Posterior convergence rate of a class of Dirichlet process mixture model for compositional data

ANDRÉS F. BARRIENTOS, ALEJANDRO JARA AND CLAUDIA WEHRHAHN

September 6, 2016

Abstract

We propose a Dirichlet process mixture of mixtures of Dirichlet models for density estimation. By assuming random sampling from a density belonging to a Hölder class, we show that the posterior distribution of the model is rate-optimal.

Keywords: Simplex, Bayesian nonparametrics, Frequentist asymptotics.

¹Andrés F. Barrientos is Research Fellow, Department of Statistical Science, Duke University, USA, Box 90251 Durham NC 27708-0251 (E-mail: afb26@stat.duke.edu). Alejandro Jara is Associate Professor, Department of Statistics, Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, Chile (E-mail: atjara@uc.cl). Claudia Wehrhahn is Ph.D. student, Department of Statistics, Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, Chile (E-mail: cawehrha@mat.puc.cl). A. Jara's research is supported by Fondecyt grant 1141193. A. F. Barrientos's research is supported by NSF ACI 1443014 grant.

1 Introduction

Models based on convex combinations of densities from parametric families underly mainstream approaches to density estimation (Silverman, 1986; Lindsay, 1995; Lo, 1984). Under a Bayesian nonparametric approach, a mixture model provides a convenient set up in that a prior distribution on densities is induced by placing a prior distribution on the mixing measure (see, e.g. Ferguson, 1983; Lo, 1984; Escobar & West, 1995). Extending the work by Petrone (1999a,b) on the unit interval, based on Bernstein polynomials, Barrientos et al. (2015) suggested to construct a probability model for densities on the d -dimensional simplex $\Delta_d = \{(y_1, \dots, y_d) \in [0, 1]^d : \sum_{i=1}^d y_i \leq 1\}$, by using mixtures of Dirichlet-densities of the form

$$b_{k,G}(\mathbf{y}) = \sum_{\mathbf{j} \in \mathcal{H}_{k,d}^0} G \left(\left(\frac{j_1-1}{k}, \frac{j_1}{k} \right] \times \dots \times \left(\frac{j_d-1}{k}, \frac{j_d}{k} \right] \right) \text{dir}(\mathbf{y} \mid \eta(k, \mathbf{j})), \quad (1)$$

where k is a discrete random variable supported on $\mathbb{N}$, G is a random probability measure defined on $\Delta_d^0 = \{\mathbf{y} \in \Delta_d : y_j > 0, \ j = 1, \dots, d\}$, $\mathbf{j} = (j_1, \dots, j_d)$,

$$\mathcal{H}_{k,d}^0 = \left\{ (j_1, \dots, j_d) \in \{1, \dots, k+d-1\}^d : \sum_{l=1}^d j_l \leq k+d-1 \right\},$$

$\text{dir}(\cdot \mid (\alpha_1, \dots, \alpha_{d+1}))$ denotes the density function of a d -dimensional Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_{d+1})$,

$$\text{dir}(\mathbf{y} \mid (\alpha_1, \dots, \alpha_{d+1})) = \frac{\Gamma(\sum_{l=1}^{d+1} \alpha_l)}{\prod_{l=1}^{d+1} \Gamma(\alpha_l)} \left(\prod_{l=1}^d y_l^{\alpha_l-1} \right) \left(1 - \sum_{l=1}^d y_l \right)^{\alpha_{d+1}-1},$$

and $\eta(k, \mathbf{j}) = (\mathbf{j}, k+d-\sum_{l=1}^d j_l)$.

By assuming that G is a stick-breaking process, Barrientos et al. (2015) showed that the induced prior distribution for densities has full weak, L_∞ and Kullback-Leibler support. By

employing results provided by Ghosal et al. (1999), Barrientos et al. (2015) showed that the posterior distribution is L_1 -consistent, under mild conditions on the distribution of k . Furthermore, they showed that if the “true” density belongs to the proposed class of mixture models, then the speed of convergence of the posterior distribution is at most $\log n / \sqrt{n}$. They also showed that when the “true” density belongs to a Hölder class with α -regularity, $\alpha \in (0, 2]$, then the speed of convergence of the posterior distribution is at most $(\log n)^{(2\alpha+d)/(2\alpha+2d)} / n^{\alpha/(2\alpha+2d)}$, which is suboptimal. A function p on Δ_d belongs to a Hölder class with α -regularity if, for $\alpha \in (0, 1]$, there exists a constant $L_\alpha(p)$ such that, for every pair of points $\mathbf{y}_1, \mathbf{y}_2 \in \Delta_d$, $|p(\mathbf{y}_1) - p(\mathbf{y}_2)| \leq L_\alpha(p) \|\mathbf{y}_1 - \mathbf{y}_2\|_2^\alpha$, where $\|\cdot\|_2$ denotes the l_2 -norm, or if for $\alpha \in (1, 2]$ its first order partial derivatives belong to a Hölder class with $(\alpha - 1)$ -regularity.

