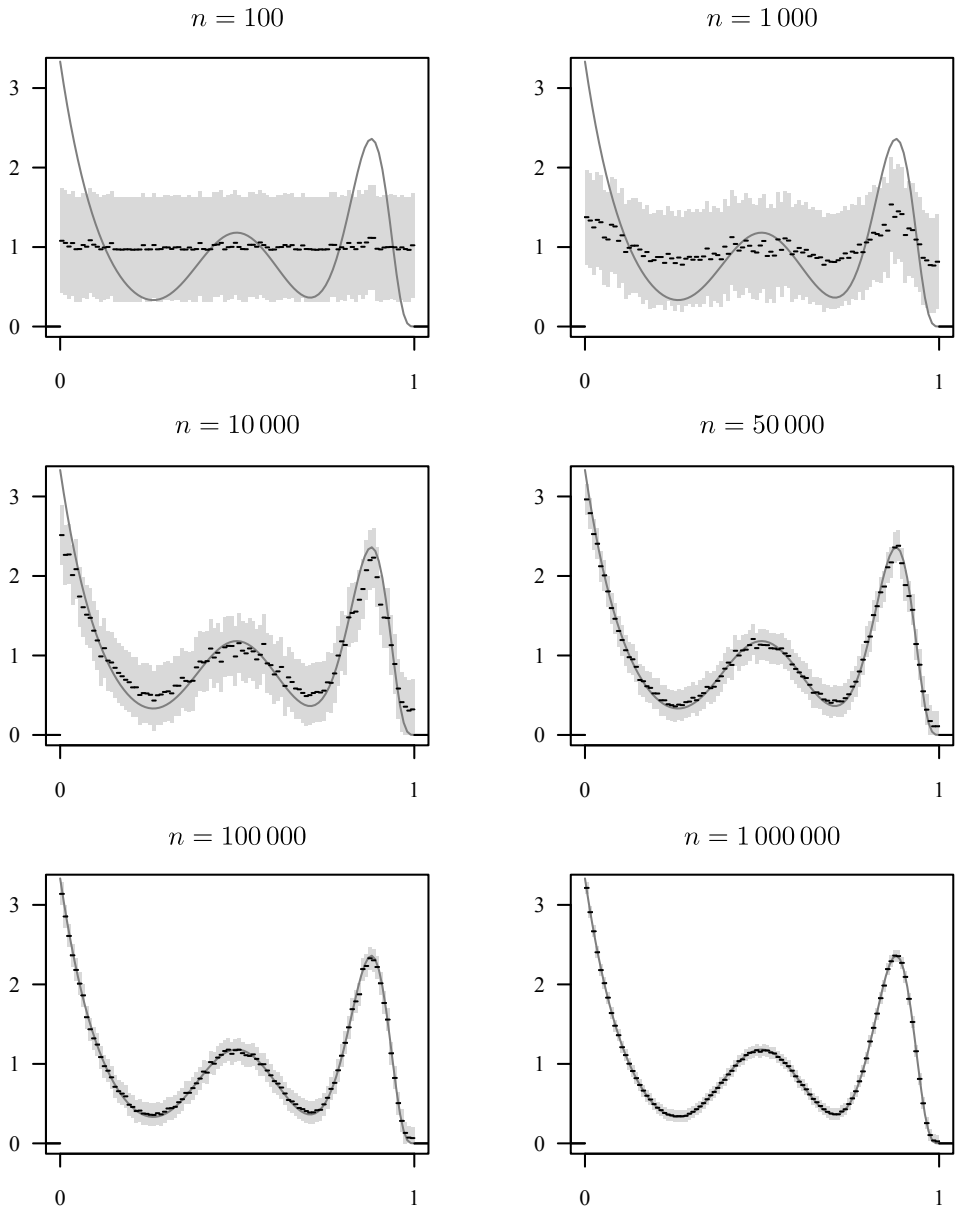


Figure 3: Posterior summaries for Example 4.1. On each graph, the black simple density is the estimate $\hat{\varphi}$, the light gray region is a credible set with credibility level of 95%, and the dark gray curve is the data generating density.

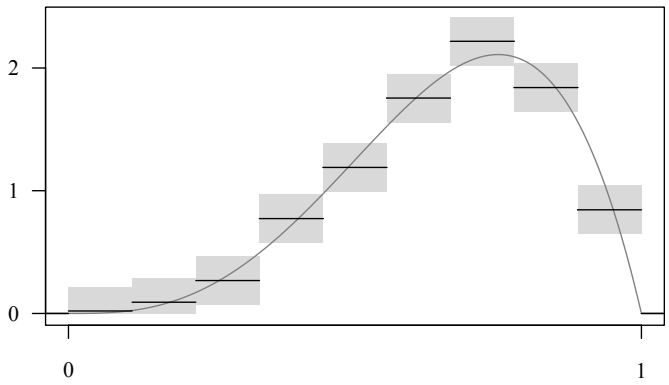
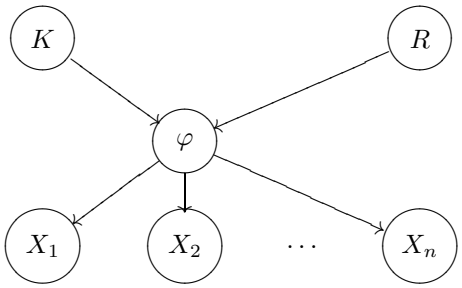


Figure 4: Posterior summaries for Example 5.1. The black simple density is the estimate $\hat{\varphi}$, the light gray region is a credible set with credibility 95%, and the dark gray curve is the data generating density.

the following graph.



