Variational Gaussian Copula Inference

Shaobo Han   **Xuejun Liao**   **David B. Dunson**† **Lawrence Carin**
Department of ECE, Department of Statistical Science†, Duke University, Durham, NC 27708, USA

Abstract

We utilize copulas to constitute a unified framework for constructing and optimizing variational proposals in hierarchical Bayesian models. For models with continuous and non-Gaussian hidden variables, we propose a semiparametric and automated variational Gaussian copula approach, in which the parametric Gaussian copula family is able to preserve multivariate posterior dependence, and the nonparametric transformations based on Bernstein polynomials provide ample flexibility in characterizing the univariate marginal posteriors.

1   Introduction

A crucial component of Bayesian inference is approximating the posterior distribution, which represents the current state of knowledge about the latent variables $\boldsymbol{x}$ after data $\boldsymbol{y}$ have been observed. When intractable integrals are involved, variational inference methods find an approximation $q(\boldsymbol{x})$ to the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}\{q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})\} = \int q(\boldsymbol{x})\log\left[q(\boldsymbol{x})/p(\boldsymbol{x}|\boldsymbol{y})\right]d\boldsymbol{x}$, providing a lower bound for the marginal likelihood.

To make inference tractable, mean-field variational Bayes (MFVB) methods (Jordan et al., 1999; Wainwright and Jordan, 2008) assume $q(\boldsymbol{x})$ is factorized over a certain partition of the latent variables $\boldsymbol{x} \equiv [\boldsymbol{x}_1, \dots, \boldsymbol{x}_J]$, $q_{\mathrm{VB}}(\boldsymbol{x}) = \prod_j q_{\mathrm{VB}}(\boldsymbol{x}_j)$, with marginal densities $q_{\mathrm{VB}}(\boldsymbol{x}_j)$ in free-form and correlations between partitions neglected. The structured mean-field approaches (Saul and Jordan, 1996; Hoffman and Blei, 2015) preserve partial correlations and apply only to models with readily identified substruc-

tures. The variational Gaussian (VG) approximation (Barber and Bishop, 1998; Opper and Archambeau, 2009) allows incorporation of correlations by postulating a multivariate Gaussian parametric form $q_{\mathrm{VG}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The VG approximation, with continuous margins of real variables, are not suitable for variables that are inherently positive or constrained, skewed, or heavy tailed. For multi-modal posteriors, a mixture of MFVB (Jaakkola and Jordan, 1998) or a mixture of uniformly-weighted Gaussians (Gershman et al., 2012) may be employed, which usually requires a further lower bound on the average over the logarithm of the mixture distribution.

To address the limitations of current variational methods in failing to simultaneously characterize the posterior dependencies among latent variables while allowing skewness, multimodality, and other characteristics, we propose a new variational copula framework. Our approach decouples the overall inference task into two subtasks: (i) inference of the copula function, which captures the multivariate posterior dependencies; (ii) inference of a set of univariate margins, which are allowed to take essentially any form. Motivated by the work on automated (black-box) variational inference (Ranganath et al., 2014; Mnih and Gregor, 2014; Titsias and Lázaro-Gredilla, 2014; Nguyen and Bonilla, 2014; Kingma and Welling, 2014), we present a stochastic optimization algorithm for *generic* hierarchical Bayesian models with continuous variables, which (i) requires minimal model-specific derivations, (ii) reproduces peculiarities of the true marginal posteriors, and (iii) identifies interpretable dependency structure among latent variables.

Using copulas to improve approximate Bayesian inference is a natural idea that has also been explored recently in other contexts (Li et al., 2015; Ferkingstad and Rue, 2015). Independently from our work, Tran et al. (2015) presented a copula augmented variational method with fixed-form marginals, and utilizes regular vines to decompose the multivariate dependency structure into bivariate copulas and a nest of trees. Our method provides complementary perspectives on nonparametric treatment of univariate marginals.

2 Variational Copula Inference Framework

Sklar's theorem (Sklar, 1959) ensures that any multivariate joint distribution Q can be written in terms of univariate marginal distributions $F_j(x) = P(X_j \leq x)$, $j = 1, \ldots, p$ and a copula which describes the dependence structures between variables, such that

$$Q(x_1, \ldots, x_p) = C[F_1(x_1), \ldots, F_p(x_p)]. \quad (1)$$

Conversely, if C is a copula and $\{F_j\}_{j=1:p}$ are distribution functions, then the function Q defined by (1) is a p-dimensional joint distribution function with marginal distributions $F_1, F_2, \ldots, F_p$, owing to the marginally closed property (Song, 2000). Assuming $Q(x_1, ..., x_p)$ has p-order partial derivatives, the joint probability density function (PDF) is $q(x_1, \ldots, x_p) = c_{\boldsymbol{\Theta}}[F_1(x_1), \ldots, F_p(x_p)] \prod_{j=1}^{p} f_j(x_j)$, where $f_j(x_j)$ is the PDF of the jth variable and it is related to the corresponding cumulative distribution function (CDF) by $F_j(x_j) = \int_{-\infty}^{x} f_j(t)dt$, $c_{\boldsymbol{\Theta}}$ is the copula density with parameter $\boldsymbol{\Theta}$.