Kruijer & Van der Vaart (2008) proposed a modification of prior models to obtain minimax posterior rates in the case of density estimation on $[0, 1]$ and $[0, 1]^2$. Following a similar approach to the one suggested by Kruijer & Van der Vaart (2008), we proposed a modification of the mixture model given by expression (1), for which we show that the posterior convergence rate is minimax optimal up to a logarithmic factor. Specifically, by assuming a Dirichlet process mixing, we show that if the “true” density belongs to a Hölder class with α -regularity, of at most $\alpha = 1$, then the convergence rate of the posterior distribution is upper bounded by $(\log n)^{(4\alpha+d)/(4\alpha+2d)} / n^{\alpha/(2\alpha+d)}$.

2 The Dirichlet process mixture model and properties

The mixture model of Kruijer & Van der Vaart (2008), for data supported on $[0, 1]$, arises by “coarsening” the mixture of beta densities model proposed by Petrone (1999a,b), which is based on Bernstein polynomials. Specifically, the model proposed by Kruijer & Van der Vaart (2008) arises by (i) defining groups of beta density basis functions from the original Bernstein polynomial model, (ii) defining a mixture of beta densities with equal weights in each group, and (iii) mixing the resulting mixtures of beta densities with weights defined by the mass assigned by the random mixing probability distribution G to the union of intervals associated with the beta density basis functions in each group. Following similar ideas, we propose a modification to the model proposed by Barrientos et al. (2015), given by expression (1), leading to the following mixture of mixtures of Dirichlet densities,

$$\begin{aligned}\tilde{b}_{k,G}(\mathbf{y}) &= \sum_{\mathbf{j} \in \mathcal{H}_{\sqrt{k},d}^0} G(B_{\mathbf{j},k}) \left\{ \sum_{\mathbf{l} \in \mathcal{L}_{k,d,\mathbf{j}}} k^{-d/2} \text{dir}(\mathbf{y} \mid \tilde{\eta}(k, \mathbf{l})) \right\}, \\ &= \sum_{\mathbf{j} \in \tilde{\mathcal{H}}_{k,d}} G(A_{\mathbf{j},k}) k^{-d/2} \text{dir}(\mathbf{y} \mid \tilde{\eta}(k, \mathbf{j})),\end{aligned}\tag{2}$$

where k is a discrete random variable supported on $\mathbb{K} = \{l \in \mathbb{N} : l^{1/2} \text{ is an integer}\}$,

$$\mathcal{L}_{k,d,\mathbf{j}} = \left\{ (l_1, \dots, l_d) \in \{1, \dots, k+d-1\}^d : l_i \in \{(j_i-1)\sqrt{k}+1, \dots, j_i\sqrt{k}\}, i=1, \dots, d \right\},$$

$$\tilde{\mathcal{H}}_{k,d} = \bigcup_{\mathbf{j} \in \mathcal{H}_{\sqrt{k},d}^0} \mathcal{L}_{k,d,\mathbf{j}} = \mathcal{H}_{k,d}^0 \bigcup \mathcal{H}_{k,d}^1,$$

$$\mathcal{H}_{k,d}^1 = \left\{ (j_1, \dots, j_d) \in \tilde{\mathcal{H}}_{k,d} : \sum_{l=1}^d j_l > k+d-1 \right\},$$

$$B_{\mathbf{j},k} = \left(\frac{j_1 - 1}{\sqrt{k}}, \frac{j_1}{\sqrt{k}} \right] \times \dots \times \left(\frac{j_d - 1}{\sqrt{k}}, \frac{j_d}{\sqrt{k}} \right],$$

$$A_{\mathbf{j},k} = \left(\frac{\mathcal{T}_k(j_1) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_1)}{\sqrt{k}} \right] \times \dots \times \left(\frac{\mathcal{T}_k(j_d) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_d)}{\sqrt{k}} \right],$$

with $\mathcal{T}_k(j) = \lceil j/\sqrt{k} \rceil$, where $\lceil \cdot \rceil$ denoting the ceiling function,

$$\tilde{\eta}(k, \mathbf{j}) = \begin{cases} \eta(k, \mathbf{j}), & \text{if } \mathbf{j} \in \mathcal{H}_{k,d}^0 \\ \mathbf{1}_{d+1}, & \text{if } \mathbf{j} \in \mathcal{H}_{k,d}^1 \end{cases},$$

and $\mathbf{1}_d$ is a d -dimensional vector of ones.

Note that $\tilde{b}_{k,G}(\cdot)$ is a polynomial function of $\mathbf{y}$ that belongs to a class that can approximate any element in the set of all probability measures defined on Δ_d with α -Hölder continuous density function, w.r.t. Lebesgue measure, and $\alpha \in (0, 1]$. The following theorem provides the order of approximation of the density function of G (w.r.t. Lebesgue measure) by $\tilde{b}_{k,G}(\mathbf{y})$, as k increases.