In the following example we follow an empirical path: instead of specifying priors for K and R, we define the likelihood of K and R by $L_x(k,\rho) = f_{X|K,R}(x \mid k,\rho)$, whose form is determined in Proposition 7.6, find the maximum $(\hat{k},\hat{\rho}) = \arg\max_{k,\rho} L_x(k,\rho)$, and use these values in the definitions of the prior, determining the posterior summaries as we did in Section 4.

Example 5.1. With a sample of size 2 000 generated from a $Beta(4,2)$ distribution, we find the maximum of the likelihood of K and R at $(\hat{k},\hat{\rho}) = (9, 1.43)$. In Figure 4 we have the posterior summaries obtained using these values in the definition of the prior. Moreover, in the left graph of Figure 5 we have the distribution function $\hat{F}$ corresponding to the estimated posterior density. For the sake of comparison, we plot in the right graph of Figure 5 some quantiles of this distribution $\hat{F}$ against the quantiles of the distribution F_0 from which we generated the data. ∎

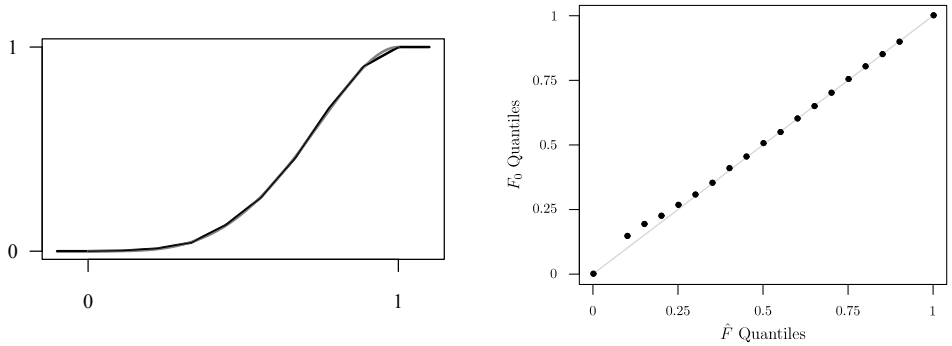


Figure 5: Example 5.1. On the left graph, the black curve is the estimated distribution function $\hat{F}$ and the gray curve is the data generating distribution function F_0. On the right graph, we have the comparison of some of the quantiles of $\hat{F}$ and F_0.

6 Smooth Estimates

It is possible to go beyond the discontinuous densities obtained as estimates in the last two sections and get smooth estimates of a simple random density φ solving a Bayesian decision problem where the states of nature are the realizations of φ and the actions are smooth densities of a suitable class.

In view of Theorem 3.2, its enough to consider the problem without data. As before, the sample space is the interval $[a, b]$, which is partitioned according with some Δ. For some density f with respect to Lebesgue measure, we denote its L_2 norm by $\|f\|_2 = \left(\int f^2 d\lambda\right)^{1/2}$.

Proposition 6.1. *For $N \geq 1$, let $g_1, \ldots, g_N$ be densities with respect to Lebesgue measure, with support $[a, b]$, such that $\|g_i\|_2 < \infty$, and let $\mathscr{D}$ be the class of densities of the form $\sum_{i=1}^N \alpha_i g_i$, with $\alpha_i \geq 0$, for $i = 1, \ldots, N$, and $\sum_{i=1}^N \alpha_i = 1$. Let $\varphi \sim \Delta(m, \Sigma)$ and define $\mathscr{S}$ as the class of densities which are realizations of φ. Define the loss function $L : \mathscr{S} \times \mathscr{D} \to \mathbb{R}$ by*

$$L(s, f) = \|s - f\|_2^2 = \int_a^b \left(s(x) - f(x)\right)^2 \, d\lambda(x) \, .$$

Then, the Bayes decision is $\hat{\varphi} = \sum_{i=1}^N \hat{\alpha}_i \, g_i$, where $\hat{\alpha}_i$ minimize globally the quadratic form

$$Q = \sum_{i,j=1}^N \alpha_i \alpha_j \, M_{ij} - \sum_{i=1}^N \alpha_i \, J_i \, ,$$

subject to the constraints $\alpha_i \geq 0$, for $i = 1, \ldots, N$, and $\sum_{i=1}^N \alpha_i = 1$, with the definitions

$$M_{ij} = \int_a^b g_i(x) g_j(x) \, d\lambda(x) \qquad e \qquad J_i = 2 \int_a^b g_i(x) \mathrm{E}[\varphi(x)] \, d\lambda(x) \, .$$

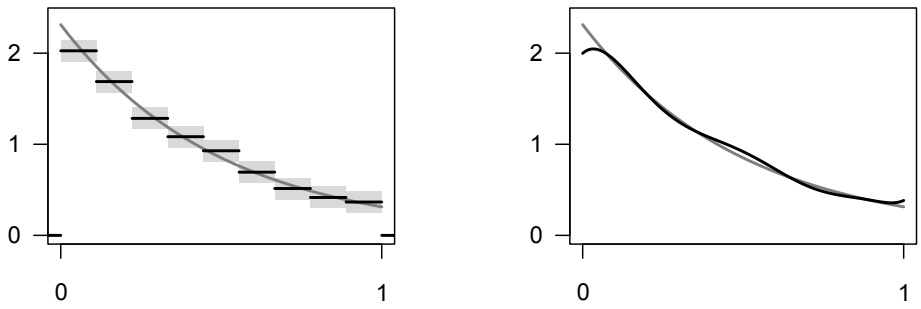


Figure 6: Example 6.2. On the right graph, the black simple density is the estimate $\hat{\varphi}$, and the light gray region is a credible set with credibility 95%. On both graphs the dark gray curve is the data generating density. On the left graph, the black smooth density is the Bayes decision of Proposition 6.1.

