University of Toronto
Department of Economics

VELUT ARBOR ÆVO

Working Paper 473

Copula Based Factorization in Bayesian Multivariate Infinite Mixture Models

By Martin Burda and Artem Prokhorov

January 28, 2013

Copula Based Factorization in Bayesian Multivariate Infinite Mixture Models*

Martin Burda[†] Artem Prokhorov[‡]

January 27, 2013

Abstract

Bayesian nonparametric models based on infinite mixtures of density kernels have been recently gaining in popularity due to their flexibility and feasibility of implementation even in complicated modeling scenarios. In economics, they have been particularly useful in estimating nonparametric distributions of latent variables. However, these models have been rarely applied in more than one dimension. Indeed, the multivariate case suffers from the curse of dimensionality, with a rapidly increasing number of parameters needed to jointly characterize each mixing component. In this paper, we propose a factorization scheme for nonparametric mixture models whereby each marginal dimension in the mixing parameter space is modeled separately, linked by a nonparametric random copula function. Specifically, we consider nonparametric univariate Gaussian mixtures for the marginals and a multivariate random Bernstein polynomial copula for the link function, under Dirichlet process priors. We show that this scheme leads to an improvement in the precision of a density estimate in finite samples, providing a suitable tool for applications in higher dimensions. We derive weak posterior consistency of the copula-based mixing scheme for general kernel types under high-level conditions, and strong posterior consistency for the specific Bernstein-Gaussian mixture model.

JEL: C11, C14, C63
Keywords: Nonparametric copula; nonparametric consistency; mixture modeling

*We would like to thank the participants of the 6^{th} Annual Bayesian Econometric Workshop of the Rimini Center for Economic Analysis, Toronto, 2012 for their insightful comments and suggestions. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and it was supported by grants from the Social Sciences and Humanities Research Council of Canada (SSHRC: www.sshrc-crsh.gc.ca).

[†]Department of Economics, University of Toronto, 150 St. George St., Toronto, ON M5S 3G7, Canada; Phone: (416) 978-4479; Email: `martin.burda@utoronto.ca`

[‡]Department of Economics, Concordia University, and CIREQ, 1455 de Maisonneuve Blvd West, Montreal, QC H3G 1M8, Canada; Phone: (514) 848-2424 ext.3908; Email: `artem.prokhorov@concordia.ca`

1. Bayesian Nonparametric Copula Kernel Mixture

Bayesian infinite mixture models are useful both as nonparametric estimation methods and as a way of uncovering latent class structure that can explain the dependencies among the model variables. Such models express a distribution as a mixture of simpler distributions without a priori restricting the number of mixture components which is stochastic and data-driven. In many contexts, a countably infinite mixture is also a more realistic model than a mixture with a small fixed number of components or "types".

Even though their theoretical foundations were developed early (Ferguson, 1973; Antoniak, 1974; Lo, 1984), infinite mixture models have only recently become computationally feasible for practical implementation on larger data sets with the development of Markov chain Monte Carlo (MCMC) methods (Escobar and West, 1995; Neal, 2000). In economics, Bayesian infinite mixture models are becoming increasingly popular with recent applications in treatment effects (Chib and Hamilton, 2002), autoregressive panel data (Hirano, 2002), finance (Jensen and Maheu, 2010), latent heterogeneity in discrete choice models (Kim, Menzefricke, and Feinberg, 2004; Burda, Harding, and Hausman, 2008), contingent valuation models (Fiebig, Kohn, and Leslie, 2009), and instrumental variables (Conley, Hansen, McCulloch, and Rossi, 2008).

However, the bulk of applications of Bayesian infinite mixture models has been univariate, or structured as conditionally independent copies of the univariate case. Indeed, the curse of dimensionality poses an acute practical problem in the multivariate case when explicit full dependence modeling is required. Infinite mixture models are formulated as expectations with respect to the parameters of a so-called mixing kernel. When the mixing kernel is defined jointly over more than one dimension, such as the multivariate Gaussian density kernel, the size of the associated parameter vector increases faster than the number of dimensions. Popular nonparametric priors, such as the Dirichlet process, can deal with this problem by selecting relatively few latent classes to form the resulting mixture in the implementation.[1] However, this relative sparsity is to the detriment of the accuracy of estimation. The prior can be set up by the analyst to generate a large number of latent classes newly proposed during the MCMC run, such as by adjusting the prior density of the concentration parameter in the Dirichlet process mixture. But many of such proposed classes will either not be accepted or quickly discarded since they are likely to be formed over regions of the parameter space that are only weakly supported by the data, leading to a more noisy estimate. Furthermore, the increase of the number of latent classes happens

[1] In contrast, frequentist kernel density estimation adds a new kernel for each data point taking the full impact of the curse of dimensionality. Frequentist latent class models based on a few mixture components need to be optimized over the number of components, typically using information criteria, which may be computationally very cumbersome – in practice, the number of components in such models is often selected ad hoc.

at the expense of rapidly increasing the number of mixing parameters. The severity of this tradeoff is further exacerbated with higher dimensions: few mixing components can provide a very inaccurate representation of the data generating process, but strengthening the prior to increase their number will yield a higher rejection rate and computational cost.

From our practical experience, the crux of this tradeoff lies in the joint parametric specification of the mixing kernel, such as in the case of the typically used multivariate Gaussian kernel. If such joint parametric mixing kernel can be decomposed into flexible building blocks, each of which can be parsimoniously determined with high probability strongly supported by the data then we should obtain a more accurate representation. However, conditional expansions such as the Cholesky factorization of the covariance matrix of the multivariate Gaussian kernel still preserve the joint parametric dependence structure of the kernel. What is needed for our purpose is a decomposition of the dependence structure itself.

In this paper, we propose a factorization scheme for Bayesian nonparametric mixture models whereby each marginal dimension in the mixing parameter space is modeled as a separate infinite mixture and these marginal models are then joined by a nonparametric copula function based on random Bernstein polynomials capturing the joint dependence structure in a flexible way. In the implementation, only a few latent classes are required for each of the marginals regardless of the overall number of dimensions. We show that this scheme leads to an improvement in the precision of a density estimate in finite samples, providing a suitable tool for applications where dependence is modeled jointly in higher dimensions. Bearing in mind Freedman's (1963) result concerning a topologically wide class of priors leading to inconsistent posteriors in Bayesian nonparametric models, we specify the conditions under which our approach yields posterior consistency; both for weak topologies for a general class of mixing kernels and strong topologies for the specific case of random Bernstein polynomial copula and Gaussian mixture marginals (Bernstein-Gaussian mixture).

In a related literature, Chen, Fan, and Tsyrennikov (2006) consider a frequentist copula sieve maximum likelihood estimator with a parametric copula and nonparametric marginals, while Panchenko and Prokhorov (2012) consider the converse problem, with parametric marginals and a nonparametric copula. In contrast, our procedure uses both nonparametric copula and marginals. Moreover, in our case the number of mixture components is data driven for any sample size within the estimation procedure.

Nonparametric copula-based mixture models have been analyzed in several specific contexts distinct from ours. Silva and Gramacy (2009) present various MCMC proposals for copula mixtures. Fuentes, Henry, and Reich (2012) analyze a spatial Dirichlet process (DP) copula model based on the stick-breaking DP representation. Rey and Roth (2012) introduce a copula mixture model to perform dependency-seeking clustering when co-occurring samples from different data sources are available. Their model features nonparametric marginals and a Gaussian copula with block-diagonal correlation matrix. Rodriguez, Dunson, and Gelfand (2010) construct a stochastic process where observations at different locations are dependent, but have a common marginal distribution. Dependence across locations is introduced by using a latent Gaussian copula model, resulting in a latent stick-breaking process. Parametric Bayesian copula models and their mixtures have also been analyzed in Pitt, Chan, and Kohn (2006), Silva and Lopes (2008), Ausin and Lopes (2010), and Giordani, Mun, and Kohn (2012), among others.

Posterior consistency[2] of Bayesian parametric models is ensured by a general theorem of Doob (1949) and consistency issues are of little concern to the parametric analyst. However, the situation changes in Bayesian nonparametric analysis: posterior consistency can fail in infinite-dimensional spaces for quite well-behaved models even for seemingly natural priors (Freedman, 1963; Diaconis and Freedman, 1986; Kim and Lee, 2001). In particular, the condition of assigning positive prior probabilities in "usual" neighborhoods of the true parameter is not sufficient to ensure consistency. The implication of Freedman's (1963) result is that the collection of all possible "good pairs" of true parameter values and priors which lead to consistency is extremely narrow when the size is measured topologically. A set F is called *meager* and considered to be topologically small if F can be expressed as a countable union of sets C_i, $i \geq 1$, whose closures $\overline{C}_i$ have empty interior. Freedman (1963) showed that the collection of the "good pairs" is meager in the product space (Ghosal, 2010). It is therefore important to provide the set of conditions that ensure a proposed nonparametric modeling scenario fits into the meager collection in infinite-dimensional spaces.