Theorem 1. *Let P be an absolutely continuous probability measure defined on Δ_d , w.r.t. Lebesgue measure, with density function p . If p is an α -Hölder continuous function, with $\alpha \in (0, 1]$, then*

$$\|\tilde{b}_{k,P} - p\|_\infty = \mathcal{O}(k^{-\alpha/2}).$$

Now, suppose we have independent observations $\mathbf{y}_1, \dots, \mathbf{y}_n$ from a common density f supported on Δ_d . We model the common density f using expression (2) (i.e., by assuming $f(\mathbf{y}) = \tilde{b}_{k,G}(\mathbf{y})$), and by placing a Dirichlet process prior on G and an appropriate prior distribution on k . The resulting model can be alternatively represented by the following Dirichlet

process mixture model,

$$\mathbf{y}_i \mid k, \boldsymbol{\theta}_i \stackrel{i.i.d.}{\sim} f_{k, \boldsymbol{\theta}_i}(\cdot) \equiv \sum_{j_1=(\lceil \sqrt{k} \boldsymbol{\theta}_{(i,1)} \rceil - 1)\sqrt{k}+1}^{\lceil \sqrt{k} \boldsymbol{\theta}_{(i,1)} \rceil \sqrt{k}} \dots \sum_{j_d=(\lceil \sqrt{k} \boldsymbol{\theta}_{(i,d)} \rceil - 1)\sqrt{k}+1}^{\lceil \sqrt{k} \boldsymbol{\theta}_{(i,d)} \rceil \sqrt{k}} \mathbb{I}_{\tilde{\mathcal{H}}_{k,d}}(\mathbf{j}) k^{-d/2} \text{dir}(\cdot \mid \tilde{\eta}(k, \mathbf{j})), \quad (3)$$

$$\boldsymbol{\theta}_i \mid G \stackrel{i.i.d.}{\sim} G, \quad (4)$$

$$G \mid M, G_0 \sim DP(MG_0), \quad (5)$$

$$k \sim \rho, \quad (6)$$

where $G \sim DP(MG_0)$ refers to G being a Dirichlet process with baseline distribution G_0 and total mass parameter M , M is a nonnegative constant, G_0 is a probability measure defined on Δ_d^0 , and ρ is a probability measure defined on $\mathbb{K}$.

The result of Theorem 1 and the weak support properties of the DP allow us to show that the posterior distribution induced by $\tilde{b}_{k,G}$ is rate-optimal. Let $\mathcal{P}(\Delta_d)$ be the set of all absolutely continuous probability measures defined on Δ_d , w.r.t. Lebesgue measure, and with continuous density. The following theorem provides sufficient conditions on the prior specification for the posterior distribution induced by expressions (3) – (6) to be rate-optimal, under certain assumptions on the “true” data generating probability measure F

Theorem 2. *Assume that $\mathbf{y}_1, \dots, \mathbf{y}_n \stackrel{i.i.d.}{\sim} F$, where $F \in \mathcal{P}(\Delta_d)$ and has a strictly positive density f which is α -Hölder continuous for $\alpha \in (0, 1]$. Assume that G_0 is a probability measure defined on Δ_d^0 with strictly positive density w.r.t. Lebesgue measure. Assume that ρ is specified such that for every $\bar{k} \in \mathbb{K}$, $B_1 \exp(-\beta_1 \bar{k}^{d/2}) \leq \rho(\bar{k}) \leq B_2 \exp(-\beta_2 \bar{k}^{d/2})$, for some positive*

constants B_1, B_2, β_1 and β_2 . Then, for a sufficiently large constant C ,

$$\Pi \left\{ P \in \mathcal{P}(\Delta_d) : H(p, f) \geq C \frac{(\log n)^{(4\alpha+d)/(4\alpha+2d)}}{n^{\alpha/(2\alpha+d)}} \mid \mathbf{y}_1, \dots, \mathbf{y}_n \right\} \xrightarrow[n \rightarrow \infty]{} 0,$$

in F^n -probability, where p is the density of P , $H(p, f)$ is the Hellinger distance between P and F and $\Pi\{\cdot \mid \mathbf{y}_1, \dots, \mathbf{y}_n\}$ is the posterior distribution induced by $\tilde{b}_{k,G}$ and expressions (5) and (6).

3 Proofs

3.1 Proof of Theorem 1

By the triangle inequality it follows that, for every probability measure P defined on Δ_d and with α -Hölder continuous density p , $\|\tilde{b}_{k,P} - p\|_\infty \leq \|b_{k,P} - p\|_\infty + \|b_{k,P} - \tilde{b}_{k,P}\|_\infty$, where $b_{k,P}$ is defined in expression (1). Barrientos et al. (2015) showed that $\|b_{k,P} - p\|_\infty = \mathcal{O}(k^{-\alpha/2})$. Thus, to prove the theorem it is sufficient to show that $\|b_{k,P} - \tilde{b}_{k,P}\|_\infty = \mathcal{O}(k^{-\alpha/2})$. This result is analogous to the one obtained by Kruijer & Van der Vaart (2008) in the proof of Lemma 3. The rest of this proof follows similar arguments to the ones employed by Barrientos et al. (2015).