We use the result of Proposition 6.1, proved in Section 7, choosing the g_i's inside a class of smooth densities that serve approximately as a basis to represent any continuous density with the specified support.

For the next example, suppose that the support of the densities is the interval $[0, 1]$. Bernstein's Theorem (see Billingsley (1995), Theorem 6.2) states that the polynomial

$$B_N(x) = \sum_{i=0}^{N} f\left(\frac{i}{N}\right) \binom{N}{i} x^i (1-x)^{N-i}$$

approximates uniformly any continuous function f defined on $[0, 1]$, when $N \to \infty$. Suppose that f is a density. If we define, for $i = 0, \ldots, N$,

$$\alpha_i = f\left(\frac{i}{N}\right) \binom{N}{i} \frac{\Gamma(i+1)\Gamma(N-i+1)}{\Gamma(N+2)},$$

we can rewrite the approximating polynomial as $B_N(x) = \sum_{i=0}^{N} \alpha_i \, g_i(x)$, where g_i is a density of a random variable with distribution $Beta(i+1, N-i+1)$. Hence, if we take a sufficiently large N, we expect that any continuous density with support $[0, 1]$ will be reasonably approximated by a mixture of these g_i's.

Example 6.2. Suppose that we have a sample of 5 000 data simulated from a truncated exponential distribution, whose density is

$$f_0(x) = \frac{2\, e^{-2(x-1)}}{e^2 - 1} \; I_{[0,1]}(x)\,.$$

Repeating the analysis made in Example 5.1, we find the maximum of the likelihood of K e R at $(\hat{k}, \hat{\rho}) = (9, 0.86)$. The left graph of Figure 6 presents the posterior summaries. After that, we solved the problem of constrained optimization of Proposition 6.1 and found the results shown in the right graph of Figure 6. ∎

7 Additional Results and Proofs

In this section we present some auxiliary propositions and give proofs to all the results stated in the paper.

Proposition 7.1. *Let $U \sim L_k(m, \Sigma)$ and denote by $\mu_{U,S_\Delta(U)}$ the joint distribution of U and $S_\Delta(U)$. Then, $\mu_{U,S_\Delta(U)} \perp \lambda_{k+1}$.*

Proof. Define the set $A = \left\{ v \in \mathbb{R}^{k+1} : \sum_{i=1}^{k} d_i v_i = v_{k+1} \right\} \in \mathscr{R}^{k+1}$. Then,

$$\mu_{U,S_\Delta(U)}(A) = P\left\{\omega : (U(\omega), S_\Delta(U(\omega))) \in A\right\} = P\left\{ \omega : \sum_{i=1}^{k} d_i U_i(\omega) = S_\Delta(U(\omega)) \right\} = 1,$$

by definition of S_Δ. On the other hand, note that $\lambda_{k+1}(A) = 0$, since this is the $(k+1)$-volume of the k-dimensional hyperplane defined by the set A. Since $\mu_{U,S_\Delta(U)}(A^c) = 0$, the result follows. ∎

Proof of Lemma 2.1. When $r \leq 0$, the result is trivial, since in this case $\mathbb{H}_r = \emptyset$, making τ_r a null measure. Suppose that $r > 0$ and let $g : \mathbb{R}^k \to \mathbb{R}^k$ be the function defined by

$$g(v) = \left(v_1, \ldots, v_{k-1}, \frac{1}{d_k} \left(v_k - \sum_{i=1}^{k-1} d_i v_i \right) \right).$$

Define $h_r : \mathbb{R}^{k-1} \to \mathbb{R}^k$ by $h_r(y) = g(y, r)$. We will show that $\pi(A \cap \mathbb{H}_r) = h_r^{-1}(A)$, for every $A \in \mathscr{R}$. Suppose that $y \in \pi(A \cap \mathbb{H}_r)$. Then, there is a $v \in A \cap \mathbb{H}_r$ such that $y = \pi(v) = (v_1, \ldots, v_{k-1})$ and

$$h_r(y) = g(y, r) = \left(v_1, \ldots, v_{k-1}, \frac{1}{d_k} \left(r - \sum_{i=1}^{k-1} d_i v_i \right) \right).$$

Since $v \in \mathbb{H}_r$, we have that $\frac{1}{d_k}\left(r - \sum_{i=1}^{k-1} d_i v_i \right) = v_k$, implying that $h_r(y) = v$. Since $v \in A$, it follows from the definition of the inverse image of h_r that $y \in h_r^{-1}(A)$ and, therefore, we conclude that $\pi(A \cap \mathbb{H}_r) \subset h_r^{-1}(A)$. To prove the other inclusion, suppose that $y \in h_r^{-1}(A)$ and define $v = h_r(y)$. Hence, $v \in A$ and by the definition of h_r we have that

$$v = g(y, r) = \left(y_1, \ldots, y_{k-1}, \frac{1}{d_k} \left(r - \sum_{i=1}^{k-1} d_i y_i \right) \right),$$

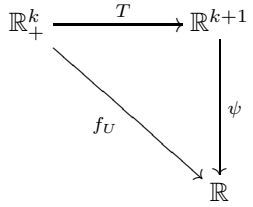
implying that $v \in \mathbb{H}_r$, because $\sum_{i=1}^{k} d_i v_i = r$. Since $v \in A \cap \mathbb{H}_r$ and $y = \pi(v)$, it follows that $y \in \pi(A \cap \mathbb{H}_r)$. Therefore, $h_r^{-1}(A) \subset \pi(A \cap \mathbb{H}_r)$. Hence, we have that $\tau_r = d_k^{-1} \lambda_k \circ h_r^{-1}$ and the usual properties of the inverse image of h_r and the Lebesgue measure entail that each τ_r is a measure over $(\mathbb{R}^k, \mathscr{R}^k)$. ∎