For the Gaussian kernel and the Dirichlet Process prior of the mixing distribution, asymptotic properties, such as consistency, and rate of convergence of the posterior distribution based on kernel mixture priors were established by Ghosal, Ghosh, and Ramamoorthi (1999), Tokdar (2006), and Ghosal and van der Vaart (2001, 2007). Similar results for Dirichlet mixture of Bernstein polynomials were shown by Petrone and Wasserman (2002), Ghosal (2001) and Kruijer and van der Vaart (2008). Petrone and Veronese (2010) derived consistency for general kernels under the strong condition that the true density is exactly of the mixture type for some compactly supported mixing distribution, or the true density itself is compactly supported and is approximated in terms of Kullback-Leibler divergence

[2]Although consistency is intrisically a frequentist property, it implies an eventual agreement among Bayesians with different priors. For a subjective Bayesian who dispenses of the notion of a true parameter, consistency has an important connection with the stability of predictive distributions of future observations – a consistent posterior will tend to agree with calculations of other Bayesians using a different prior distribution in the sense of weak topology. For an objective Bayesian who assumes the existence of an unknown true model, consistency can be thought of as a validation of the Bayesian method as approaching the mechanism used to generate the data (Ghosal and van der Vaart, 2011). In this sense, it is desirable to establish the conditions under which infinite mixture models have a consistent posterior, as the use of models that are inconsistent can be regarded as ill-advised.

by its convolution with the chosen kernel. Wu and Ghosal (2008, 2010) showed consistency for a class of location-scale kernels and kernels with bounded support. Shen and Ghosal (2011) and Canale and Dunson (2011) showed consistency for DP location-scale mixtures based on multivariate Gaussian kernel. We extend on these results by analyzing a kernel composed of a copula density with location-scale marginals. We use some of these results as building blocks in deriving posterior consistency of our mixture model.

The remainder of the paper is organized as follows. In Section 2 we provide the details of our proposed copula-based factorization scheme, along with a specific kernel choices of Gaussian mixtures for the marginals and Bernstein random polynomial copula for the link function. In Section 3, we derive posterior consistency in the weak topology for general classes of kernels under high level conditions and the Bernstein-Gaussian case under low level conditions. In Section 3 we show strong posterior consistency for the Bernstein-Gaussian scenario. In Section 4 we present the results of a Monte Carlo experiment, comparing the accuracy of estimation of a multivariate joint density between our Bernstein-Gaussian copula mixture and the popular multivariate Gaussian mixture. Section 5 concludes.

2. Factorization for Infinite Mixture Models

2.1. Setup

Throughout, we will use the notation and terminology of Wu and Ghosal (2008), henceforth WG, where applicable. Let $\mathcal{X}$ be the sample space with elements x, Θ the space of the mixing parameter θ, and Φ the space of the hyperparameter ϕ. Let $\mathcal{D}(\mathcal{X})$ denote the space of probability measures F on $\mathcal{X}$. Denote by $\mathcal{M}(\Theta)$ the space of probability measures on Θ and let P be the mixing distribution on Θ with density p and a prior Π on $\mathcal{M}(\Theta)$ with weak support $supp(\Pi)$. Denote the prior for ϕ by μ, and the support of μ by $supp(\mu)$, with μ independent of P. Let $K(x;\theta,\phi)$ be a kernel on $\mathcal{X}\times\Theta\times\Phi$, such that $K(x;\theta,\phi)$ is a jointly measurable function with the property that for all $\theta\in\Theta$ and $\phi\in\Phi$, $K(\cdot;\theta,\phi)$ is a probability density on $\mathcal{X}$.

P, μ and $K(x;\theta,\phi)$ induce a prior on $\mathcal{D}(\mathcal{X})$ via the map

$$(\phi,P)\mapsto f_{P,\phi}(x)\equiv\int K(x;\theta,\phi)dP(\theta) \tag{2.1}$$

(the so-called type II mixture prior in WG). Denote such composite prior by Π^*.

2.2. Copula Kernel

We consider a kernel with the structure

$$K(x;\theta,\phi)=K_c(F(x;\theta_m,\phi_m);\theta_c,\phi_c)K_m(x;\theta_m,\phi_m) \tag{2.2}$$

where

$$K_m(x;\theta_m,\phi_m) = \prod_{s=1}^{d} K_{ms}(x_s;\theta_{ms},\phi_{ms}) \tag{2.3}$$

is the product of univariate kernels of the marginals in d dimensions, $K_c(F(x;\theta_m,\phi_m);\theta_c,\phi_c)$ is a copula density kernel, $\theta = \{\theta_m,\theta_c\}$, and $\phi = \{\phi_m,\phi_c\}$. The arguments of $K_c(\cdot)$ consist of a d-vector of distribution functions of the marginals

$$F(x;\theta_m,\phi_m) = \int_{-\infty}^{x} K_m(t;\theta_m,\phi_m)dt$$

with copula parameter vectors θ_c and ϕ_c.

The copula counterpart of the mixed joint density (2.1) takes the form

$$f_{P,\phi}(x) = \int K_c(F(x;\theta_m,\phi_m);\theta_c,\phi_c)K_m(x;\theta_m,\phi_m)dP(\theta) \tag{2.4}$$

The mixing parameter θ enters through both the marginals and the copula which complicates the analysis relative to cases when $K(x;\theta,\phi)$ in (2.1) is a single kernel, such as the multivariate Gaussian.

2.3. **Random Bernstein Polynomial Copula Density**

Let $P(\theta) = P_c(\theta_c) \times P_m(\theta_m)$. Let $[0,1]^d$ denote the unit cube in $\mathbb{R}^d$ where d is a positive integer. For a distribution function $P_c : [0,1]^d \to \mathbb{R}$, a multivariate Bernstein polynomial of order $\mathbf{k} = (k_1,\ldots,k_d)$ associated with P_c is defined as

$$B_{\mathbf{k},P_c}(\mathbf{u}) = \sum_{j_1=0}^{k_1} \cdots \sum_{j_d=0}^{k_d} P_c(\mathbf{j}/\mathbf{k}) \prod_{s=1}^{d} q_{j_s k_s}(u_s) \tag{2.5}$$

where $\mathbf{u} = (u_1,\ldots,u_d) \in [0,1]^d$, $\mathbf{j} = (j_1,\ldots,j_d)$, $\mathbf{k} = (k_1,\ldots,k_d)$, $q_{j_s k_s}(u_s) = \binom{k_s}{j_s} u_s^{j_s}(1 - u_s)^{k_s - j_s}$ for $0 \le u_s \le 1$, $j_s = 0,\ldots,k_s$, $k_s = 1,2,\ldots$ and $s = 1,\ldots,d$. The order $\mathbf{k}$ controls the smoothness of $B_{\mathbf{k},P_c}$, with a smaller k_s associated with a smoother function along the dimension s. For $P_c(0,1,\ldots,1) = P_c(1,\ldots,1,0) = 0$, $B_{\mathbf{k},P_c}(\mathbf{u})$ is a probability distribution function on $[0,1]^d$ and is referred to as the Bernstein distribution associated with P_c. As $\min\{\mathbf{k}\} \to \infty$, $B_{\mathbf{k},P_c}(\mathbf{u})$ converges to P_c at each continuity point of P_c and if P_c is continuous then the convergence is uniform on the unit cube $[0,1]^d$ (Sancetta and Satchell, 2004; Zheng, 2011).

The derivative of (2.5) is the multivariate Bernstein density function

$$
\begin{aligned}
b_{\mathbf{k},P_c}(\mathbf{u}) &= \frac{\partial^d}{\partial\omega_1\cdots\partial\omega_d} B_{\mathbf{k},P_c}(\mathbf{u}) \\
&= \sum_{j_1=1}^{k_1}\cdots\sum_{j_d=1}^{k_d} w_{\mathbf{k},P_c}(\mathbf{j}) \prod_{s=1}^{d} \beta(u_s; j_s, k_s - j_s + 1)
\end{aligned} \tag{2.6}
$$

where $w_{\mathbf{k},P_c}(\mathbf{j}) = \mathbf{\Delta}^{(1,\ldots,1)} P_c((\mathbf{j}-\mathbf{1})/\mathbf{k})$ are mixing weights derived using the forward difference operator $\mathbf{\Delta}$, and $\beta(\cdot;\gamma,\delta)$ denotes the probability density function of the Beta distribution with parameters γ and δ. Let Cube$(\mathbf{j},\mathbf{k})$ denote the d-dimensional cube of the form $((j_1-1)/k_1, j_1/k_1] \times \ldots \times ((j_d-1)/k_d, j_d/k_d]$ with the convention that if $j_s = 0$ then the interval $((j_s-1)/k_s, j_s/k_s]$ is replaced by the point $\{0\}$. The mixing weights $w_{\mathbf{k},P_c}(\mathbf{j})$ are the probabilities of Cube$(\mathbf{j},\mathbf{k})$ under P_c. The Bernstein density function $b_{\mathbf{k},P_c}(\mathbf{u})$ can thus be viewed as a mixture of beta densities, and is a probability density function over $[0,1]^d$.