Let $\mathbf{z}_j = (z_{j,1}, \dots, z_{j,d+1}) \stackrel{i.i.d.}{\sim} \text{Multinomial}(1, \boldsymbol{\pi})$, $1 \leq j \leq k-1$, where $\boldsymbol{\pi} = (\mathbf{y}, 1 - \sum_{l=1}^d y_l)$. Set $\mathbf{S}_k = \sum_{j=1}^{k-1} \mathbf{z}_j$ and $\mathbf{W}_k = (S_{k,1}, \dots, S_{k,d})$. By the multivariate version of the mean value theorem proposed by Ash (2008), it is possible to ensure the existence of $(0, 1)^d$ -valued random vectors $\mathbf{c}_{\mathbf{W}_k}$ and $\tilde{\mathbf{c}}_{\mathbf{W}_k}$ such that,

$$P \left(\left(\frac{\mathcal{T}_k(S_{k,1} + 1) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(S_{k,1} + 1)}{\sqrt{k}} \right) \times \dots \times \left(\frac{\mathcal{T}_k(S_{k,d} + 1) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(S_{k,d} + 1)}{\sqrt{k}} \right) \right) =$$

$$\frac{1}{k^{d/2}P} \left(\frac{\mathcal{T}_k(\mathbf{W}_k + \mathbf{1}_d) - \mathbf{1}_d + \tilde{\mathbf{c}}_{\mathbf{W}_k}}{\sqrt{k}} \right),$$

and

$$P \left(\left(\frac{S_{k,1}}{k}, \frac{S_{k,1}+1}{k} \right] \times \dots \times \left(\frac{S_{k,d}}{k}, \frac{S_{k,d}+1}{k} \right] \right) = k^{-d} p \left(\frac{\mathbf{W}_k + \mathbf{c}_{\mathbf{W}_k}}{k} \right).$$

Also, recall that $\tilde{b}_{k,P}(\mathbf{y})$ can be expressed as the following mixture of Dirichlet densities

$$\begin{aligned} \tilde{b}_{k,P}(\mathbf{y}) &= \sum_{\mathbf{j} \in \mathcal{H}_{k,d}^0} k^{-d/2} P \left(\left(\frac{\mathcal{T}_k(j_1) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_1)}{\sqrt{k}} \right] \times \dots \times \left(\frac{\mathcal{T}_k(j_d) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_d)}{\sqrt{k}} \right] \right) \text{dir}(\mathbf{y} \mid \eta(k, \mathbf{j})) \\ &\quad + \sum_{\mathbf{j} \in \mathcal{H}_{k,d}^1} k^{-d/2} P \left(\left(\frac{\mathcal{T}_k(j_1) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_1)}{\sqrt{k}} \right] \times \dots \times \left(\frac{\mathcal{T}_k(j_d) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_d)}{\sqrt{k}} \right] \right) \text{dir}(\mathbf{y} \mid \mathbf{1}_{d+1}). \end{aligned}$$

Now, given the relationship between the Dirichlet density and the multinomial distribution, it

follows that, for every $\mathbf{j} \in \mathcal{H}_{k,d}^0$,

$$\begin{aligned} &\text{dir}(\mathbf{y} \mid \eta(k, \mathbf{j})) \\ &= \frac{(k+d-1)!}{(j_1-1)! \dots (j_d-1)! (k+d-\sum_{l=1}^d j_l-1)!} y_1^{j_1-1} \dots y_d^{j_d-1} \left(1 - \sum_{l=1}^d y_l \right)^{k+d-\sum_{l=1}^d j_l-1}, \\ &= \frac{(k+d-1)!}{(k-1)!} \frac{(k-1)!}{t_1! \dots t_d! (k-\sum_{l=1}^d t_l-1)!} y_1^{t_1} \dots y_d^{t_d} \left(1 - \sum_{l=1}^d y_l \right)^{k-\sum_{l=1}^d t_l-1}, \\ &= \frac{(k+d-1)!}{(k-1)!} \text{Mult}(t_1, \dots, t_d \mid k-1, \mathbf{y}), \end{aligned}$$

where $t_l = j_l - 1$, $l = 1, \dots, d$, $\text{Mult}(\cdot \mid k-1, \mathbf{y})$ denotes the probability mass function of a

Multinomial distribution with parameters $k-1$ and $\mathbf{y}$. Thus, one part of $\tilde{b}_{k,P}$ can be expressed

as a expected value of a multinomial distribution, i.e.,

$$\begin{aligned} \tilde{b}_{k,P}(\mathbf{y}) &= \frac{(k+d-1)!}{(k-1)!} E_{\mathbf{S}_k} \left[k^{-d/2} P \left(\left(\frac{\mathcal{T}_k(S_{k,1}+1)-1}{\sqrt{k}}, \frac{\mathcal{T}_k(S_{k,1}+1)}{\sqrt{k}} \right) \times \dots \times \right. \right. \\ &\quad \left. \left. \left(\frac{\mathcal{T}_k(S_{k,d}+1)-1}{\sqrt{k}}, \frac{\mathcal{T}_k(S_{k,d}+1)}{\sqrt{k}} \right) \right) \right] \end{aligned} \quad (7)$$