Lemma 7.2. *Let $U \sim L_k(m, \Sigma)$. Let ξ, defined by $\xi(A) = \lambda_k\{u \in \mathbb{R}^k_+ : (u, S_\Delta(u)) \in A\}$, be a measure over $(\mathbb{R}^{k+1}, \mathscr{R}^{k+1})$. Denote by $\mu_{U,S_\Delta(U)}$ the joint distribution of U and $S_\Delta(U)$. Then, we have that $\mu_{U,S_\Delta(U)} \ll \xi$, with Radon-Nikodym derivative $d\mu_{U,S_\Delta(U)}/d\xi = f_{U,S_\Delta(U)}$ given by*

$$f_{U,S_\Delta(U)}(u, r) = f_U(u)\, I_{\mathbb{H}_r}(u)\,,$$

where $u \in \mathbb{R}^k$ and $r \in \mathbb{R}$.

Proof. Define the function $T : \mathbb{R}^k_+ \to \mathbb{R}^{k+1}$ by $T(u) = (u, S_\Delta(u))$. Note that $\xi = \lambda_k \circ T^{-1}$. Define the function $\psi : \mathbb{R}^{k+1} \to \mathbb{R}$ by $\psi(u, r) = f_U(u)\, I_{\mathbb{H}_r}(u)$, with $u \in \mathbb{R}^k$ and $r \in \mathbb{R}$. The diagram



commutes, since $\psi(T(u)) = \psi(u, S_\Delta(u)) = f_U(u)\, I_{\mathbb{H}_{S_\Delta(U)}}(u) = f_U(u)$, for every $u \in \mathbb{R}^k_+$. For every $A \in \mathscr{R}^{k+1}$, we have that

$$\begin{aligned}
\mu_{U,S_\Delta(U)}(A) &= P\{\omega : (U(\omega), S_\Delta(U(\omega))) \in A\} = P\{\omega : U(\omega) \in T^{-1}(A)\} \\
&= \int_{T^{-1}(A)} f_U(u)\, d\lambda_k(u) = \int_{T^{-1}(A)} \psi(T(u))\, d\lambda_k(u) \\
&= \int_A \psi(u, r)\, d\xi(u, r) = \int_A f_U(u)\, I_{\mathbb{H}_r}(u)\, d\xi(u, r)\,,
\end{aligned}$$

where the fifth equality is obtained transforming by T, $u \in \mathbb{R}^k$ and $r \in \mathbb{R}$. It follows that $\mu_{U,S_\Delta(U)} \ll \xi$ and the Radon-Nikodym derivative has the desired expression. ∎

Lemma 7.3. *Let ξ be the measure defined on Lemma 7.2 and let $\{\tau_r\}_{r\in\mathbb{R}}$ be the family of measures defined on Lemma 2.1. Then, for every measurable nonnegative $\psi : \mathbb{R}^{k+1} \to \mathbb{R}$, we have that*

$$\int_{\mathbb{R}^{k+1}} \psi(u, r)\, d\xi(u, r) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}^k} \psi(u, r)\, d\tau_r(u) \right) d\lambda(r)\,,$$

where $u \in \mathbb{R}^k$ and $r \in \mathbb{R}$.

Proof. Define $f : \mathbb{R}^k \to \mathbb{R}^k$ by $f(u) = (u_1, \ldots, u_{k-1}, \sum_{i=1}^k d_i u_i)$. Hence, f is a differentiable function whose inverse is the differentiable function g defined on Lemma 2.1. The value of the Jacobian on the point $v \in \mathbb{R}^k$ is $J_g(v) = d_k^{-1}$. Let $A \in \mathscr{R}^k$, $y \in \mathbb{R}^{k-1}$, $r \in \mathbb{R}$, and define h_r as in Lemma 2.1. When $r > 0$, we have already shown in the course of the proof of Lemma 2.1 that $\pi(A \cap \mathbb{H}_r) = h_r^{-1}(A)$, for every $A \in \mathscr{R}^k$. Remembering

that, by definition, $\mathbb{H}_r \subset \mathbb{R}_+^k$, it follows that $\pi(A \cap \mathbb{H}_r) = h_r^{-1}(A \cap \mathbb{R}_+^k)$ and we conclude that $I_{\pi(A\cap\mathbb{H}_r)}(y) = I_{A\cap\mathbb{R}_+^k}(g(y,r))$. Now suppose that $r \leq 0$. In this case, since $\mathbb{H}_r = \emptyset$, we have that $I_{\pi(A\cap\mathbb{H}_r)}(y) = I_\emptyset(y) = 0$. As for the value of $I_{A\cap\mathbb{R}_+^k}(g(y,r))$, consider two subcases: since

$$g(y,r) = \left(y_1, \ldots, y_{k-1}, \frac{1}{d_k}\left(r - \sum_{i=1}^{k-1} d_i y_i\right)\right),$$

if any of the $y_i \leq 0$, then $I_{A\cap\mathbb{R}_+^k}(g(y,r)) = 0$, otherwise, we have $\frac{1}{d_k}\left(r - \sum_{i=1}^{k-1} d_i y_i\right) < 0$ and again it happens that $I_{A\cap\mathbb{R}_+^k}(g(y,r)) = 0$. Therefore, we conclude that in this case also $I_{\pi(A\cap\mathbb{H}_r)}(y) = I_{A\cap\mathbb{R}_+^k}(g(y,r))$. Hence, for $A \in \mathscr{R}^k$ and $B \in \mathscr{R}$, we have that