It can also be viewed as a smoothed version of a discrete, or empirical, copula. As an example, let $d = 2$ and define a $k_1 \times k_2$ doubly stochastic matrix M, such that $M'\mathbf{1} = M\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is a vector of ones. It can be shown that $\frac{1}{k_1 k_2} M$ is a discrete bivariate copula density on

$$
\left[\frac{1}{k_1}, \frac{2}{k_1}, \ldots, \frac{k_1}{k_1}\right] \times \left[\frac{1}{k_2}, \frac{2}{k_2}, \ldots, \frac{k_2}{k_2}\right]
$$

Now consider smoothing $\frac{M}{k_1 k_2}$. Define a vector of smoothing functions

$$
\mathbf{f}_j(u) = \left(f_{j1}, \ldots, f_{jk_j}\right)'
$$

for $j = 1, 2$, such that $\int f_i(u) d(u) = 1, i = 1, ..., k_j$, and $\mathbf{1}'\mathbf{f}_j(u) = 1$, for any $u \in [0,1]$. The function $b(u_1, u_2) = \frac{1}{k_1 k_2} \mathbf{f}_1(u_1)' M \mathbf{f}_2(u_2)$ is a smoothed version of the discrete copula $\frac{M}{k_1 k_2}$. It can be written as $\sum_{j=1}^{k_1}\sum_{l=1}^{k_2} \frac{m_{jl}}{k_1 k_2} f_{1j}(u_1) f_{2l}(u_2)$, where m_{jl} is the relevant element of M. Then, the Bernstein copula density in (2.6) is obtained if $\mathbf{f}_j(u)$ contains β-densities with parameters $\{(a,b) : a + b = k_j + 1\}$, that is, if

$$
\mathbf{f}_j(u) = (\beta(u; 1, k_j), \beta(u; 2, k_j - 1), \ldots, \beta(u; k_j, 1))'
$$

where $\beta(u; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1} (1-u)^{b-1}$.

Being a mixture of (a product of) β-densities, the Bernstein copula density assigns no weight outside $[0,1]^d$ and thus avoids any boundary problems. It is a density by construction; at the same time, it does not impose symmetry, contrary to the conventional kernels such as multivariate Gaussian (Wu and Ghosal, 2008, 2010). As a density corresponding to $B_{\mathbf{k},P_c}(\mathbf{u})$, $b_{\mathbf{k},P_c}(\mathbf{u})$ converges, as $\min\{\mathbf{k}\} \to \infty$, to $p_c(\mathbf{u})$ at every point on $[0,1]^d$ where $\partial^d/\partial\omega_1\cdots\partial\omega_d P_c(\mathbf{u}) = p_c(\mathbf{u})$ exists, and if p_c is continuous and bounded then the convergence is uniform (Lorentz, 1986). Uniform approximation results for the univariate and bivariate Bernstein density estimator can be found in Vitale (1975) and Tenbusch (1994).

Petrone (1999a,b) proposed a class of prior distributions on the set of densities defined on $(0, 1]$ based on the univariate Bernstein polynomials and Petrone and Wasserman (2002) showed weak and strong consistency of the Bernstein polynomial posterior for the space of univariate densities on $(0, 1]$. Zheng, Zhu, and Roy (2010) and Zheng (2011) extended the settings to the multivariate case, where $\mathbf{k}$ is an $\mathbb{N}^d$-valued random variable and P_c is a random probability distribution function, yielding $B_{\mathbf{k},P_c}$ as a random function. A random Bernstein density $b_{\mathbf{k},P_c}(\mathbf{u})$ thus features random mixing weights $w_{\mathbf{k},P_c}(\mathbf{j})$ with $\mathbf{w}_{\mathbf{k},P_c} = \{w_{\mathbf{k},P_c}(\mathbf{j}) : j_s = 1, \ldots, k_s, s = 1, \ldots, d\}$ belonging to the $\prod_{s=1}^{d} k_s - 1$ dimensional simplex

$$\mathsf{s}_{\mathbf{k}} = \left\{ \mathbf{w}_{\mathbf{k},P_c}(\mathbf{j}) : w_{\mathbf{k},P_c}(\mathbf{j}) \geq 0, \sum_{j_1=1}^{k_1} \cdots \sum_{j_d=1}^{k_d} \mathbf{w}_{\mathbf{k},P_c}(\mathbf{j}) = 1 \right\}$$

We adopt the multivariate Bernstein density function as a particular case of the copula density kernel in (2.2):

$$K_c(F(x;\theta_m,\phi_m);\theta_c,\phi_c) = b_{\mathbf{k},P_c}(F(x;\theta_m,\phi_m)) \tag{2.7}$$

Let $\{u_{s1}, u_{s2}, \ldots\}$ denote a sequence of exchangeable random variables with values in $[0, 1]$ for $s = 1, \ldots, d$. Conditional on the marginal parameters θ_m and ϕ_m, $u_{si} = F_s(x_{si};\theta_{ms},\phi_{ms})$. A multivariate version of the hierarchical mixture model based on the Bernstein-Dirichlet prior proposed in Petrone (1999b), p. 383, can be specified as follows:

$$\begin{aligned}
u_{si}|y_{si}, P_c, k_s &\sim \beta(u_{si}; j_s, k_s - j_s + 1) \\
\text{if } y_{si} &\in ((j_s - 1)/k_s, j_s/k_s], \text{ for each } s = 1, \ldots, d \\
\mathbf{y}_i|P_c, \mathbf{k} &\sim P_c \\
P_c|\mathbf{k} &\sim DP(\alpha_c, P_{c0}) \\
\mathbf{k} &\sim \mu(\mathbf{k})
\end{aligned}$$

where $\mathbf{y}_i = (y_{1i}, \ldots, y_{di})$ are latent random variables determining the hidden labels associated with $\mathbf{u}_i = (u_{1i}, \ldots, u_{di})$. In our case, $\theta_c = \{\mathbf{y}_i\}_{i=1}^n$ and $\phi_c = \mathbf{k}$. P_{c0}, a probability measure on $[0, 1]^d$ that is absolutely continuous with respect to the Lebesgue measure, is the baseline of the Dirichlet process $DP(\alpha_c, P_{c0})$ with concentration parameter α_c. We set P_{c0} to be uniform on $[0, 1]^d$ and, following Petrone (1999a), $\alpha_c = 1$. We further specify $\mu(\mathbf{k})$ as the Dirichlet distribution $Dir\left(\{j_s/k_s\}_{j_s=1}^{k_s}; 1/k_s\right)$ for $k_s \leq k_s^{\max}$ for each $s = 1, \ldots, d$. The posterior then follows directly from Petrone (1999a) who also proposes a sampling algorithm for the posterior that we follow in the implementation.

For the marginal univariate kernel in (2.3) we consider a product of univariate Gaussian kernels

$$K_{ms}(x_{si};\theta_{ms},\phi_{ms}) = (2\pi)^{-1/2}\sigma_s^{-1} \exp\left(-(x_s - \nu_s)^2/(2\sigma_s^2)\right) \tag{2.8}$$

with $\theta_{ms} = \{\nu_s, \sigma_s^2\}$ and ϕ_{ms} being a vacuous parameter. The prior structure for the marginal is then

$$\begin{aligned} x_{si} &\sim N(x_{si}; \theta_{ms}) \\ \theta_{ms}|P_m &\sim P_m \\ P_m &\sim DP(\alpha_m, P_{m0}) \\ \alpha_m &\sim Gamma(\alpha_{m01}, \alpha_{m02}) \end{aligned}$$

with P_{m0} for $\{\nu_s, \sigma_s^2\}$ composed of $N(\nu_{s0\nu}, \sigma_{s0\nu}^2)$ and $InvGamma(\gamma_{s01}, \gamma_{s02})$, respectively.

3. Weak Consistency

In this section we first specify general high-level conditions on $K_c(u; \cdot)$, $K_m(x; \cdot)$, f_0, and the associated priors under which $f_{P,\phi}(x)$ is consistent at f_0. These general conditions cover a wide range of copula and marginal kernels. We then verify that the general conditions are satisfied under a set of low-level conditions for the specific case of the Bernstein polynomial copula and Gaussian marginals.

Schwartz (1965) showed that posterior consistency at a "true density" f_0 holds if the prior assigns positive probabilities to a specific type of neighborhoods of f_0 defined by Kullback-Leibler divergence measure (the so-called Kullback-Leibler property) and the size of the model is restricted in some appropriate sense. For the weak topology, the size condition for the weak consistency holds automatically (Ghosal, Ghosh, and Ramamoorthi 1999, Theorem 4.4.2). Thus the Kullback-Leibler (K-L) property is tantamount to weak posterior consistency.

Let $\mathcal{F}$ denote the set of all possible joint densities with respect to probability measures in $\mathcal{D}$. Define a Kullback-Leibler (K-L) neighborhood of a density $f \in \mathcal{F}$ of size ε by

$$\mathcal{K}_\varepsilon(f) \equiv \{g \in \mathcal{F} : \mathcal{K}(f; g) < \varepsilon\}$$

where

$$\mathcal{K}(f; g) = \int f \log(f/g)$$

is the K-L divergence between f and g. By convention, we say that the K-L property holds at $f_0 \in \mathcal{F}$ or f_0 is in the K-L support of Π^*, and write $f_0 \in$KL(Π^*), if $\Pi^*(\mathcal{K}_\varepsilon(f_0)) > 0$ for every $\varepsilon > 0$.

WG's Theorem 1 and Lemmas 2 and 3 specify high-level conditions under which the K-L property holds for a mixture density $f_{P,\phi}(x)$ of the generic kernel $K(x; \theta, \phi)$ in (2.1). They further show that these conditions are satisfied under lower-level conditions for specific kernel types, such as the location-scale kernel, gamma kernel, random histogram, and the Bernstein polynomial kernel.