Sklar's theorem allows separation of the marginal distributions $F_j(x_j)$ from the dependence structure, which is appropriately expressed in the copula function C. As a modeling tool, the specified copula function and margins can be directly fitted to the observed data $\boldsymbol{y}$ (Liu et al., 2009; Wauthier and Jordan, 2010; Lopez-Paz et al., 2013) with their parameters optimized via Bayesian or maximum likelihood estimators (see Smith (2013) and the references therein). In contrast, our goal is to use a copula as an *inference engine* for full posterior approximation. All the unknowns (variables/parameters) in the user-specified hierarchical model are encapsulated into a vector $\boldsymbol{x}$, and the optimal variational approximation $q_{\mathrm{VC}}(\boldsymbol{x})$ to the true posterior $p(\boldsymbol{x}|\boldsymbol{y})$ is found under the Sklar's representation. This approach provides users with full modeling freedom and does not require conditional conjugacy between latent variables; thus the approach is applicable to general models. Within some tractable copula family $C \in \mathcal{C}$, and assuming $F(\cdot)$ and $C(\cdot)$ to be differentiable, we construct the variational proposal as $q_{\mathrm{C}}(\boldsymbol{x}) = c(\boldsymbol{u}) \prod_{j=1}^{p} f_j(x_j)$, where $\boldsymbol{u} = F(\boldsymbol{x}) = [F_1(x_1), \ldots, F_p(x_p)]$, such that the approximation satisfies

$$q_{\mathrm{C}}^{\star}(\boldsymbol{x}) = \underset{q_{\mathrm{C}}(\boldsymbol{x})}{\arg\min} \, \mathrm{KL}\{q_{\mathrm{C}}(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})\}$$
$$= \underset{q_{\mathrm{C}}(\boldsymbol{x})}{\arg\min} \, \mathrm{KL}\{q_{\mathrm{C}}(\boldsymbol{x})||p(\boldsymbol{x})\} - \mathbb{E}_{q_{\mathrm{C}}(\boldsymbol{x})}[\ln p(\boldsymbol{y}|\boldsymbol{x})],$$

where $p(\boldsymbol{y}|\boldsymbol{x})$ is the likelihood and $p(\boldsymbol{x})$ is the prior. Letting the true posterior $p(\boldsymbol{x}|\boldsymbol{y})$ in Sklar's representation be $p(\boldsymbol{x}|\boldsymbol{y}) = c^{\star}(\boldsymbol{v}) \prod_{j} f_j^{\star}(x_j)$, where $\boldsymbol{v} =$

$[F_1^{\star}(x_1), \ldots, F_p^{\star}(x_p)]$, $c^{\star}(\boldsymbol{v})$ and $\{f_j^{\star}(x_j)\}_{j=1:p}$ are the true underlying copula density and marginal posterior densities, respectively, the KL divergence decomposes into additive terms (derivations are provided in Supplementary Material),

$$\mathrm{KL}\{q_{\mathrm{C}}(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})\} = \mathrm{KL}\{c(\boldsymbol{u})||c^{\star}[F^{\star}(F^{-1}(\boldsymbol{u}))]\}$$
$$+ \sum_{j} \mathrm{KL}\{f_j(x_j)||f_j^{\star}(x_j)\}. \quad (2)$$

Classical methods, such as MFVB and the VG approximation are special cases of the proposed VC inference framework. We next compare their KL divergence under Sklar's representation and offer a reinterpretation of them under the proposed framework.