$$\begin{aligned}
\xi(A\times B) &= \lambda_k\{u \in \mathbb{R}_+^k : u \in A, S_\Delta(u) \in B\} = \int_{\mathbb{R}^k} I_{A\cap\mathbb{R}_+^k}(u)\, I_B(S_\Delta(u))\, d\lambda_k(u) \\
&= \int_{\mathbb{R}^k} I_{A\cap\mathbb{R}_+^k}(g(y,r))\, I_B(r)\, |J_g(y,r)|\, d\lambda_k(y,r) \\
&= \int_{\mathbb{R}^k} d_k^{-1}\, I_{\pi(A\cap\mathbb{H}_r)}(y) I_B(r)\, d\lambda_k(y,r) \\
&= \int_B \left(d_k^{-1}\int_{\pi(A\cap\mathbb{H}_r)} d\lambda_{k-1}(y)\right) d\lambda(r) = \int_B \tau_r(A)\, d\lambda(r)\,,
\end{aligned}$$

where $y \in \mathbb{R}^{k-1}$ and $r \in \mathbb{R}$, the third equality is obtained transforming by f, and the penultimate is a consequence of Tonelli's Theorem. The result follows from the Product Measure Theorem and Fubini's Theorem (see Ash (2000), Theorems 2.6.2 e 2.6.4). ∎

Lemma 7.4. *Let $U \sim L_k(m,\Sigma)$. Let $\{\tau_r\}_{r\in\mathbb{R}}$ be the family of measure defined on Lemma 2.1. Let $\mu_{S_\Delta(U)}$ be the distribution of $S_\Delta(U)$. Then, $\mu_{S_\Delta(U)} \ll \lambda$ with Radon-Nikodym derivative $d\mu_{S_\Delta(U)}/d\lambda = f_{S_\Delta(U)}$ given by $f_{S_\Delta(U)}(r) = \int_{\mathbb{R}^k} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u)$.*

Proof. Let $A \in \mathscr{R}$, $u \in \mathbb{R}^k$, and $r \in \mathbb{R}$. Let ξ be the measure defined on Lemma 7.2. We have that

$$\begin{aligned}
\mu_{S_\Delta(U)}(A) &= P\{\omega : S_\Delta(U(\omega)) \in A\} = P\{\omega : U(\omega) \in \mathbb{R}^k, S_\Delta(U(\omega)) \in A\} \\
&= \mu_{U,S_\Delta(U)}(\mathbb{R}^k \times A) = \int_{\mathbb{R}^k\times A} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\xi(u,r) \\
&= \int_A \left(\int_{\mathbb{R}^k} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u)\right) d\lambda(r)\,,
\end{aligned}$$

where the penultimate equality follows from Lemma 7.2, and the last equality follows from Lemma 7.3. Hence, $\mu_{S_\Delta(U)} \ll \lambda$ and the Radon-Nikodym derivative has the desired expression. ∎

Proof of Theorem 2.2. Let $\mu_{U,S_\Delta(U)}$ be the joint distribution of U and $S_\Delta(U)$, and let $\mu_{S_\Delta(U)}$ be the distribution of $S_\Delta(U)$. For $A \in \mathscr{R}^k$ and $B \in \mathscr{R}$, by the definition of conditional distribution, we have that

$$\begin{aligned}\mu_{U,S_\Delta(U)}(A \times B) = P\{U \in A, S_\Delta(U) \in B\} &= \int_B \mu_{U|S_\Delta(U)}(A \mid r)\, d\mu_{S_\Delta(U)}(r) \\ &= \int_B \mu_{U|S_\Delta(U)}(A \mid r)\, \frac{d\mu_{S_\Delta(U)}}{d\lambda}(r)\, d\lambda(r)\,,\end{aligned}$$

where we have used the Leibniz rule for the Radon-Nikodym derivatives. On the other hand, by Lemmas 7.2 and 7.3, we have that

$$\begin{aligned}\mu_{U,S_\Delta(U)}(A \times B) &= \int_{A\times B} f_U(u)\, I_{\mathbb{H}_r}(u)\, d\xi(u,r) \\ &= \int_B \left(\int_A f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u) \right) d\lambda(r)\,,\end{aligned}$$

with $u \in \mathbb{R}^k$ and $r \in \mathbb{R}$. Both expressions for $\mu_{U,S_\Delta(U)}(A \times B)$ are compatible if

$$\mu_{U|S_\Delta(U)}(A \mid r) = \frac{\displaystyle\int_A f_U(u)\, I_{\mathbb{H}_r}(u)\, d\tau_r(u)}{f_{S_\Delta(U)}(r)}\,,$$

for almost every r $[\lambda]$. Therefore, we have that $\mu_{U|S_\Delta(U)}(\,\cdot \mid r) \ll \tau_r$, for almost every $r > 0$ $[\lambda]$, with Radon-Nikodym derivative $d\mu_{U|S_\Delta(U)}/d\tau_r = f_{U|S_\Delta(U)}(\,\cdot \mid r)$ given by

$$f_{U|S_\Delta(U)}(u \mid r) = \frac{f_U(u)}{f_{S_\Delta(U)}(r)}\, I_{\mathbb{H}_r}(u)\,,$$

as desired. The fact that $\mu_{U|S_\Delta(U)}(\mathbb{H}_r \mid r) = 1$ follows immediately. ∎