Our analysis will proceed similarly by showing that the high-level conditions of WG's Theorem 1 and Lemmas 2 and 3 are satisfied for the copula kernel in (2.2). Our copula kernel is a composite function of the special cases treated in WG – the location scale kernel and Bernstein polynomial kernel – in that an integral of the former enters as an argument of the latter. Hence WG conditions derived for each kernel separately need to be further developed and linked together to show consistency of the resulting composite copula kernel which we do here. In the Technical Appendix, we restate WG Theorem 1, and Lemmas 2 and 3 for the sake of completeness using our notation so that we can seamlessly refer to the given assumptions throughout the text.

3.1. **The K-L Property of a General Copula Mixture**

The copula kernel $K_c(\cdot)$ contains an integral expression and it is difficult to impose low-level conditions on it directly without specifying its functional form. Hence we will state the following Theorem in terms of relatively high-level conditions which will be subsequently verified for a specific functional form of the copula kernel. The Theorem can be somewhat loosely regarded as the copula kernel counterpart of WG Theorem 2, but under higher-level conditions. Whenever ϕ and θ share a common general prior, we will drop ϕ from the notation without loss of generality.

Assume the following set of conditions hold.

B1. For some $0 < \overline{f} < \infty$, $0 < f_0(x) \leq \overline{f}$ for all x;
B2. For some $0 < \overline{p} < \infty$, $0 < p(\theta) < \overline{p}$ for all θ;
B3. $K_m(x; \theta_m)$ is continuous in x, positive, bounded and bounded away from zero everywhere;
B4. $K_c(\cdot)$, $K_m(\cdot)$, $\log f_{P_\varepsilon}(x)$, $\log K_c(\cdot) K_m(\cdot)$, and $\inf_{\theta \in D} K_c(\cdot) K_m(\cdot)$ are f_0-integrable, the latter for some closed $D \supset supp(P_\varepsilon)$;
B5. For some $0 < \overline{K} < \infty$, $\int K_c\left(\int_{-\infty}^{x} K_m(t; \theta_m) dt; \theta_c\right) K_m(x; \theta_m) d\theta = \overline{K}$ for all x;
B6. The weak support of Π is $\mathcal{M}(\Theta)$.

Condition B1 requires the true density to be bounded and bounded away from zero. Conditions B2 and B3 specify regularity conditions and f_0-integrability of the kernels and their integrals. There is a variety of different copula and location-scale kernel choices that satisfy these conditions. Conditions B4 and B5 provide integrability and boundedness restrictions on the copula and marginal densities. Condition B6 is relatively weak and does not make any specific assumptions on Π other than requiring that it has large weak support. Thus, Π includes a wide class of priors such as the Dirichlet process or the Pólya tree process.

Theorem 1. *Let $f_0(x)$ be the true density and Π^* an induced prior on $\mathcal{F}(\mathcal{X})$ with the kernel function*

$$K(x;\theta) = K_c\left(\int_{-\infty}^{x} K_m(t;\theta_m)dt;\theta_c\right) K_m(x;\theta_m)$$

implying $P \sim \Pi$, and given P, $\theta \sim P$. If $K_c(\cdot)$, $K_m(\cdot)$, and $f_0(x)$ satisfy conditions B1–B6 then $f_0 \in KL(\Pi^*)$.

Proof of Theorem 1:

The proof is based on invoking WG Theorem 1, stated in the Technical Appendix, and showing that its conditions A1 and A3 are met. Since in Theorem 1 ϕ is subsumed as a part of θ, a separate case for Condition A2 is redundant. To satisfy Condition A1, it suffices to show that for each ε there exists $f_{P_{\varepsilon r}}(x)$ such that

$$\lim_{r\to\infty} f_0(x)\log\frac{f_0(x)}{f_{P_{\varepsilon r}}(x)} = 0 \tag{3.1}$$

pointwise and that

$$\left| f_0(x)\log\frac{f_0(x)}{f_{P_{\varepsilon r}}(x)}\right| < C < \infty \tag{3.2}$$

for all $r > r^*$ for some $r^* < \infty$. Then, by the Dominated Convergence Theorem (DCT),

$$\lim_{r\to\infty}\int f_0(x)\log\frac{f_0(x)}{f_{P_{\varepsilon r}}(x)}dx = 0$$

which implies that Condition A1 is satisfied.

For each ε, define the probability density on a compact truncation of the parameter space as

$$p_{\varepsilon r}(\theta) = \begin{cases} v_r p(\theta), & \|\theta\| < r,\ r \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

and

$$v_r^{-1} = \int_{\|\theta\|<r} p(\theta)d\theta,$$

Let $P_{\varepsilon r}$ denote the probability measure corresponding to $p_{\varepsilon r}$. Thus, by construction,

$$\lim_{r\to\infty} p_{\varepsilon r}(\theta) = p_\varepsilon(\theta) \tag{3.4}$$

Define

$$f_{P_{\varepsilon r}}(x) = \int K_c\left(\int_{-\infty}^{x} K_m(t;\theta_m)dt;\theta_c\right) K_m(x;\theta_m)p_{\varepsilon r}(\theta)d\theta \tag{3.5}$$

Using conditions B2, B3, and (3.4), another invocation of the DCT yields

$$\lim_{r\to\infty} f_{P_{\varepsilon r}}(x) = f_0(x) \tag{3.6}$$

pointwise in x for each θ. (3.6) then implies (3.1) using composition of limits and condition B1.

In order to show (3.2), we will provide bounds for the sequence $f_{P_{\varepsilon r}}(x)$ in r. First, using (3.3) and conditions B2 and B5,

$$\begin{aligned} f_{P_{\varepsilon r}}(x) &= v_r \int_{\|\theta\|<r} K_c\left(\int_{-\infty}^{x} K_m(t;\theta_m)dt;\theta_c\right) K_m(x;\theta_m)p(\theta)d\theta \\ &< v_r\overline{p} \int_{\|\theta\|<r} K_c\left(\int_{-\infty}^{x} K_m(t;\theta_m)dt;\theta_c\right) K_m(x;\theta_m)d\theta \\ &\leq v_r\overline{p}\overline{K} \qquad (3.7) \end{aligned}$$

Due to (3.7), for all r, since $f_{P_{\varepsilon r}}(x) \to f_0(x)$ and $v_r \to 1$ as $r \to \infty$, there exists an r^* such that for all $r > r^*$,

$$\frac{f_0(x)}{v_r\overline{p}\overline{\overline{K}}} < 1 \qquad (3.8)$$

Combining (3.7) and (3.8),

$$\log\frac{f_0(x)}{f_{P_r}(x)} \geq \log\frac{f_0(x)}{\overline{p}v_r} \qquad (3.9)$$

Second, by Conditions B2, B3 and (3.5) there exists a function $g(x)$ such that $f_{P_{\varepsilon r}}(x) \geq g(x) > 0$ for all r and $x \in \mathcal{X}$, and hence

$$\log\frac{f_0(x)}{f_{P_{\varepsilon r}}(x)} \leq \log\frac{f_0(x)}{g(x)} \qquad (3.10)$$

The inequalities (3.9) and (3.10) combined yield (3.2), thus completing verification of Condition A1.

To show Condition A3, it suffices to verify conditions A7–A9 of WG Lemma 3, stated in the Technical Appendix. Let P_ε in the statement of Theorem 1 be chosen to be $P_{\varepsilon r}$ which is compactly supported. By Condition B6, $P_\varepsilon \in supp(\Pi)$. Condition A7 is satisfied by condition B4. Condition A8 is satisfied by Condition B3. Condition A9 is satisfied by condition B5. □

Now consider the case of the full copula kernel specification of (2.2) where $\phi = \{\phi_m, \phi_c\}$ is a hyperparameter with prior $\mu = \mu_m \times \mu_c$ separate from P. Assume that μ and P are a-priori independently distributed. Now the prior Π^* for density functions on $\mathcal{X}$ is induced by $\Pi \times \mu$ via the mapping $(P, \phi) \mapsto f_{P,\phi}$ where $f_{P,\phi}$ is given in (2.4). In this case condition B6 is replaced by the following condition:

B6'. The weak support of Π is $\mathcal{M}(\Theta \times \Phi)$.