2.1 Special Case 1: Mean-field VB

The mean-field proposal corresponds to the independence copula $C_{\Pi}(\boldsymbol{u}) = \prod_{j=1}^{J} u_j$ with free-form marginal densities $f_j(\boldsymbol{x}_j)$. Given $c_{\Pi}(\boldsymbol{u}) = 1$ we have $q_{\Pi}(\boldsymbol{x}) = c_{\Pi}(\boldsymbol{u}) \prod_{j} f_j(\boldsymbol{x}_j) = \prod_{j} f_j(\boldsymbol{x}_j) = q_{\mathrm{VB}}(\boldsymbol{x})$. If MFVB is not fully factorized, i.e. $J < p$, the independence copula is the only copula satisfying the marginal closed property, according to the impossibility theorem (Nelsen, 2007). MFVB assumes an independence copula and only optimizes the free-form margins,

$$\mathrm{KL}\{q_{\mathrm{VB}}(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})\} = \mathrm{KL}\{c_{\Pi}(\boldsymbol{u})||c^{\star}[F^{\star}(F^{-1}(\boldsymbol{u}))]\}$$
$$+ \sum_{j} \mathrm{KL}\{f_j(x_j)||f_j^{\star}(x_j)\}. \quad (3)$$

The lowest achievable KL divergence in MFVB is $\mathrm{KL}\{q_{\mathrm{VB}}(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})\} = \mathrm{KL}\{c_{\Pi}(\boldsymbol{u})||c^{\star}(\boldsymbol{u})\}$, which is achieved when the true posterior marginals are found, i.e. $F_j \equiv F_j^{\star}, \forall j$, in which case the overall KL divergence is reduced to the KL divergence between the independence copula and the true copula. As is shown in (3), the objective function contains two terms, both involving marginal CDFs $\{F_j\}_{j=1:p}$. Since in general $c^{\star} \neq c_{\Pi}$, the optimal F minimizing the first term will not be equal to $F^{\star}$. Therefore, minimizing (3) will not lead to the correct marginals and this partially explains the reason why MFVB usually cannot find the true marginal posteriors in practice (e.g., variances can be severely underestimated (Neville et al., 2014)), even though it allows for free-form margins.

2.2 Special Case 2: VG Approximation

In fixed-form variational Bayes (Honkela et al., 2010), such as VG approximation, the multivariate Gaussian proposal $q_{\mathrm{VG}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written as $q_{\mathrm{VG}}(\boldsymbol{x}) = c_{\mathrm{G}}(\boldsymbol{u}|\boldsymbol{\Upsilon}) \prod_{j=1}^{p} \phi_j(x_j; \mu_j, \sigma_j^2)$. VG not only assumes the true copula function is a Gaussian copula (Song, 2000) with parameter $\boldsymbol{\Upsilon} = \boldsymbol{D}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{D}^{-1/2}$,

$\boldsymbol{D} = \text{diag}(\boldsymbol{\Sigma})$, but is also restricted to univariate Gaussian marginal densities $\{\phi_j(x_j; \mu_j, \sigma_j^2)\}_{j=1:p}$,

$$\text{KL}\{q_{\text{VG}}(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})\} = \text{KL}\{c_{\text{G}}(\boldsymbol{u})||c^\star[F^\star(\Phi^{-1}(\boldsymbol{u}))]\}$$
$$+ \sum_j \text{KL}\{\phi_j(x_j)||f_j^\star(x_j)\}. \quad (4)$$

We can see in (4) that if the margins are misspecified, even if the true underlying copula is a Gaussian copula, there could still be a discrepancy $\sum_j \text{KL}\{\phi_j(x_j)||f_j^\star(x_j)\}$ between margins, and $\text{KL}\{c_{\text{G}}(\boldsymbol{u})||c^\star[F^\star(\Phi^{-1}(\boldsymbol{u}))]\}$ is not zero.

Concerning analytical tractability and simplicity, in the sequel we concentrate on variational Gaussian copula (VGC) proposals constructed via Gaussian copula with continuous margins, i.e. $q_{\text{VGC}}(\boldsymbol{x}) = c_{\text{G}}(\boldsymbol{u}|\boldsymbol{\Upsilon}) \prod_{j=1}^p f_j(x_j)$, where $\boldsymbol{u} = [F_1(x_1), \ldots, F_p(x_p)]$. Our VGC method extends MFVB and VG, and improves upon both by allowing simultaneous updates of the Gaussian copula parameter $\boldsymbol{\Upsilon}$ and the adaptation of marginal densities $\{f_j(x_j)\}_{j=1:p}$. First, the univariate margins in VGC is not restricted to be Gaussian. Second, the Gaussian copula in VGC is more resistant to local optima than the independence copula assumed in MFVB and alleviates its variance underestimation pitfall, as is demonstrated in Section 6.3.