Proof of Lemma 3.1. Let α_h be the measures over $(\mathbb{R}^n, \mathscr{R}^n)$ defined by $\alpha_h(A) = \int_A \left(\prod_{i=1}^k h_i^{c_i} \right) d\lambda_n(x)$, for each $h \in \mathbb{H}_1$. Let $B = B_1 \times \cdots \times B_n$, with $B_i \in \mathscr{R}$, for $i = 1, \ldots, n$. By the hypothesis of conditional independence and Tonelli's Theorem, we have that

$$\begin{aligned}\mu_{X|H}(B \mid h) = \prod_{j=1}^n \mu_{X_j|H}(B_j \mid h) = \prod_{j=1}^n \int_{B_j} f(x_j)\, d\lambda(x_j) &= \int_B \left(\prod_{j=1}^n f(x_j) \right) d\lambda_n(x) \\ = \int_B \left(\prod_{j=1}^n \sum_{i=1}^k h_i\, I_{[t_{i-1},t_i)}(x_j) \right) d\lambda_n(x) &= \int_B \left(\prod_{i=1}^k h_i^{c_i} \right) d\lambda_n(x) = \alpha_h(B)\,.\end{aligned}$$

Hence, $\mu_{X|H}(\,\cdot \mid h)$ and α_h agree on the π-system of product sets that generate $\mathscr{R}^n$. Therefore, by Theorem A.26 of Schervish (1995), both measures agree on the whole sigma-field $\mathscr{R}^n$. It follows that $\mu_{X|H}(\,\cdot \mid h) \ll \lambda_n$, almost surely $[\mu_H]$, and the Radon-Nikodym derivative has the desired expression. ∎

Proof of Theorem 3.2. By Bayes Theorem, for each $A \in \mathscr{R}^k$, we have that

$$\begin{aligned}\mu_{H|X}(A \mid x) &= C_0 \int_A f_{X|H}(x \mid h)\, d\mu_H(h) = C_0 \int_A \left(\prod_{i=1}^k h_i^{c_i}\right) d\mu_H(h) \\ &= C_0 \int_A \left(\prod_{i=1}^k h_i^{c_i}\right) \frac{d\mu_H}{d\tau_1}(h)\, d\tau_1(h) \\ &= \frac{C_0}{f_{S_\Delta(U)}(1)} \int_A \left(\prod_{i=1}^k h_i^{c_i}\right) f_U(h)\, I_{\mathbb{H}_1}(h)\, d\tau_1(h)\,,\end{aligned}$$

where we have used the expression of the likelihood obtained on Lemma 3.1, the Leibniz rule for the Radon-Nikodym derivatives, the expression of $d\mu_H/d\tau_1$ in Definition 2.3, and the constant C_0 is such that$\mu_{H|X}(\mathbb{H}_1 \mid x) = 1$. The remainder of the proof depends on some matrix algebra. Let I be the identity matrix. Since, by definition, Σ is symmetric , we have that $I = I^\top = (\Sigma\Sigma^{-1})^\top = (\Sigma^{-1})^\top\Sigma^\top = (\Sigma^{-1})^\top\Sigma$. Therefore, we have that $(\Sigma^{-1})^\top = \Sigma^{-1}$. Write $l = \log h$. Since the scalar $l^\top\Sigma^{-1}m$ is equal to its transpose $(l^\top\Sigma^{-1}m)^\top = m^\top\Sigma^{-1}l$, we have that $(l-m)^\top\Sigma^{-1}(l-m) = l^\top\Sigma^{-1}l - 2m^\top\Sigma^{-1}l + m^\top\Sigma^{-1}m$. Defining $d = \Sigma\, c$, we have

$$\begin{aligned}&\left(\prod_{i=1}^k h_i^{c_i}\right) \exp\left(-\frac{1}{2}(l-m^*)^\top\Sigma^{-1}(l-m^*)\right) \\ &\quad= \exp\left(-\frac{1}{2}\left(-2d^\top\Sigma^{-1}l + l^\top\Sigma^{-1}l - 2m^\top\Sigma^{-1}l + m^\top\Sigma^{-1}m\right)\right) \\ &\quad= C_1 \exp\Bigg(-\frac{1}{2}\left(-2d^\top\Sigma^{-1}l + l^\top\Sigma^{-1}l - 2m^\top\Sigma^{-1}l + m^\top\Sigma^{-1}m\right) \\ &\qquad\qquad + 2m^\top\Sigma^{-1}d + d^\top\Sigma^{-1}d\Bigg)\,,\end{aligned}$$

with $C_1 = \exp\left(-(1/2)\left(-2m^\top\Sigma^{-1}d - d^\top\Sigma^{-1}d\right)\right)$. Define $m^* = m + d$. Since the scalar $d^\top\Sigma^{-1}m = (d^\top\Sigma^{-1}m)^\top = m^\top\Sigma^{-1}d$, we have that $(m^*)^\top\Sigma^{-1}m^* = m^\top\Sigma^{-1}m + 2m^\top\Sigma^{-1}d + d^\top\Sigma^{-1}d$. Hence, we obtain

$$\begin{aligned}&\left(\prod_{i=1}^k h_i^{c_i}\right) \exp\left(-\frac{1}{2}(l-m^*)^\top\Sigma^{-1}(l-m^*)\right) \\ &\qquad= C_1 \exp\left(-\frac{1}{2}\left(l^\top\Sigma^{-1}l - 2(m^*)^\top\Sigma^{-1}l + (m^*)^\top\Sigma^{-1}m^*\right)\right) \\ &\qquad= C_1 \exp\left(-\frac{1}{2}(l-m^*)^\top\Sigma^{-1}(l-m^*)\right)\,.\end{aligned}$$

Using this result in the expression of $\mu_{H|X}$ together with the expression of f_U, we have

$$\mu_{H|X}(A \mid x) = C_2 \int_A f_{U^*}(h)\, I_{\mathbb{H}_1}(h)\, d\tau_1(h)\,,$$

where $C_2 = (C_0\, C_1)/f_{S_\Delta(U)}(1)$ and f_{U^*} is a density of the random vector $U^* \sim L_k(m^*, \Sigma)$. We conclude that, given that $X = x$, the vector H has the distribution of the heights of the steps of a simple random density $\varphi^* \sim \Delta(m^*, \Sigma)$, as desired. ∎