Then the following Theorem applies:

Theorem 2. *Under conditions B1-B5 and B6', for $f_{P,\phi}$ given in (2.4), $f_0 \in KL(\Pi^*)$.*

Proof of Theorem 2:

The proof of the theorem is virtually identical to the proof of Theorem 1 for verifying Conditions A1 and A3. In contrast to Theorem 1, the weak support of Π is now $\mathcal{M}(\Theta \times \Phi)$ and the Condition A2 of WG Theorem 1 now also needs to be satisfied. Condition A2 holds under conditions A4–A6 given WG Lemma 2, stated in the Technical Appendix. Conditions A4 and A5 are satisfied by our Condition B6'. Condition A6 is satisfied by our Conditions B4 and B5. In summary, under the Conditions A1–A3 we can now invoke WG Theorem 1 completing the proof. □

3.2. **The K-L Property of the Random Bernstein Polynomial Copula**

We will now show that the conditions B1–B5, and B6' of Theorems 1 and 2 are satisfied by the specific cases of the Bernstein random polynomials for the copula (2.7) along with the Gaussian marginal kernel (2.8), under their respective priors. Condition B1 is assumed for the unknown true density function and will be maintained throughout. Conditions B2 and B6' are satisfied by construction for the Dirichlet distribution and Dirichlet process priors considered. Condition B3 is satisfied by the marginal Gaussian kernel provided that its variance stays strictly positive, which is attained by the inverse gamma prior for the inverse variance. For Condition B4, f_0-integrability of $K_c(\cdot)$, $K_m(\cdot)$, $\log f_{P_\varepsilon}(x)$, and $\log K_c(\cdot)\, K_m(\cdot)$ is given by the exponential tails of the Gaussian $K_m(\cdot)$, compact support of $K_c(\cdot)$ and a maintained assumption on f_0. Integrability of $K_c(\cdot)\, K_m(\cdot)$ with respect to θ in Condition B5 holds as long as the variance parameter in $K_m(\cdot)$ is strictly positive, as specified by its prior.

4. **Strong Consistency**

The conditions discussed in previous section ensure weak convergence of the Bayesian estimate of the density to the true density f_0, that is, they guarantee that asymptotically the posterior accumulates in weak neighborhoods of f_0. However, as argued in Barron, Schervish and Wasserman (1989), among others, there are many densities in the weak neighborhood that do not resemble f_0 and consistency in a stronger sense is more appropriate.

4.1. **General conditions for strong posterior consistency**

Define the Hellinger distance

$$d_H(f,g) \equiv \left| \int \left| \sqrt{f} - \sqrt{g} \right|^2 \right|^{1/2}$$

and a strong ε-neighborhood of f_0

$$V_\varepsilon(f_0) = \{f : d_H(f_0, f) < \epsilon\}$$

A posterior is said to be Hellinger consistent (or strongly consistent) at f_0 if $\Pi(V_\varepsilon|\mathbf{x}_1, \ldots, \mathbf{x}_n) \to 1$ in P_{f_0}-probability for any Hellinger neighborhood V_ε of f_0. That is, the posterior will attach a high probability (asymptotically equal to 1) to any strong neighborhood of the true f_0. As before, the approach due to Schwartz (1965) suggests that posterior consistency is obtained by putting appropriate size restrictions on the model and conditions on the support of the prior defined by the Kullback-Leibler property. The K-L property holds by the arguments of the previous section. However, the size condition does not hold automatically for the strong topology and one has to resort to the technique of truncating the parameter space, depending on the sample size (Canale and Dunson, 2011; Ghosal, 2010, Section 2.4). The size condition is imposed by controlling the size of the truncated parameter space, also known as the sieve.

Let $\mathcal{F}_n \subset \mathcal{F}, n \geq 1$, denote a sieve for $\mathcal{F}$, the set of joint densities we consider. A set of pairs of functions $(l_q, u_1), \ldots, (l_r, u_r)$ is called an ε-bracketing for the set $\mathcal{F}_n$ if $d_H(l_j, u_j) \leq \varepsilon$ for $j = 1, \ldots, r$ and for every $f \in \mathcal{F}$, there exists $1 \leq j \leq r$ such that $l_j \leq f \leq u_j$. Let $N_{[]}(\delta, \mathcal{F}_n, d)$ denote the minimum number of brackets of size δ needed to cover $\mathcal{F}_n$, that is,

$$N_{[]}(\delta, \mathcal{F}_n, d) = \min_k \{\mathcal{F}_n \subset \cup_{i=1}^k \{f : l \leq f \leq u, d(l,u) < \delta\}, \mathcal{F}_n \subset \mathcal{F}\} \tag{4.1}$$

and let $J_{[]}(\delta, \mathcal{F}_n, d) = \log N_{[]}(\delta, \mathcal{F}_n, d)$. The number $N_{[]}(\delta, \mathcal{F}_n, d)$ is known as the δ-bracketing number of $\mathcal{F}_n$ with respect to the metric d and its logarithm $J_{[]}(\delta, \mathcal{F}_n, d)$ is known as the metric bracketing entropy (Kolmogorov and Tikhomirov, 1961).

Based on the results of Wasserman (1998) and Barron, Schervish, and Wasserman (1999), Petrone and Wasserman (2002) list the appropriate sufficient conditions for Hellinger consistency as follows:

C1. $\Pi(K_\varepsilon(f_0)) > 0$, i.e. K-L property holds, any ε;
C2. $\Pi(\mathcal{F}_n^c) < c_1 \exp(-nc_2)$ for some constants $c_1, c_2 > 0$;
C3. The entropy $J_{[]}(\delta, \mathcal{F}_n, d_H)$ satisfies the following condition: for every $\varepsilon > 0$ there exist constants c_3 and c_4 such that

$$\int_{\varepsilon^2/2^8}^{\varepsilon\sqrt{2}} \sqrt{J_{[]}(u/c_3, \mathcal{F}_n, d_H)} du \leq c_4 \sqrt{n} \varepsilon^2$$

In essence these conditions (specifically, conditions C2 and C3) balance the size of the sieve $\mathcal{F}_n$ and the prior probability of the sieve components $\Pi(\mathcal{F}_n)$: the integrated bracketing entropy of $\mathcal{F}_n$ has to grow slower than linearly with $\sqrt{n}$, while the prior probability assigned outside $\mathcal{F}_n$ has to decrease exponentially with n. Weaker conditions using covering numbers, rather than bracketing numbers, are available from Ghosal, Ghosh, and Ramamoorthi (1999). However, since the bracketing number bounds the covering number Kosorok (2008, p. 160-163), the fundamental idea of the balancing between the size of the model and the prior probability assigned outside the sieve remains the same.

In the multivariate case, the metric entropy of the sieve components may quickly increase with increasing dimensions. Wu and Ghosal (2010) and Canale and Dunson (2011) argue that this substantially restricts the set of useable sieves in the multivariate setting: if the sieve is too small, this will restrict the prior; if it is too large, this may lead to an exploding entropy. This makes it difficult to provide general results on posterior consistency in multivariate settings. In what follows we focus on the case of the Bernstein-Gaussian prior.

4.2. **Hellinger consistency of Bernstein-Gaussian posterior**

We build on the results of Petrone and Wasserman (2002) by constructing a suitable sieve in the space of multivariate densities and showing that conditions C2-C3 hold for it. We start by rewriting our mixed density (2.4) using the Bernstein-Gaussian setting as follows:

$$f(\theta,\phi) = \frac{1}{\sqrt{2\pi}} \sum_{j_1=1}^{k_1} \cdots \sum_{j_d=1}^{k_d} w_{\mathbf{k},P_c}(\mathbf{j}) \prod_{s=1}^{d} \beta\left[F(x;\theta_{ms},\phi_{ms}); j_s, k_s - j_s + 1\right] \frac{1}{\sigma_s} e^{-\frac{(x_s-\nu_s)^2}{2\sigma_s^2}}$$

where, as before, θ contains $\theta_c = w_{\mathbf{k},P_c}$ and $\theta_{ms} = \{\upsilon_s, \sigma_s^2\}, s = 1, ..., d$, while ϕ contains only $\phi_c = \mathbf{k} = (k_1, \ldots, k_d)$ since ϕ_{ms} is empty. The advantage of the Bernstein copula based density is that the weights $w_{\mathbf{k},P_c}$ come from the simplex $\mathsf{s}_{\mathbf{k}}$, and the density can be viewed as an infinite mixture of multivariate parametric densities, as $\min\{\mathbf{k}\} \to \infty$, similar to the stick breaking representation of Sethuraman (1994).

Specifically, let $1 \leq k' \leq \prod_{j=1}^{d} k_j \equiv K'$ index the set $\{j_1, \ldots, j_d\}_{j_i=1}^{k_i}, i = 1, \ldots, d$, and let $\pi(k')$ stand for $w_{\mathbf{k},P_c}$. Then, $f(\theta,\phi)$ can be written as follows

$$f(\theta,\phi) = \sum_{k'=1}^{K'} \pi(k') g_{k'}(\theta_m, \phi_m),$$

where

$$g_{k'}(\theta_m,\phi_m) = \prod_{s=1}^{d} \beta\left[F(x;\theta_{ms},\phi_{ms}); j_s, k_s - j_s + 1\right] \frac{1}{\sigma_s} e^{-\frac{(x_s-\nu_s)^2}{2\sigma_s^2}}$$

and $K' \to \infty$ as $\min\{\mathbf{k}\} \to \infty$. We can now follow the standard proof based on bracketing entropy bounds for a simplex (Petrone and Wasserman, 2002; Genovese and Wasserman, 2000). Let $\mathcal{B}_{k'}$ denote the class of Bernstein densities of order k', where the order is indexed by k' instead of $\mathbf{k} = \{k_1, \ldots, k_d\}$. In our case, the role of the sieve $\mathcal{F}_n$ is played by the set of Bernstein densities of order K'_n or lower, that is, $\mathcal{F}_n = \cup_{k'=1}^{K'_n} \mathcal{B}_{k'}$. (The subscript n distinguishes a sample specific value of an index, e.g., K'_n, k_{jn}, from a generic value, e.g., K', k_j.)