3 Variational Gaussian Copula Approximation

A Gaussian copula function with $p \times p$ correlation matrix $\boldsymbol{\Upsilon}$ is defined as $C_{\text{G}}(u_1, \ldots, u_p|\boldsymbol{\Upsilon}) = \Phi_p(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)|\boldsymbol{\Upsilon}) : [0,1]^p \to [0,1]$ where $\Phi(\cdot)$ is a shorthand notation of the CDF of $\mathcal{N}(0,1)$, and $\Phi_p(\cdot|\boldsymbol{\Upsilon})$ is the CDF of $N_p(\boldsymbol{0}, \boldsymbol{\Upsilon})$. The Gaussian copula density is

$$c_{\text{G}}(u_1, \ldots, u_p|\boldsymbol{\Upsilon}) = \frac{1}{\sqrt{|\boldsymbol{\Upsilon}|}} \exp\left\{-\frac{\boldsymbol{z}^T(\boldsymbol{\Upsilon}^{-1} - \boldsymbol{I}_p)\boldsymbol{z}}{2}\right\},$$

where $\boldsymbol{z} = [\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)]^T$.

In the proposed VGC approximation, the variational proposal $q_{\text{VGC}}(\boldsymbol{x})$ is constructed as a product of Gaussian copula density and continuous marginal densities. The evidence lower bound (ELBO) of VGC approximation is

$$\mathcal{L}_{\text{C}}[q_{\text{VGC}}(\boldsymbol{x})] = \int \left[c_{\text{G}}[F(\boldsymbol{x})] \times \prod_{j=1}^p f_j(x_j)\right] \ln p(\boldsymbol{y}, \boldsymbol{x}) d\boldsymbol{x}$$

$$+ H[c_{\text{G}}(\boldsymbol{u})] + \sum_{j=1}^p H[f_j(x_j)], \quad (5)$$

where $u_j = F_j(x_j)$, $H[f(x)] = -\int f(x) \ln f(x) dx$.

However, directly optimizing the ELBO in (5) w.r.t. the Gaussian copula parameter $\boldsymbol{\Upsilon}$ and the univariate marginals $\{f_j(x_j)\}_{j=1:p}$ often leads to a non-trivial variational calculus problem. For computational convenience, we present several equivalent proposal constructions based on Jacobian transformation and reparameterization.

3.1 Equivalent Variational Proposals

We incorporate auxiliary variables $\boldsymbol{z}$ by exploiting the latent variable representation of the Gaussian copula: $x_j = F_j^{-1}(u_j)$, $u_j = \Phi(z_j)$, $\boldsymbol{z} \sim N_p(\boldsymbol{0}, \boldsymbol{\Upsilon})$. Letting $g_j(\cdot) = F_j^{-1}(\Phi(\cdot))$ be bijective monotonic non-decreasing functions, $x_j = g_j(z_j)$, $\forall j$, the Jacobian transformation gives

$$q_{\text{VGC}}(\boldsymbol{x}) = \int \left[\prod_{j=1}^p \delta(x_j - g_j(z_j))\right] q_{\text{G}}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{\Upsilon}) d\boldsymbol{z}$$

$$= q_{\text{G}}(g^{-1}(\boldsymbol{x}); \boldsymbol{0}, \boldsymbol{\Upsilon}) \left[\prod_{j=1}^p \frac{d}{dx_j} g_j^{-1}(x_j)\right],$$

where $\delta(\cdot)$ is the Dirac delta function.

It is inconvenient to directly optimize the correlation matrix $\boldsymbol{\Upsilon}$ of interest, since $\boldsymbol{\Upsilon}$ is a positive semi-definite matrix with ones on the diagonal and off-diagonal elements between $[-1, 1]$. We adopt the parameter expansion (PX) technique (Liu et al., 1998; Liu and Wu, 1999), which has been applied in accelerating variational Bayes (Qi and Jaakkola, 2006) and the sampling of correlation matrix (Talhouk et al., 2012). Further considering $\widetilde{z}_j = t_j^{-1}(z_j) = \mu_j + \sigma_{jj} z_j$, $\widetilde{\boldsymbol{z}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \boldsymbol{D} \boldsymbol{\Upsilon} \boldsymbol{D}^T$, $\boldsymbol{D} = [\text{diag}(\sigma_{jj})]_{j=1:p}$, thus $x_j = g(z_j) = g(t(\widetilde{z}_j)) := h(\widetilde{z}_j)$, where $h_j(\cdot) = g_j \circ t_j(\cdot)$ are also bijective monotonic non-decreasing functions, the variational proposal is further written as

$$q_{\text{VGC}}(\boldsymbol{x}) = \int \left[\prod_{j=1}^p \delta(x_j - h_j(\widetilde{z}_j))\right] q_{\text{G}}(\widetilde{\boldsymbol{z}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\widetilde{\boldsymbol{z}}$$

$$= q_{\text{G}}(h^{-1}(\boldsymbol{x}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) \left[\prod_{j=1}^p \frac{d}{dx_j} h_j^{-1}(x_j)\right].$$