$$\begin{aligned} &+ \sum_{\mathbf{j} \in \mathcal{H}_{k,d}^1} k^{-d/2} P \left(\left(\frac{\mathcal{T}_k(j_1)-1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_1)}{\sqrt{k}} \right) \times \dots \times \left(\frac{\mathcal{T}_k(j_d)-1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_d)}{\sqrt{k}} \right) \right) \text{dir}(\mathbf{y} \mid \mathbf{1}_{d+1}), \\ &= \frac{(k+d-1)!}{(k-1)!k^d} E_{\mathbf{S}_k} \left[p \left(\frac{\mathcal{T}_k(\mathbf{W}_k + \mathbf{1}_d) - \mathbf{1}_d + \tilde{\mathbf{c}}_{\mathbf{W}_k}}{\sqrt{k}} \right) \right] \\ &+ \sum_{\mathbf{j} \in \mathcal{H}_{k,d}^1} k^{-d} p \left(\frac{\mathcal{T}_k(\mathbf{j}) + \tilde{\mathbf{c}}_{\mathbf{j}}}{\sqrt{k}} \right) \text{dir}(\mathbf{y} \mid \mathbf{1}_{d+1}), \end{aligned} \quad (8)$$

where $\tilde{\mathbf{c}}_{\mathbf{j}}$ is determined by the multivariate mean value theorem. From expression (7) and the

fact that $b_{k,P}(\mathbf{y}) = \frac{(k+d-1)!}{(k-1)!k^d} E_{\mathbf{S}_k} \left[p \left(\frac{\mathbf{W}_k + \mathbf{c}_{\mathbf{W}_k}}{k} \right) \right]$ (see, Barrientos et al., 2015, Section 6.1), it

follows that

$$\begin{aligned} &\|\tilde{b}_{k,P} - b_{k,P}\|_{\infty} \\ &\leq \frac{(k+d-1)!}{(k-1)!k^d} \sup_{\mathbf{y} \in \Delta_d} \left| E_{\mathbf{S}_k} \left[p \left(\frac{\mathcal{T}_k(\mathbf{W}_k + \mathbf{1}_d) - \mathbf{1}_d + \tilde{\mathbf{c}}_{\mathbf{W}_k}}{\sqrt{k}} \right) \right] - p \left(\frac{\mathbf{W}_k + \mathbf{c}_{\mathbf{W}_k}}{k} \right) \right| \\ &\quad + k^{-d} \sup_{\mathbf{y} \in \Delta_d} \left| \sum_{\mathbf{j} \in \mathcal{H}_{k,d}^1} p \left(\frac{\mathcal{T}_k(\mathbf{j}) + \tilde{\mathbf{c}}_{\mathbf{j}}}{\sqrt{k}} \right) \text{dir}(\mathbf{y} \mid \mathbf{1}_{d+1}) \right|, \\ &\leq \frac{(k+d-1)!}{(k-1)!k^d} L_{\alpha}(p) \sup_{\mathbf{y} \in \Delta_d} E_{\mathbf{S}_k} \left[\left\| \frac{\mathcal{T}_k(\mathbf{W}_k + \mathbf{1}_d) - \mathbf{1}_d + \tilde{\mathbf{c}}_{\mathbf{W}_k}}{\sqrt{k}} - \frac{\mathbf{W}_k + \mathbf{c}_{\mathbf{W}_k}}{k} \right\|^{\alpha} \right] \\ &\quad + \|p\|_{\infty} \|\text{dir}(\cdot \mid \mathbf{1}_{d+1})\|_{\infty} k^{-d} \text{Cardinality}(\mathcal{H}_{k,d}^1), \\ &\leq \frac{(k+d-1)!}{(k-1)!k^d} L_{\alpha}(p) \sup_{\mathbf{y} \in \Delta_d} E_{\mathbf{S}_k} \left[\sum_{j=1}^d \left(\frac{\mathcal{T}_k(S_{k,j}+1)-1 + \tilde{c}_{S_{k,j}}}{\sqrt{k}} - \frac{S_{k,j} + c_{S_{k,j}}}{k} \right)^2 \right]^{\alpha/2} \\ &\quad + \|p\|_{\infty} \|\text{dir}(\cdot \mid \mathbf{1}_{d+1})\|_{\infty} k^{-d} \text{Cardinality}(\mathcal{H}_{k,d}^1), \end{aligned}$$

where $L_{\alpha}(p)$ is the Hölder constant for the function p . The proof of the theorem is completed

by noticing that

$$\begin{aligned} \left(\frac{\mathcal{T}_k(S_{k,j} + 1) - 1 + \tilde{c}_{S_{k,j}}}{\sqrt{k}} - \frac{S_{k,j} + c_{S_{k,j}}}{k} \right)^2 &\leq \left(\frac{3}{\sqrt{k}} \right)^2 + 2 \left(\frac{3}{\sqrt{k}} \right) \left(\frac{1}{\sqrt{k}} + \frac{1}{k} \right), \\ &\quad + \left(\frac{1}{\sqrt{k}} + \frac{1}{k} \right)^2, \\ &\leq \frac{25}{k}, \end{aligned}$$

$$\text{Cardinality}(\mathcal{H}_{k,d}^1) < k^{d/2} \sum_{l=1}^{d-1} \left[\prod_{j=1}^{d-1} \frac{(\sqrt{k} + j - l)}{(d-1)!} \right],$$

and that for k large enough, $(k + d - 1)! / ((k - 1)! k^d) \leq 2$.