Theorem 3. *Suppose that there exists $K'_n \to \infty$ such that $K'_n = o(n)$ and such that $\sum_{k'=K'_n}^{\infty} \pi(k') \le c_1 \exp(-nc_2)$ for some $c_1 > 0$ and $c_2 > 0$. Then under condition C1, the posterior from the Bernstein-Gaussian prior is Hellinger consistent at f_0.*

Proof of Theorem 3: The proof follows quite closely the arguments of Theorem 3 of Petrone and Wasserman (2002). With condition C1 holding, it remains to show that conditions C2 and C3 hold. First, by assumption, $\Pi(\mathcal{F}_n^c) = \sum_{k'=K'_n}^{\infty} \pi(k') < c_1 \exp(-nc_2)$. So, condition C2 holds. Then, by Theorem 2 of Genovese and Wasserman (2000), there exists a constant $c > 0$ such that, for all $0 < \varepsilon \le 1$,

$$N_{[]}(\varepsilon, \mathcal{B}_{k'}, d_H) \le (c/\varepsilon)^{k'}$$

Thus, $N_{[]}(\varepsilon, \mathcal{F}_n, d_H) \le \sum_{k'=1}^{K'_n} N_{[]}(\varepsilon, \mathcal{B}_{k'}, d_H) \le K'_n (c/\varepsilon)^{K'_n} \le (2c/\varepsilon)^{K'_n}$, which, after integrating and noting that $K'_n = o(n)$, gives condition C3 for all large n. $\square$

Clearly the index change is done here only for convenience of the proof. The rate assumption on the bound for index k' implies an assumption on $\mathbf{k}$. For example, if the number of grid points k_{jn} is $o(\sqrt[d]{n})$ in each of the d dimensions, then $K'_n = \prod_{j=1}^{d} k_{jn} = o(n)$. This point is used in the following corollary.

Corollary 4. *If $\min\{\mathbf{k}\} \to \infty$ so that $k_j = o(\sqrt[d]{n}), j = 1, ..., d$, and the other assumptions of Theorem 3 hold, then the posterior from the Bernstein-Gaussian prior is Hellinger consistent at f_0.*

5. Monte Carlo Experiment

In this Section we perform a Monte Carlo experiment, comparing the accuracy of multivariate density estimation between our Bernstein-Gaussian mixture and a multivariate Gaussian mixture, which is arguably the default option in applications. Our data generating process (DGP) was inspired by the univariate multimodal densities of Marron and Wand (1992). The density from which our artificial data were generated is a κ-mixture of multivariate skew-Normal density kernels (Azzalini and Capitanio, 1999). A d-dimensional random variable x is said to have a multivariate skew-Normal distribution, denoted by $x \sim SN_d(\Omega, \eta)$, if it is continuous with density

$$f_d^{SN}(x) = 2\xi_d(x;\Omega)\Xi(\eta' x) \tag{5.1}$$

where $\xi_d(x;\Omega)$ is the d-dimensional normal density with zero mean and correlation matrix Ω, $\Xi(\cdot)$ is the univariate $N(0,1)$ distribution function, and η is a d-dimensional vector, referred to as the shape parameter. The location parameter $f_d^{SN}(x)$ is treated separately. When $\eta = 0$, (5.1) reduces to the $N_d(0,\Omega)$ density. We used the following result (Proposition 1 in Azzalini and Capitanio, 1999) for random number generation from $SN_d(\Omega,\eta)$. If

$$\begin{pmatrix} z_0 \\ z \end{pmatrix} \sim N_{d+1}\left(0,\Omega^*\right), \quad \Omega^* = \begin{pmatrix} 1 & \delta^T \\ \delta & \Omega \end{pmatrix} \tag{5.2}$$

where $z_0 \in \mathbb{R}$, $z \in \mathbb{R}^d$, and Ω^* is a correlation matrix, then

$$x = \begin{cases} z & \text{if } z_0 > 0 \\ -z & \text{otherwise} \end{cases}$$

is distributed $SN_d(\Omega,\eta)$, where $\eta = \left(1 - \delta^T\Omega^{-1}\delta\right)^{-1/2}\Omega^{-1}\delta$. Correspondingly, for a given shape parameter η, δ can be derived from η for use in (5.2) as $\delta = \left(1 + \eta^T\Omega\eta\right)^{-1/2}\Omega\eta$. Moreover, for the scaling matrix A, it holds that

$$A^T x \sim SN_d(A^T\Omega A, A^{-1}\eta)$$

(Azzalini and Capitanio, 1999, Proposition 3).

We will first describe the DGP for $d = 2$ and then elaborate on it as generalized to higher dimensions. We set the number of DGP kernels as $\kappa = 5$ with equal weights, and used the first kernel, multivariate Normal with $-1/d$ correlations in Ω but without skewness ($\eta = 0$), centered at zero as the anchor for the DGP. In the remaining kernels, we specified the correlations in Ω as $1/\kappa$, the shape vector $\eta = 5$, and the scale matrix $A = diag(a_s)$, $a_1 = 3/2$, $a_s = 4/5$ for $s = 2,\ldots,d$. In order to achieve kernel locations evenly spread in space with distinct multimodal features but with tails overlapping with the first kernel to avoid complete separation, we divided 2π radians (360°) into $(\kappa - 1)$ equal angles and shifted out the kernel center locations $\overline{x}$ under these angles until the ratio of the center kernel to the new kernel at $\overline{x}$ fell to 10^{-3}. For higher dimensions $d > 2$ we kept the same DGP parameters, and shifted out every odd dimension under the same angle as the first dimension, and every even dimension under the same angle as the second dimension. For $d = 2$, this DGP generated the following joint density.

FIGURE 1. Simulation DGP Density for $d = 2$

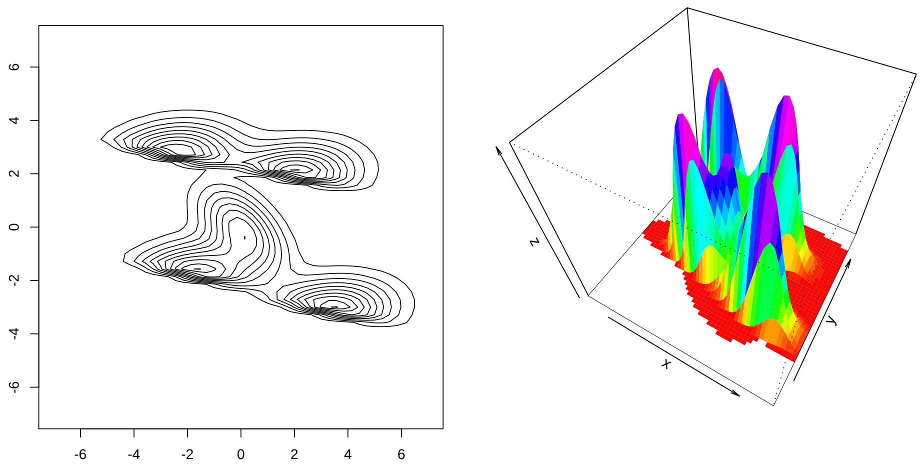
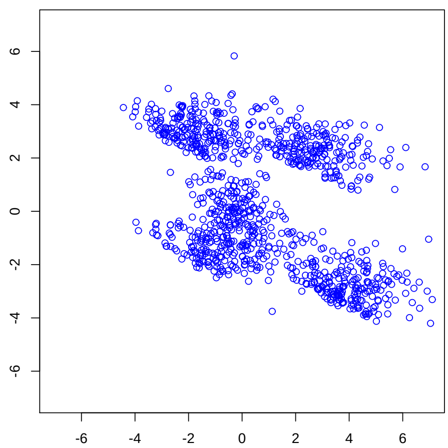


FIGURE 2. Scatterplot of a simulated sample for $d = 2$



For this DGP, we compared the precision of the density estimate between our Bernstein-Gaussian mixture model (BG) and the multivariate Normal mixture model (MVN). In the BG implementation, we used the Neal (2000) Algorithm 8 for the marginals and a multivariate version of the Petrone (1999a) algorithm for the copula part, compiled using Intel fortran 95 on a 2.8 GHz Unix machine. For the MVN implementation, we used the DPpackage in R (Jara, Hanson, Quintana, Müller, and Rosner, 2011) on a 3.0 GHz Windows machine. In each case we drew 10,000 MCMC steps with a 1,000 burnin section. Each case took in the order of hours to run. Past the burnin section the Markov chain draws were very stable regardless of the length of additional output and hence we have not performed an explicit comparison per unit of run time; what matters is the precision of each estimate provided that each can be obtained relatively quickly. For each model, we computed the mean absolute deviation (MAD) measure as the absolute value difference between the DGP density and the estimated density evaluated at each data point and summed over the data. The results, averaged over 10 different samples from the DGP, for 2, 3 and 4 dimensional densities, are presented in Table 1 below.