Given the transformations $\{h_j\}_{j=1:p}$, $q_{\text{G}}(\widetilde{\boldsymbol{z}}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be further reparameterized by the Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^T$ (Challis and Barber, 2013; Titsias and Lázaro-Gredilla, 2014), where $\boldsymbol{C}$ is a square lower triangular matrix. Table 1 summarizes four translatable representations of variational proposals.

3.2 VGC with Fixed-form Margins

The ELBO under Sklar's representation (5) is therefore translated into the Jacobian representation

Table 1: Equivalent Representations of Variational Gaussian Copula (VGC) Proposals

	Posterior Formulation	Optimization Space	
R0	Original	Multivariate (non-Gaussian) density $q(\boldsymbol{x})$	
R1	Sklar's Representation	Copula density $c_{\mathrm{G}}(\boldsymbol{u}|\boldsymbol{\Upsilon})$	Univariate marginals $\{f_j(x_j)\}_{j=1:p}$
R2	Jacobian Transform	Gaussian density $q(\widetilde{\boldsymbol{z}}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Upsilon})$	Monotone functions $\{g_j(z_j)\}_{j=1:p}$
R3	Parameter Expansion	Gaussian density $q(\widetilde{\boldsymbol{z}}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}\boldsymbol{C}^T)$	Monotone functions $\{h_j(\widetilde{z}_j)\}_{j=1:p}$

$$\mathcal{L}_{\mathrm{C}}[q_{\mathrm{VGC}}(\boldsymbol{x})] = \mathbb{E}_{\mathcal{N}(\widetilde{\boldsymbol{z}};\boldsymbol{\mu},\boldsymbol{\Sigma})}[\ell_s(\widetilde{\boldsymbol{z}}) - \ln q_{\mathrm{G}}(\widetilde{\boldsymbol{z}})],$$

$$\ell_s(\widetilde{\boldsymbol{z}}, h) = \ln p(\boldsymbol{y}, h(\widetilde{\boldsymbol{z}})) + \sum_{j=1}^{p} \ln h_j'(\widetilde{z}_j). \qquad (6)$$

The monotonic transformations $h_j(\cdot) = F_j^{-1}[\Phi(t(\cdot))]$ can be specified according to the desired parametric form of marginal posterior, if the inverse CDF F_j^{-1} is tractable. For example, the multivariate log-normal posterior can be constructed via a Gaussian copula with log-normal (LN) margins,

$$q_{\mathrm{VGC\text{-}LN}}(\boldsymbol{x}) = C_{\mathrm{G}}(\boldsymbol{u}|\boldsymbol{\Upsilon}) \prod_{j=1}^{p} \mathrm{LN}(x_j; \mu_j, \sigma_j^2). \qquad (7)$$

This also corresponds to imposing exponential transform on Gaussian variables, $\boldsymbol{x} = h(\widetilde{\boldsymbol{z}}) = \exp(\widetilde{\boldsymbol{z}})$, $\widetilde{\boldsymbol{z}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In this case, $\{\mu_j, \sigma_j^2\}_{j=1:p}$ controls the location and dispersion of the marginal density; $h(\cdot)$ does not have any additional parameters to control the shape and $\ln h'(\widetilde{z}_j) = \widetilde{z}_j$ takes a simple form. VGC-LN is further discussed in Section 6.2 and Section 6.3.

Given the copula function C, we only need to find p one-dimensional margins. However, without knowing characteristics of the latent variables, specifying appropriate parametric form for margins is a difficult task in general cases. First, the marginals might exhibit multi-modality, high skewness or kurtosis, which are troublesome for particular parametric marginals to capture. Second, a tractable inverse CDF with optimizable arguments/parameters, as required here, are available only in a handful of cases. Instead of using some arbitrary parametric form, we construct bijective transform functions via kernel mixtures, which lead to highly flexible (ideally free-form) marginal proposals.