3.2 Proof of Theorem 2

To prove this theorem it is sufficient to show that all the conditions of Theorem 2.1 in Ghosal (2001) are satisfied. In our context, the conditions of Theorem 2.1 are satisfied if there is a sequence $(\mathcal{F}_n)_{n \geq 1}$ of subsets of the parameter space and there are positive sequences $(\tilde{\epsilon}_n)_{n \geq 1}$ and $(\bar{\epsilon}_n)_{n \geq 1}$ satisfying $\tilde{\epsilon}_n, \bar{\epsilon}_n \rightarrow 0$ and $n \min\{\tilde{\epsilon}_n^2, \bar{\epsilon}_n^2\} \rightarrow \infty$, such that for certain positive constants a_1, a_2, a_3, a_4 , the following conditions hold:

$$\Pi \{P \in \mathcal{P}(\Delta_d) : P \in N(\tilde{\epsilon}_n, f)\} \geq a_1 \exp(-a_2 n \tilde{\epsilon}_n^2), \quad (9)$$

$$\Pi \{P \in \mathcal{P}(\Delta_d) : P \in \mathcal{F}_n^C\} \leq a_3 \exp(-(a_2 + 4)n \tilde{\epsilon}_n^2), \quad (10)$$

$$\log(D(\bar{\epsilon}_n, \mathcal{F}_n, H)) \leq a_4 n \bar{\epsilon}_n^2, \quad (11)$$

where Π is the prior distribution induced by $\tilde{b}_{k,G}$ and expressions (5) – (6),

$$N(\epsilon, f) = \left\{ Q \in \mathcal{P}(\Delta_d) : \int_{\Delta_d} f(\mathbf{y}) \log \left(\frac{f(\mathbf{y})}{q(\mathbf{y})} \right) d\mathbf{y} \leq \epsilon, \int_{\Delta_d} f(\mathbf{y}) \left(\log \frac{f(\mathbf{y})}{q(\mathbf{y})} \right)^2 d\mathbf{y} \leq \epsilon \right\},$$

q is the density function of the probability measure Q , $\mathcal{F}_n^C$ is the complement of the set $\mathcal{F}_n$, and $D(\epsilon, \mathcal{F}, H)$ is the ϵ -packeting number defined to be the maximum number of points in $\mathcal{F}$ such that the H -distance between each pair is at least ϵ .

Let $\tilde{\epsilon}_n = (\log n/n)^{\alpha/(2\alpha+d)}$, $k_n = \left\lceil M_1/\tilde{\epsilon}_n^{1/\alpha} \right\rceil - 1$, for some $M_1 > 1$, $\boldsymbol{\pi}_{k_n^2}^P = \left(\pi_{k_n^2, \mathbf{j}}^P \right)_{\mathbf{j} \in \tilde{\mathcal{H}}_{k_n^2, d}}$,

and

$$\tilde{b}_{k_n^2, P}(\cdot) = \sum_{\mathbf{j} \in \tilde{\mathcal{H}}_{k_n^2, d}} \pi_{k_n^2, \mathbf{j}}^P \text{dir}(\cdot \mid \tilde{\eta}(k_n^2, \mathbf{j})),$$

where

$$\pi_{k_n^2, \mathbf{j}}^P = k^{-d/2} P \left(\left(\frac{\mathcal{T}_{k_n^2}(j_1) - 1}{k_n}, \frac{\mathcal{T}_{k_n^2}(j_1)}{k_n} \right] \times \dots \times \left(\frac{\mathcal{T}_{k_n^2}(j_d) - 1}{k_n}, \frac{\mathcal{T}_{k_n^2}(j_d)}{k_n} \right] \right).$$

By Theorem 1, there exists a positive constant, C_1 , such that $\|f - \tilde{b}_{k_n^2, F}\|_\infty \leq C_1/k_n^\alpha$. Therefore,

if P is such that $\left\| \boldsymbol{\pi}_{k_n^2}^P - \boldsymbol{\pi}_{k_n^2}^F \right\|_1 \leq \tilde{\epsilon}_n^{1+2d/\alpha}$, then