As the results indicate, the BG mixture model improved on the MVN mixture model by reducing the MAD measure to less than a half, with the rate of improvement increasing with higher dimensionality of the underlying density. This outcome can be visually corroborated by examining the plots of the two dimensional MVN estimate and BG estimate in Figures 3 and 4 below. The DGP is based on four clusters of skewed multivariate Normal density kernels centered around a Normal kernel. The MVN estimate in Figure 3 concentrates very few multivariate Normal kernels in the cluster locations and appears to misrepresent the degree of skewness in the DGP. Moreover, the relative weights of the estimated clusters apparent from the 3D density plot does not reflect well the DGP cluster weights. In contrast, the contour plot of the BG estimate in Figure 4 appears to capture more accurately the degree of skewness present in at least three of the estimated clusters. Their height in the 3D density plot seems to reflect their relative equal weight similarly to the DGP. Furthermore, the BG nonparametrically estimated copula function is given in Figure 5.

6. Conclusions

In this paper, we propose a factorization scheme for Bayesian nonparametric mixture models based on modeling separately the marginals as univariate infinite mixtures, linked by a nonparametric random copula function. We show that this scheme leads to an improvement in the precision of a density estimate in finite samples. Our approach should thus be useful for practical applications. We show weak posterior consistency of the copula-based mixing scheme for general kernel types under high-level conditions, and strong posterior consistency for a model with univariate Gaussian marginal mixing kernels and random Bernstein polynomial for the copula, under the Dirichlet process prior.

Table 1. Mean Absolute Deviation Comparison of BG vs MVN mixture

Dimensions	MAD BG	MAD MVN	Ratio
2	15.007	35.788	0.419
	(9.392)	(43.015)	
3	17.270	46.553	0.371
	(10.171)	(58.701)	
4	21.149	59.996	0.353
	(10.405)	(71.956)	

Standard deviations over 10 simulated data sets are given in brackets.

FIGURE 3. MVN Mixture Density Estimate for $d = 2$

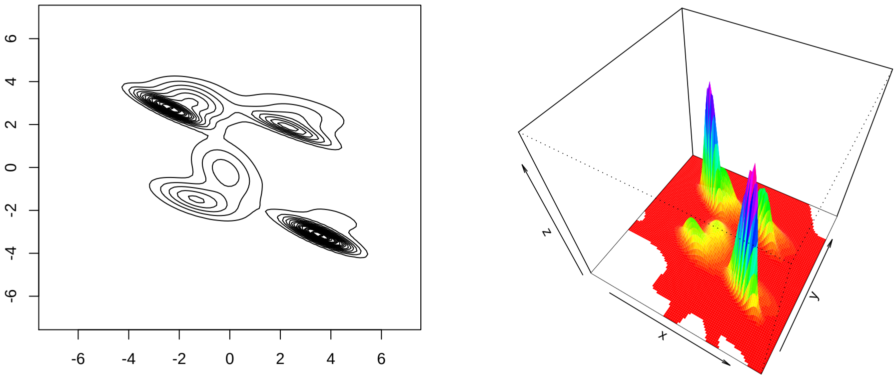


FIGURE 4. BG Copula Mixture Density Estimate for $d = 2$

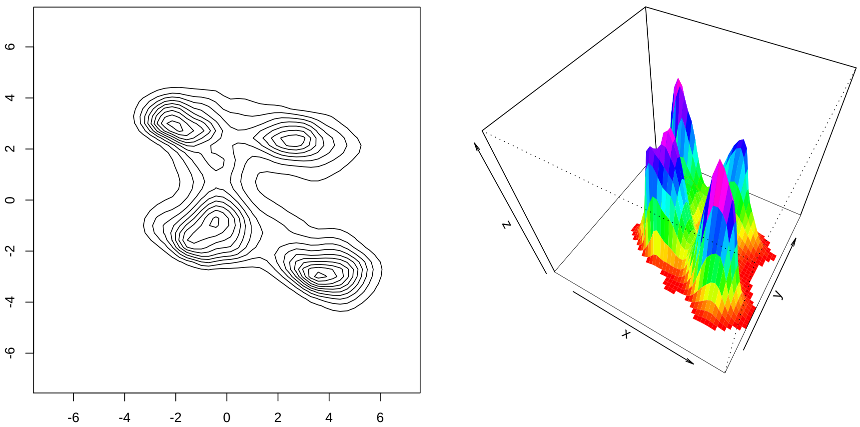
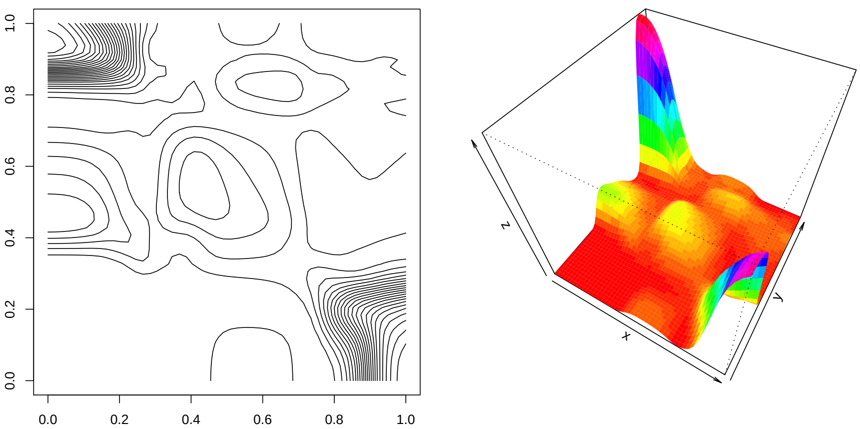


FIGURE 5. BG Nonparametric Copula Estimate for $d = 2$



7. Technical Appendix

Here we restate WG Theorem 1, and Lemmas 2 and 3 for the sake of completeness since we refer to their assumptions throughout the text.

WG Theorem 1:

Theorem 5. *Let f_0 be the true density with the composition as in (2.1), μ and Π be the priors for ϕ and P, and Π^* be the prior induced by μ and Π on $\mathcal{D}(\mathcal{X})$. If for any $\varepsilon > 0$, there exists P_ε, ϕ_ε, $A \subset \Phi$ with $\mu(A) > 0$ and $\mathcal{W} \subset \mathcal{M}(\Theta)$ with $\Pi(\mathcal{W}) > 0$, such that*
A1. $\int f_0 \log\left(\frac{f_0(x)}{f_{P_\varepsilon,\phi_\varepsilon}(x)}\right) < \varepsilon$;
A2. $\int f_0 \log\left(\frac{f_{P_\varepsilon,\phi_\varepsilon}(x)}{f_{P_\varepsilon,\phi}(x)}\right) < \varepsilon$ *for every* $P \in \mathcal{W}$,
A3. $\int f_0 \log\left(\frac{f_{P_\varepsilon,\phi}(x)}{f_{P,\phi}(x)}\right) < \varepsilon$ *for every* $P \in \mathcal{W}$, $\phi \in A$,
then $f_0 \in KL(\Pi^*)$.

The proof given in WG is based on A1-A3 showing that $\int f_0 \log\left(\frac{f_0(x)}{f_{P,\phi}(x)}\right) < 3\varepsilon$, and hence

$$\Pi^*\left\{f : f \in \mathcal{K}_{3\varepsilon}(f_0)\right\} \geq \Pi^*\left\{f_{P,\phi} : P \in \mathcal{W}, \phi \in A\right\} = (\Pi \times \mu)(\mathcal{W} \times A) > 0.$$

WG Lemma 2:

Lemma 6. *Let f_0, Π, μ, and Π^* be the same as in Theorem 1. If for any $\varepsilon > 0$, there exist P_ε, a set D containing supp(P_ε), and $\phi_\varepsilon \in$ supp(μ) such that A1 holds and the kernel function K satisfies*
A4. for any given x and θ, the map $\phi \mapsto K(x;\theta,\phi)$ is continuous on the interior of the support of μ;
A5. $\int_{\mathcal{X}} \left\{\left|\log \frac{\sup_{\theta\in D} K(x;\theta,\phi_\varepsilon)}{\inf_{\theta\in D} K(x;\theta,\phi)}\right| + \left|\log \frac{\sup_{\theta\in D} K(x;\theta,\phi)}{\inf_{\theta\in D} K(x;\theta,\phi_\varepsilon)}\right|\right\} f_0(x)dx < \infty$ *for every* $\phi \in N(\phi_\varepsilon)$, *where $N(\phi_\varepsilon)$ is an open neighborhood of ϕ_ε;*
A6. for any given $x \in \mathcal{X}$, $\theta \in D$ and $\phi \in N(\phi_\varepsilon)$, there exists $g(x,\theta)$ such that $g(x,\theta) \geq K(x;\theta,\phi)$, and $\int g(x,\theta)dP_\varepsilon(\theta) < \infty$;
then there exists a set $A \subset \Phi$ such that A2 holds.