4 Bernstein Polynomials based Monotone Transformations

The marginal densities in VGC can be recovered through Jacobian transformation,

$$f_j(x_j) = q_{\mathrm{G}}(h_j^{-1}(x_j); \mu_j, \sigma_2^2) \frac{d}{dx_j} h_j^{-1}(x_j)$$

$$= q_{\mathrm{G}}(h_j^{-1}(x_j); \mu_j, \sigma_2^2) \frac{1}{h_j'(h_j^{-1}(x_j))}, \qquad (8)$$

where the $[h_j'(h_j^{-1}(x_j))]^{-1}$ term is interpreted as a marginal-correction term. To guarantee analytical tractability, we require $h(\cdot)$ to be (i) bijective; (ii) monotonic non-decreasing; (iii) having unbounded/constrained range; (iv) differentiable with respect to both its argument and parameters; and (v) sufficiently flexible. We propose a class of continuous and smooth transformations $h(\cdot)$ constructed via kernel mixtures that automatically have these desirable properties.

4.1 Continuous Margins Constructed via Bernstein Polynomials

The Bernstein polynomials (BPs) have a uniform convergence property for continuous functions on unit interval $[0, 1]$ and have been used for nonparametric density estimation (Petrone, 1999). It seems more natural to use kernel mixtures directly as the variational proposal. However, the difficulty lies in tackling the term $f(F^{-1}(\cdot))$ involving the inverse CDF of mixtures (not analytical) and the need of a further lower bound on the entropy of mixtures. In this paper, we overcome this issue by using a sandwich-type construction of the transform $h(\widetilde{z})$[1] which maps from $(-\infty, \infty)$ to some target range building upon BP,

$$h(\widetilde{z}) = \Psi^{-1}[B(\Phi(\widetilde{z}); k, \boldsymbol{\omega})],$$

$$B(u; k, \boldsymbol{\omega}) = \sum_{r=1}^{k} \omega_{r,k} I_u(r, k-r+1), \qquad (9)$$

where $I_u(r, k-r+1)$ is the regularized incomplete beta function. $\Phi(\cdot)$ is the standard normal CDF mapping from $(-\infty, \infty)$ to $[0, 1]$, and $\Psi^{-1}(\cdot)$ is some predefined tractable inverse CDF with fixed parameters; for example, the inverse CDF of the exponential distribution helps map from $[0, 1]$ to $(0, \infty)$ for positive variables. $B(u; k, \boldsymbol{\omega})$ relocates the probability mass on the unit interval $[0, 1]$. The degree k is an unknown smoothing parameter, and $\boldsymbol{\omega}$ is the unknown mixture weights on the probability simplex $\Delta_k = \{(\omega_1, \ldots, \omega_k) : \omega_i \geq 0, \sum_i \omega_i = 1\}$. The proposed sandwich-type transformation avoids the difficulty of specifying any particular types of marginals, while still leads to tractable derivations presented in Section 5.

---

[1] The index j on $\widetilde{z}$ is temporarily omitted for simplicity, and is added back when necessary.

4.2 Variational Inverse Transform

Considering a 1-d variational approximation problem (x is a scalar, the true posterior $f(x)$ is known up to the normalizing constant), fix $q(\widetilde{z}) = \mathcal{N}(0, 1)$, thus $u = \Phi(\widetilde{z}) \sim \mathcal{U}[0, 1]$, we can learn the monotonic transformation $\xi(\cdot) = Q^{-1}(\cdot)$ on the base uniform distribution $q_0(u)$ by solving a variational problem,

$$\xi^\star(\cdot) = \arg\min_\xi \mathrm{KL}\{q(x) || f(x)\}, \quad x = \xi(u) = Q^{-1}(u),$$

i.e., if we generate $u \sim \mathcal{U}[0, 1]$, then $x = \xi^\star(u) \sim Q^\star$. $Q^\star$ is closest to the true distribution F with the minimum KL divergence. This can be interpreted as the variational counterpart of the inverse transform sampling (Devroye, 1986), termed as variational inverse transform (VIT). Our BP-based construction $\xi(\cdot) = Q^{-1}(\cdot) = \Psi^{-1}(B(u; k, \boldsymbol{\omega}))$ is one appropriate parameterization scheme for the inverse probability transformation $Q^{-1}(\cdot)$. VIT-BP offers two clear advantages. First, as opposed to fixed-form variational Bayes, it does not require any specification of parametric form for $q(x)$. Second, the difficult task of calculating the general inverse CDFs $Q^{-1}(\cdot)$ is lessened to the much easier task of calculating the predefined tractable inverse CDF $\Psi^{-1}(\cdot)$. Some choices of $\Psi(\cdot)$ include CDF of $\mathcal{N}(0, 1)$ for variables in $(-\infty, \infty)$, Beta(2, 2) for truncated variables in $(0, 1)$.