$$\begin{aligned} \|\tilde{b}_{k_n^2, F} - \tilde{b}_{k_n^2, P}\|_\infty &\leq \left\| \sum_{\mathbf{j} \in \tilde{\mathcal{H}}_{k_n^2, d}} \pi_{k_n^2, \mathbf{j}}^F \text{dir}(\cdot \mid \tilde{\eta}(k_n^2, \mathbf{j})) - \sum_{\mathbf{j} \in \tilde{\mathcal{H}}_{k_n^2, d}} \pi_{k_n^2, \mathbf{j}}^P \text{dir}(\cdot \mid \tilde{\eta}(k_n^2, \mathbf{j})) \right\|_\infty \\ &\leq \|\text{dir}(\cdot \mid \tilde{\eta}(k_n^2, \mathbf{j}))\|_\infty \left\| \boldsymbol{\pi}_{k_n^2}^P - \boldsymbol{\pi}_{k_n^2}^F \right\|_1, \\ &\leq k_n^{2d} \tilde{\epsilon}_n^{1+2d/\alpha}. \end{aligned}$$

and, for some $C_2 > 0$,

$$\|f - \tilde{b}_{k_n^2, P}\|_\infty \leq \|f - \tilde{b}_{k_n^2, F}\|_\infty + \|\tilde{b}_{k_n^2, F} - \tilde{b}_{k_n^2, P}\|_\infty \leq \frac{C_1}{k_n^\alpha} + k_n^{2d} \tilde{\epsilon}_n^{1+2d/\alpha} \leq C_2 \tilde{\epsilon}_n.$$

This implies that $H^2(f, \tilde{b}_{k_n^2, P}) \leq \|1/f\|_\infty C_2^2 \tilde{\epsilon}_n^2$ (Please see, Kruijer & Van der Vaart, 2008, Lemma 5). Thus, this results along with expression (8.6) in Ghosal et al. (2000) and the fact that $\tilde{b}_{k_n^2, P}$ is bounded away from zero for k_n sufficiently large, imply that for some $C_3 > 0$,

$\tilde{b}_{k_n, P}^2 \in N(C_3 \tilde{\epsilon}_n, f)$. Therefore, by Lemma A.2 in Ghosal (2001),

$$\begin{aligned} \Pi \{P \in \mathcal{P}(\Delta_d) : P \in N(C_3 \tilde{\epsilon}_n, f)\} &\geq B_1 \exp(-\beta_1 k_n^d) C_4 \exp(-c_1 k_n^d \log(2/\tilde{\epsilon}_n^{1+2d/\alpha})), \\ &\geq B_1 C_4 \exp(-c_2 k_n^d \log n), \\ &\geq B_1 C_4 \exp(-c_3 n \tilde{\epsilon}_n^2), \end{aligned}$$

where C_4, c_1, c_2 and c_3 , are suitable positive constants. Thus, the condition (9) is satisfied. Now,

let $\{s_n\}_{n \geq 1}$ be a sequence such that

$$E_1 n^{1/(d+2\alpha)} (\log n)^{2\alpha/(d(d+2\alpha))} \leq s_n \leq E_2 n^{1/(d+2\alpha)} (\log n)^{2\alpha/(d(d+2\alpha))},$$

where E_1 and E_2 are positive constants, and $\mathcal{F}_n := \bigcup_{r=1}^{s_n} \mathcal{B}_r$, where,

$$\mathcal{B}_r = \left\{ \tilde{B}_{\bar{k}, P} : \bar{k} = r^2, P \text{ is a probability measure defined on } \Delta_d \right\},$$

and $\tilde{B}_{\bar{k}, P}$ denotes the probability measure associated with $\tilde{b}_{\bar{k}, P}$. Condition (10) is satisfied since,

for some E_1 large enough,

$$\begin{aligned} \Pi \{P \in \mathcal{P}(\Delta_d) : P \in \mathcal{F}_n^C\} &\leq \sum_{r=s_n+1}^{\infty} \rho(r^2), \\ &\leq B_2 \exp(-\beta_2 E_1 n^{d/(2\alpha+d)} (\log n)^{2\alpha/(2\alpha+d)}), \\ &\leq B_2 \exp(-(c_3 + 4)n \tilde{\epsilon}_n^2). \end{aligned}$$

Finally, set $\bar{\epsilon}_n = (\log n)^{(4\alpha+d)/(4\alpha+2d)} / n^{\alpha/(2\alpha+d)}$ and notice that

$$\begin{aligned} \|\tilde{b}_{r^2, F} - \tilde{b}_{r^2, P}\|_{\infty} &\leq \left\| \sum_{\mathbf{j} \in \tilde{\mathcal{H}}_{k_n^2, d}} \pi_{r^2, \mathbf{j}}^F \text{dir}(\cdot \mid \tilde{\eta}(r^2, \mathbf{j})) - \sum_{\mathbf{j} \in \tilde{\mathcal{H}}_{k_n^2, d}} \pi_{r^2, \mathbf{j}}^P \text{dir}(\cdot \mid \tilde{\eta}(r^2, \mathbf{j})) \right\|_{\infty}, \\ &\leq \|\text{dir}(\cdot \mid \tilde{\eta}(r^2, \mathbf{j}))\|_{\infty} \sum_{\mathbf{j} \in \tilde{\mathcal{H}}_{k_n^2, d}} |\pi_{r^2, \mathbf{j}}^F - \pi_{r^2, \mathbf{j}}^P|, \\ &= \|\text{dir}(\cdot \mid \tilde{\eta}(r^2, \mathbf{j}))\|_{\infty} \sum_{\mathbf{j} \in \mathcal{H}_{r, d}^0} |w_{r, \mathbf{j}}^F - w_{r, \mathbf{j}}^P|, \\ &\leq r^{2d} \|\mathbf{w}_r^P - \mathbf{w}_r^F\|_1, \end{aligned}$$