Proposition 7.5. *Suppose that the random variables $X_1, \ldots, X_{n+1}$ are modeled according to Lemma 3.1. Denote by μ_{X_i} the distribution of X_i, for $i = 1, \ldots, n+1$. For convenience, use the notations $X^{(n)} = (X_1, \ldots, X_n)$ and $x^{(n)} = (x_1, \ldots, x_n) \in \mathbb{R}^k$. Then, for every $A \in \mathscr{R}$, we have*

(a) $\mu_{X_i}(A) = \displaystyle\int_A \mathrm{E}[\varphi(y)]\, d\lambda(y)$, *for* $i = 1, \ldots, n+1$;

(b) $\mu_{X_{n+1}|X^{(n)}}(A \mid x^{(n)}) = \displaystyle\int_A \mathrm{E}[\varphi(y) \mid X^{(n)} = x^{(n)}]\, d\lambda(y)$, *almost surely* $[\mu_{X^{(n)}}]$.

Proof. By Definition 2.3, we have

$$\mathrm{E}[\varphi(y)] = \mathrm{E}\left[\sum_{i=1}^{k} H_i\, I_{[t_{i-1},t_i)}(y)\right] = \int_{\mathbb{R}^k} f(y)\, d\mu_H(h)\,,$$

where $h \in \mathbb{R}^k$ and $f(y) = \sum_{i=1}^k h_i\, I_{[t_{i-1},t_i)}(y)$, for $y \in \mathbb{R}$. In an analogous manner, we have

$$\mathrm{E}[\varphi(y) \mid X^{(n)} = x^{(n)}] = \int_{\mathbb{R}^k} f(y)\, d\mu_{H|X^{(n)}}(h \mid x^{(n)})\,.$$

For item (a), note that

$$\begin{aligned}
\mu_{X_i}(A) = P\{X_i \in A, H \in \mathbb{R}^k\} &= \int_{\mathbb{R}^k} \mu_{X_i|H}(A \mid h)\, d\mu_H(h) \\
&= \int_{\mathbb{R}^k} \left(\int_A f(y)\, d\lambda(y)\right) d\mu_H(h) = \int_A \left(\int_{\mathbb{R}^k} f(y)\, d\mu_H(h)\right) d\lambda(y) \\
&= \int_A \mathrm{E}[\varphi(y)]\, d\lambda(y)\,,
\end{aligned}$$

where the fourth equality follows from Tonelli's Theorem. For item (b), for each $B \in \mathscr{R}^n$, we have

$$P\{X_{n+1} \in A, X^{(n)} \in B\} = \int_B \mu_{X_{n+1}|X^{(n)}}(A \mid x^{(n)})\, d\mu_{X^{(n)}}(x^{(n)})\,.$$

On the other hand, we have

$$\begin{aligned}
P\{X_{n+1} \in A, X^{(n)} \in B\} &= P\{X_{n+1} \in A, X^{(n)} \in B, H \in \mathbb{R}^k\} \\
&= \int_{B\times\mathbb{R}^k} \mu_{X_{n+1}|X^{(n)},H}(A \mid x^{(n)}, h)\, d\mu_{X^{(n)},H}(x^{(n)}, h) \\
&= \int_{B\times\mathbb{R}^k} \mu_{X_{n+1}|H}(A \mid h)\, d\mu_{X^{(n)},H}(x^{(n)}, h) \\
&= \int_B \left(\int_{\mathbb{R}^k} \mu_{X_{n+1}|H}(A \mid h)\, d\mu_{H|X^{(n)}}(h \mid x^{(n)}) \right) d\mu_{X^{(n)}}(x^{(n)}) \\
&= \int_B \left(\int_{\mathbb{R}^k} \left(\int_A f(y)\, d\lambda(y) \right) d\mu_{H|X^{(n)}}(h \mid x^{(n)}) \right) d\mu_{X^{(n)}}(x^{(n)}) \\
&= \int_B \left(\int_A \left(\int_{\mathbb{R}^k} f(y)\, d\mu_{H|X^{(n)}}(h \mid x^{(n)}) \right) d\lambda(y) \right) d\mu_{X^{(n)}}(x^{(n)}) \\
&= \int_B \left(\int_A \mathrm{E}[\varphi(y) \mid X^{(n)} = x^{(n)}]\, d\lambda(y) \right) d\mu_{X^{(n)}}(x^{(n)}) \,,
\end{aligned}$$

where the third equality follows from the hypothesis of conditional independence and Theorem B.61 of Schervish (1995), the fourth equality is a consequence of Theorem 2.6.4 of Ash (2000), and the sixth equality is due to Tonelli's Theorem. Comparing both expressions for $P\{X_{n+1} \in A, X^{(n)} \in B\}$, we get the desired result. ∎

Proposition 7.6. *Let $\mu_K = P \circ K^{-1}$ over $(\mathbb{N}, 2^{\mathbb{N}})$ be the distribution of K and let $\mu_R = P \circ R^{-1}$ over $(\mathbb{R}, \mathscr{R})$be the distribution of R. Denote by $\mu_{K,R}$ the joint distribution of K and R, which by the independence of K and R is equal to the product measure $\mu_K \times \mu_R$, and let $\mu_{K,R,H}$ be the joint distribution of K, R and H. In the hierarchical model described on Section 5, we have that $\mu_{X|K,R}(\,\cdot \mid k, \rho) \ll \lambda_n$, almost surely $[\mu_{K,R}]$, with Radon-Nikodym derivative*

$$\frac{d\mu_{X|K,R}}{d\lambda_n}(x \mid k, \rho) = f_{X|K,R}(x \mid k, \rho) = \int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_{H|K,R}(h \mid k, \rho) \,,$$

for the $f_{X|H}$ defined on Lemma 3.1.