To be consistent with VIT, we shall set $\Phi(\cdot)$ in (9) to be $\Phi(\cdot | \mu, \sigma^2)$, instead of $\Phi(\cdot | 0, 1)$, such that u is always uniformly distributed. Ideally, BP itself suffices to represent arbitrary continuous distribution function on the unit interval. However, it might require a higher order k. As is demonstrated in Section 6.1, this requirement can be alleviated by incorporating auxiliary parameters $\{\mu, \sigma^2\}$ in VGC-BP, which potentially help in changing location and dispersion of the probability mass.

5 Stochastic VGC

The derivations of deterministic VGC updates are highly model-dependent. First, due to the cross terms often involved in the log likelihood/prior, the corresponding Gaussian expectations and their derivatives may not be analytically tractable. Second, owing to the non-convex nature of many problems, only locally optimal solutions can be guaranteed. In contrast, stochastic implementation of VGC only requires the evaluation of the log-likelihood and log-prior along with their derivatives, eliminating most model-specific derivations, and it provides a chance of escaping local optima by introducing randomness in gradients.

5.1 Coordinate transformations

Applying the coordinate transformations[2] of stochastic updates, $\widetilde{z} = \boldsymbol{\mu} + \boldsymbol{C}\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, introduced in (Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014), the gradient of the ELBO w.r.t. variational parameter $(\boldsymbol{\mu}, \boldsymbol{C})$ can be written as

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}_{\mathrm{C}} = \mathbb{E}_{q_G(\widetilde{z})} \left[\nabla_{\widetilde{z}} \ell_s(\widetilde{z}, h) - \nabla_{\widetilde{z}} \ln q_{\mathrm{G}}(\widetilde{z}) \right],$$
$$\nabla_{\boldsymbol{C}} \mathcal{L}_{\mathrm{C}} = \mathbb{E}_{q_G(\widetilde{z})} \left[(\nabla_{\widetilde{z}} \ell_s(\widetilde{z}, h) - \nabla_{\widetilde{z}} \ln q_{\mathrm{G}}(\widetilde{z})) \boldsymbol{\epsilon}^T \right], \quad (10)$$

where the stochastic gradient terms

$$\nabla_{\widetilde{z}_j} \ell_s(\widetilde{z}) = \nabla_{\widetilde{z}_j} \ln p(\boldsymbol{y}, h(\widetilde{z})) + \nabla_{\widetilde{z}_j} \ln h_j'(\widetilde{z}_j)$$
$$= \frac{\partial \ln p(\boldsymbol{y}, \boldsymbol{x})}{\partial x_j} h_j'(\widetilde{z}_j) + \nabla_{\widetilde{z}_j} \ln h_j'(\widetilde{z}_j).$$

According to the chain rule, the first derivative of $h(\cdot)$ w.r.t $\widetilde{z}$ is,

$$h'(\widetilde{z}) = \frac{d\Psi^{-1}[B(\Phi(\widetilde{z}); k, \boldsymbol{\omega})]}{dB(\Phi(\widetilde{z}); k, \boldsymbol{\omega})} \frac{dB(\Phi(\widetilde{z}); k, \boldsymbol{\omega})}{d\Phi(\widetilde{z})} \frac{d\Phi(\widetilde{z})}{d\widetilde{z}}$$
$$= \frac{b(\Phi(\widetilde{z}); k, \boldsymbol{\omega}) \phi(\widetilde{z})}{\psi(h(\widetilde{z}))}, \quad (11)$$

where $b(u; k, \boldsymbol{\omega}) = \sum_{r=1}^k \omega_{r,k} \beta(u; r, k - r + 1)$, $\beta(x; a, b)$ is the beta density $\beta(x; a, b) = \Gamma(a + b)/(\Gamma(a)\Gamma(b))x^{a-1}(1 - x)^{b-1}$. Therefore, $\ln h'(\widetilde{z}) = \ln b(\Phi(\widetilde{z}); k, \boldsymbol{\omega}) + \ln \phi(\widetilde{z}) - \ln \psi(h(\widetilde{z}))$ and $\nabla_{\widetilde{z}_j} \ln h_j'(\widetilde{z}_j) = h_j''(\widetilde{z}_j)/h_j'(\widetilde{z}_j)$ all take analytical expressions, where

$$h_j''(\widetilde{z}_j) = [\rho_1'(\widetilde{z}_j)\rho_2(\widetilde{z}_j)\rho_3(\widetilde{z}_j) + \rho_1(\widetilde{z}_j)\rho_2'(\widetilde{z}_j)\rho_3(\widetilde{z}_j)$$
$$- \rho_1(\widetilde{z}_j)\rho_2(\widetilde{z}_j)\rho_3'(\widetilde{z}_j)]/[\rho_3(\widetilde{z}_j)]^2,$$