WG Lemma 3, with A9 from Wu and Ghosal (2009):

Lemma 7. *Let f_0, Π, μ, and Π^* be the same as in Theorem 1. If for any $\varepsilon > 0$, there exist $P_\varepsilon \in$ supp(Π), $\phi_\varepsilon \in$ supp(μ), and $A \subset \Phi$ with $\mu(A) > 0$ such that Conditions A1 and A2 hold and for some closed $D \supset$ supp(P_ε), the kernel function K and prior Π satisfy*
A7. for any $\phi \in A$, $\log \frac{f_{P_\varepsilon,\phi}(x)}{\inf_{\theta\in D} K(x;\theta,\phi)} f_0(x)dx < \infty$;
A8. $c \equiv \inf_{x\in C} \inf_{\theta\in D} K(x;\theta,\phi) > 0$, *for any compact* $C \subset \mathcal{X}$;
A9. for any given $\phi \in A$ and compact $C \subset \mathcal{X}$, there exists E containing D in its interior such that the family of maps $\{\theta \mapsto K(x;\theta,\phi), x \in C\}$ is uniformly equicontinuous on D;
then there exists $\mathcal{W} \subset \mathcal{M}(\Theta)$ such that Condition A3 holds and $\Pi(\mathcal{W}) > 0$.

References

Antoniak, C. E. (1974): "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 1, 1152–1174.

Ausin, M., and H. Lopes (2010): "Time-varying joint distribution through copulas," *Computational Statistics and Data Analysis*, 54(11), 2383–2399, cited By (since 1996) 8.

Barron, A., M. J. Schervish, and L. Wasserman (1999): "The Consistency of Posterior Distributions in Nonparametric Problems," *The Annals of Statistics*, 27(2), 536–561.

Burda, M., M. C. Harding, and J. A. Hausman (2008): "A Bayesian Mixed Logit-Probit Model for Multinomial Choice," *Journal of Econometrics*, 147(2), 232–246.

Canale, A., and D. B. Dunson (2011): "Bayesian multivariate mixed-scale density estimation," .

Chen, X., Y. Fan, and V. Tsyrennikov (2006): "Efficient Estimation of Semiparametric Multivariate Copula Models," *Journal of the American Statistical Association*, 101, 1228–1240.

Chib, S., and B. Hamilton (2002): "Semiparametric bayes analysis of longitudinal data treatment models," *Journal of Econometrics*, 110, 67–89.

Conley, T., C. Hansen, R. McCulloch, and P. Rossi (2008): "A Semi-Parametric Bayesian Approach to the Instrumental Variable Problem," *Journal of Econometrics*, 144, 276–305.

Diaconis, P., and D. Freedman (1986): "On the Consistency of Bayes Estimates," *Annals of Statistics*, 14(1), 1–26.

Doob, J. L. (1949): "Application of the theory of martingales," *Colloque International Centre Nat. Rech. Sci., Paris*, p. 2228.

Escobar, M. D., and M. West (1995): "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90(430), 577–588.

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

Fiebig, D. G., R. Kohn, and D. S. Leslie (2009): "Nonparametric estimation of the distribution function in contingent valuation models," *Bayesian Analysis*, 4(3), 573–597.

Freedman, D. (1963): "On the asymptotic behavior of Bayes estimates in the discrete Markov processes," *Annals of Mathematical Statistics*, 34, 1386–1403.

Fuentes, M., J. Henry, and B. Reich (2012): "Nonparametric spatial models for extremes: application to extreme temperature data," *Extremes*, pp. 1–27, cited By (since 1996) 2; Article in Press.

Genovese, C., and L. Wasserman (2000): "Rates of convergence for the Gaussian mixture sieve," *The Annals of Statistics*, 28(4), 1105–1127.

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

Ghosal, S. (2010): "The Dirichlet process, related priors and posterior asymptotics," in *Bayesian Nonparametrics*, ed. by N. L. Hjort, C. Holmes, P. Mueller, and S. G. Walker. Cambridge University Press.

Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi (1999): "Posterior Consistency of Dirichlet Mixtures in Density Estimation," *The Annals of Statistics*, 27(1), 143–158.

Ghosal, S., and A. van der Vaart (2001): "Entropies and rates of convergence for Bayes and maximum likelihood estimation for mixture of normal densities," *Annals of Statistics*, 29(5), 1233–1263.

——— (2007): "Posterior convergence rates of Dirichlet mixtures at smooth densities," *Annals of Statistics*, 35(2), 697–723.

——— (2011): "Theory of Nonparametric Bayesian Inference," Cambridge University Press.

Giordani, P., X. Mun, and R. Kohn (2012): "Efficient estimation of covariance matrices using posterior mode multiple shrinkage," *Journal of Financial Econometrics*, 11(1), 154–192, cited By (since 1996) 0.

Hirano, K. (2002): "Semiparametric bayesian inference in autoregressive panel data models," *Econometrica*, 70, 781–799.

Jara, A., T. Hanson, F. Quintana, P. Müller, and G. Rosner (2011): "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *Journal of Statistical Software*, 40(5), 1–30.

Jensen, M., and J. M. Maheu (2010): "Bayesian semiparametric stochastic volatility modeling," *Journal of Econometrics*, 157(2), 306–316.

Kim, J. G., U. Menzefricke, and F. Feinberg (2004): "Assessing Heterogeneity in Discrete Choice Models Using a Dirichlet Process Prior," *Review of Marketing Science*, 2(1), 1–39.

Kim, Y., and J. Lee (2001): "On posterior consistency of survival models," *Annals of Statistics*, 29, 666–686.

Kolmogorov, A., and V. Tikhomirov (1961): "Epsilon-entropy and epsilon-capacity of sets in function spaces," *AMS Translations Series 2 [Translated from Russian (1959) Uspekhi Mat.Nauk 14, 3-86]*, 17, 277–364.

Kosorok, M. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Springer Series in Statistics. Springer.

Kruijer, W., and A. van der Vaart (2008): "Posterior convergence rates for Dirichlet mixtures of Beta densities," *Journal of Statistical Planning and Inference*, 138, 1981–1992.

Lo, A. Y. (1984): "On a class of Bayesian nonparametric estimates I: Density estimation," *Annals of Statistics*, 12, 351–357.

Lorentz, G. (1986): *Bernstein Polynomials*. University of Toronto Press.

Neal, R. M. (2000): "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9(2), 249–265.

Panchenko, V., and A. Prokhorov (2012): "Efficient Estimation of Parameters in Marginals," *Concordia University Working Paper*.

Petrone, S. (1999a): "Bayesian density estimation using bernstein polynomials," *Canadian Journal of Statistics*, 27(1), 105–126.

——— (1999b): "Random Bernstein Polynomials," *Scandinavian Journal of Statistics*, 26(3), 373–393.

Petrone, S., and P. Veronese (2010): "Feller operators and mixture priors in Bayesian nonparametrics," *Statistica Sinica*, 20, 379–404.

Petrone, S., and L. Wasserman (2002): "Consistency of Bernstein Polynomial Posteriors," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(1), 79–100.

PITT, M., D. CHAN, AND R. KOHN (2006): "Efficient Bayesian inference for Gaussian copula regression models," *Biometrika*, 93(3), 537–554, cited By (since 1996) 37.

REY, M., AND V. ROTH (2012): "Copula mixture model for dependency-seeking clustering," vol. 1, pp. 927–934. cited By (since 1996) 0.

RODRIGUEZ, A., D. DUNSON, AND A. GELFAND (2010): "Latent stick-breaking processes," *Journal of the American Statistical Association*, 105(490), 647–659, cited By (since 1996) 4.

SANCETTA, A., AND S. SATCHELL (2004): "The Bernstein Copula And Its Applications To Modeling And Approximations Of Multivariate Distributions," *Econometric Theory*, 20(03), 535–562.

SCHWARTZ, L. (1965): "On Bayesian procedures," *Probability Theory and Related Fields (Z. Wahrscheinlichkeitstheorie)*, 4, 10–26.

SETHURAMAN, J. (1994): "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

SHEN, W., AND S. GHOSAL (2011): "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures," .

SILVA, R., AND R. B. GRAMACY (2009): "MCMC Methods for Bayesian Mixtures of Copulas," in *Proceedings of the* 12^{th} *International Conference on Artificial Intelligence and Statistics 2009, Clearwater Beach, Florida, USA*, pp. 512–519.

SILVA, R. D. S., AND H. F. LOPES (2008): "Copula, marginal distributions and model selection: a Bayesian note," *Statistics and Computing*, 18, 313–320.

TENBUSCH, A. (1994): "Two-dimensional Bernstein polynomial density estimators," *Metrika*, 41(1), 233–253, Metrika.

TOKDAR, S. T. (2006): "Posterior Consistency of Dirichlet Location-scale Mixture of Normals in Density Estimation and Regression," *Sankhya*, 68, 90–110.

VITALE, R. (1975): "A Bernstein polynomial approach to density function estimation," in *Statistical inference and related topics*, ed. by M. Puri.

WASSERMAN, L. (1998): "Asymptotic properties of nonparametric Bayesian procedures," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, ed. by D. Dey, P. Mueller, and D. Sinha. Springer, New York.

WU, Y., AND S. GHOSAL (2008): "Kullback Leibler property of kernel mixture priors in Bayesian density estimation," *Electronic Journal of Statistics*, 2, 298–331.

——— (2010): "The -consistency of Dirichlet mixtures in multivariate Bayesian density estimation," *Journal of Multivariate Analysis*, 101(10), 2411–2419.

ZHENG, Y. (2011): "Shape restriction of the multi-dimensional Bernstein prior for density functions," *Statistics and Probability Letters*, 81(6), 647–651.

ZHENG, Y., J. ZHU, AND A. ROY (2010): "Nonparametric Bayesian inference for the spectral density function of a random field," *Biometrika*, 97, 238–245.