Proof. Let $A \in \mathscr{R}^n$ and $B \in 2^{\mathbb{N}} \otimes \mathscr{R}$. By the definition of conditional distribution, we have

$$P\{X \in A, (K, R) \in B\} = \int_B \mu_{X|K,R}(A \mid k, \rho)\, d\mu_{K,R}(k, \rho) \,.$$

On the other hand, by arguments similar to those used in the proof of Proposition 7.5,

we have

$$\begin{aligned}
&P\{X \in A,(K,R) \in B\} \\
&\qquad = P\{X \in A, (K,R) \in B, H \in \mathbb{R}^k\} \\
&\qquad = \int_{B\times\mathbb{R}^k} \mu_{X|K,R,H}(A \mid k,\rho,h)\, d\mu_{K,R,H}(k,\rho,h) \\
&\qquad = \int_{B\times\mathbb{R}^k} \mu_{X|H}(A \mid h)\, d\mu_{K,R,H}(k,\rho,h) \\
&\qquad = \int_B \left(\int_{\mathbb{R}^k} \mu_{X|H}(A \mid h)\, d\mu_{H|K,R}(h \mid k,\rho) \right) d\mu_{K,R}(k,\rho) \\
&\qquad = \int_B \left(\int_{\mathbb{R}^k} \left(\int_A f_{X|H}(x \mid h)\, d\lambda_n(x) \right) d\mu_{H|K,R}(h \mid k,\rho) \right) d\mu_{K,R}(k,\rho) \\
&\qquad = \int_B \left(\int_A \left(\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_{H|K,R}(h \mid k,\rho) \right) d\lambda_n(x) \right) d\mu_{K,R}(k,\rho)\,.
\end{aligned}$$

Comparing both expressions for $P\{X \in A, (K,R) \in B\}$, we have

$$\mu_{X|K,R}(A \mid k,\rho) = \int_A \left(\int_{\mathbb{R}^k} f_{X|H}(x \mid h)\, d\mu_{H|K,R}(h \mid k,\rho) \right) d\lambda_n(x)\,,$$

almost surely $[\mu_{K,R}]$, and the result follows. ∎

Proof of Proposition 6.1. By Tonelli's Theorem, the expected loss is

$$\mathrm{E}[L(\varphi,f)] = \int_a^b f^2(x)\, d\lambda(x) - 2\int_a^b f(x)\mathrm{E}[\varphi(x)]\, d\lambda(x) + C_0\,,$$

where we have defined the positive constant $C_0 = \int_a^b \mathrm{E}[\varphi^2(x)]\, d\lambda(x)$. By hypothesis, each f has the form $f(x) = \sum_{i=1}^N \alpha_i\, g_i(x)$, leading us to

$$\mathrm{E}[L(\varphi,f)] = \sum_{i,j=1}^N \left(\alpha_i\alpha_j \int_a^b g_i(x)g_j(x)\, d\lambda(x) \right) - 2\sum_{i=1}^N \left(\alpha_i \int_a^b g_i(x)\mathrm{E}[\varphi(x)]\, d\lambda(x) \right) + C_0\,,$$

where we have used the linearity of the integral. Therefore, minimizing the expected loss is the same as solving the problem of constrained minimization of the quadratic form Q. For the matrix $M = (M_{ij})$, note that, for every nonnull $y = (y_1,\ldots,y_N)^\top \in \mathbb{R}^N$, we have

$$\begin{aligned}
y^\top My = \sum_{i,j=1}^N y_i y_j M_{ij} &= \sum_{i,j=1}^N \left(y_i y_j \int_a^b g_i(x)g_j(x)\, d\lambda(x) \right) \\
&= \int_a^b \sum_{i,j=1}^N \left(y_i\, g_i(x)\, y_j\, g_j(x) \right)\, d\lambda(x) = \int_a^b \left(\sum_{i=1}^N y_i\, g_i(x) \right)^2\, d\lambda(x) > 0\,,
\end{aligned}$$

where we have used the linearity of the integral. Therefore, the matrix M is positive definite, yielding (see Bazaraa and Shetty (2006)) that the quadratic form Q is convex and the problem of constrained minimization of Q has a single global solution $(\hat{\alpha_1}, \ldots, \hat{\alpha_N})$. Since the Bayes decision is the f that minimizes the expected loss, the result follows. ∎

References

Ash, R. B. (2000). *Probability and Measure Theory*. Massachusetts: Harcourt/Academic Press, 3rd edition.

Bazaraa, M. S. and Shetty, C. M. (2006). *Nonlinear Programming: Theory and Algorithms*. New Jersey: Wiley-Interscience, 3rd edition.

Billingsley, P. (1995). *Probability and Measure*. New Jersey: Wiley-Interscience, 3rd edition.

Blackwell, D. (1973). "Discreteness of Ferguson Selections." *The Annals of Statistics*, 1(2): 356–358.

Ferguson, T. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1(2): 209–230.

Gosh, J. K. and Ramamoorthi, R. V. (2002). *Bayesian Nonparametrics*. New York: Springer.

Lenk, P. J. (1988). "The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities." *Journal of the American Statistical Association*, 83(402): 509–516.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer, 2nd edition.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.

Thorburn, D. (1986). "A Bayesian Approach to Density Estimation." *Biometrika*, 73(1): 65–75.