where

$$w_{r,\mathbf{j}}^P := \left\{ P \left(\left(\frac{j_1 - 1}{r}, \frac{j_1}{r} \right] \times \dots \times \left(\frac{j_d - 1}{r}, \frac{j_d}{r} \right] \right) \right\}_{\mathbf{j} \in \mathcal{H}_{r,d}^0},$$

and $\mathbf{w}_r^P := \{w_{r,\mathbf{j}}^P\}_{\mathbf{j} \in \mathcal{H}_{r,d}^0} \in \Delta_{\frac{(r+d-1)!}{d!(r-1)!}}$. Therefore, if $\|\mathbf{w}_r^P - \mathbf{w}_r^F\|_1 \leq \bar{\epsilon}_n^2/(r^2 + d - 1)^d$, then

$\|\tilde{b}_{r^2,F} - \tilde{b}_{r^2,P}\|_\infty \leq \bar{\epsilon}_n^2$, i.e., the number of L_1 -balls of radius $\bar{\epsilon}_n^2$ needed to cover the set $\mathcal{B}_r$ can

be upper bounded by the numbers of L_1 -balls of radius $\bar{\epsilon}_n^2/(r^2 + d - 1)^d$ required to cover the

simplex space $\Delta_{\frac{(r+d-1)!}{d!(r-1)!}}$. Thus, by Lemma A.4 in Ghosal & van der Vaart (2001), it follows that

$$\begin{aligned} D(\bar{\epsilon}_n, \mathcal{F}_n, H) &\leq \sum_{r=1}^{s_n} D(\bar{\epsilon}_n^2, \mathcal{B}_r, \|\cdot\|_1), \\ &\leq \sum_{r=1}^{s_n} D\left(\frac{\bar{\epsilon}_n^2}{(r^2 + d - 1)^d}, \Delta_{\frac{(r+d-1)!}{d!(r-1)!}}, \|\cdot\|_1\right), \\ &\leq \sum_{r=1}^{s_n} \left(\frac{5(r^2 + d - 1)^d}{\bar{\epsilon}_n^2} \right)^{\frac{(r+d-1)!}{d!(r-1)!} - 1}, \\ &\leq s_n \left(\frac{5(s_n^2 + d - 1)^d}{\bar{\epsilon}_n^2} \right)^{(s_n + d - 1)^d}. \end{aligned}$$

This implies that

$$\begin{aligned} \log(D(\bar{\epsilon}_n, \mathcal{F}_n, H)) &\leq \log s_n + (s_n + d - 1)^d \log \left(\frac{5(s_n^2 + d - 1)^d}{\bar{\epsilon}_n^2} \right), \\ &\leq C_5 \log n + C_5 s_n^d \log \left(\frac{5c_4 s_n^{2d}}{\bar{\epsilon}_n^2} \right), \\ &\leq C_5 s_n^d \log n, \\ &\leq C_5 n \bar{\epsilon}_n^2, \end{aligned}$$

where C_5 and c_4 are suitable positive constants. Therefore, condition (11) is also satisfied,

which implies that the proof of the theorem is completed.

Acknowledgements

A. F. Barrientos's research is supported by NSF ACI 1443014 grant. The second author was supported by Fondecyt 1141193 grant. The authors thank one referee for valuable comments that have led to substantial improvements in the paper.

References

- ASH, J. M. (2008). Mean value theorems for differences. *Elemente der Mathematik* 63 122–125.
- BARRIENTOS, A. F., JARA, A. & QUINTANA, F. A. (2015). Bayesian density estimation for compositional data using random Bernstein polynomials. *Journal of Statistical Planning and Inference* 166 116–125.
- ESCOBAR, M. D. & WEST, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90 577–588.
- FERGUSON, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In D. Siegmund, J. Rustage & G. G. Rizvi, eds., *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*. Bibliohound, 287–302.
- GHOSAL, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics* 29 1264–1280.

- GHOSAL, S., GHOSH, J. K. & RAMAMOORTHY, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* 27 143–158.
- GHOSAL, S., GHOSH, J. K. & VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* 28 500–531.
- GHOSAL, S. & VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* 29 1233–1263.
- KRUIJER, W. & VAN DER VAART, A. W. (2008). Posterior convergence rates for Dirichlet mixtures of beta densities. *Journal of Statistical Planning and Inference* 138 1981–1992.
- LINDSAY, B. (1995). *Mixture Models: Theory, Geometry and Applications*. California, USA: IMS Hayward.
- LO, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *The Annals of Statistics* 12 351–357.
- PETRONE, S. (1999a). Bayesian density estimation using Bernstein polynomials. *The Canadian Journal of Statistics* 27 105–126.
- PETRONE, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics* 26 373–393.
- SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York, USA: Chapman and Hall.