where $\rho_1(\widetilde{z}_j) = b(u_j; k, \boldsymbol{\omega}^{(j)})$, $\rho_2(\widetilde{z}_j) = \phi(\widetilde{z}_j)$, $\rho_3(\widetilde{z}_j) = \psi(h_j(\widetilde{z}_j))$, $\rho_1'(\widetilde{z}_j) = \phi(\widetilde{z}_j) \sum_{r=1}^k \omega_{r,k}^{(j)} \beta'(u_j; r, k - r + 1)$, $\rho_2'(\widetilde{z}_j) = -\widetilde{z}_j \phi(\widetilde{z}_j)$, $\rho_3'(\widetilde{z}_j) = \psi'(h_j(\widetilde{z}_j))h_j'(\widetilde{z}_j)$, $u_j = \Phi(\widetilde{z}_j)$, $\phi(\cdot)$ is the PDF of $\mathcal{N}(0, 1)$, $\psi(\cdot)$ and $\psi'(\cdot)$ are the predefined PDF and its derivative respectively. Defining $\beta(x; a, 0) = \beta(x; 0, b) = 0$, the derivative is written as a combination of two polynomials of lower degree $\beta'(x; a, b) = (a + b - 1)[\beta(x; a - 1, b) - \beta(x; a, b - 1)]$.

In stochastic optimization, the gradients expressed in terms of expectations are approximated using Monte Carlo integration with finite samples. The gradients contain expectations on additive terms. Note that Rezende et al. (2014) and Titsias and Lázaro-Gredilla (2014) ignore the stochasticity in the entropy term

---

[2]If necessary, the Gaussian copula can be replaced with other appropriate parametric forms. The coordinate transformation supports many other distributions as well, for example, those described in Appendix C.2. of Rezende et al. (2014).

Algorithm 1 (VGC-BP) Stochastic Variational Gaussian Copula Inference with Bernstein Polynomials

---

Input: observed data $\boldsymbol{y}$, user specified model $\ln p(\boldsymbol{y}, \boldsymbol{x})$ and first-order derivatives $\nabla_{\boldsymbol{x}} \ln p(\boldsymbol{y}, \boldsymbol{x})$, Bernstein polynomials degree k, predefined $\Psi(\cdot)$ and $\Phi(\cdot)$

Initialize variational parameter $\Theta_0 = \left(\boldsymbol{\mu}_0, \boldsymbol{C}_0, \{\boldsymbol{\omega}_0^{(j)}\}_{j=1:p} \right)$, $t = 0$.

repeat
    $t = t + 1$,
    Sample $\widetilde{\boldsymbol{\epsilon}} \sim q_{\mathrm{G}}\left(\widetilde{\boldsymbol{\epsilon}}, \boldsymbol{0}, \boldsymbol{I}_p\right)$, and set $\widetilde{\boldsymbol{z}} = \boldsymbol{\mu}_{t-1} + \boldsymbol{C}_{t-1}\boldsymbol{\epsilon}$,
    $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \lambda_t[\nabla_{\widetilde{\boldsymbol{z}}}\ell_s(\widetilde{\boldsymbol{z}}, h) - \nabla_{\widetilde{\boldsymbol{z}}}\ln q_{\mathrm{G}}(\widetilde{\boldsymbol{z}})]$,   % Update $\boldsymbol{\mu}_{t-1}$ with stepsize λ_t
    $\boldsymbol{C}_t = \boldsymbol{C}_{t-1} + \eta_t[\nabla_{\widetilde{\boldsymbol{z}}}\ell_s(\widetilde{\boldsymbol{z}}, h) - \nabla_{\widetilde{\boldsymbol{z}}}\ln q_{\mathrm{G}}(\widetilde{\boldsymbol{z}})]\boldsymbol{\epsilon}^T$,   % Update $\boldsymbol{C}_{t-1}$ with stepsize η_t
    for $j = 1$ **to** p **do**
        $\boldsymbol{\omega}_t^{(j)} = \mathcal{P}(\boldsymbol{\omega}_{t-1}^{(j)} + \xi_t^{(j)}\nabla_{\boldsymbol{\omega}^{(j)}}\ell_s(\widetilde{\boldsymbol{z}}, h))$,   % Update $\boldsymbol{\omega}_{t-1}^{(j)}$ with stepsize $\xi_t^{(j)}$ and gradient projection $\mathcal{P}$
    end for
until convergence criterion is satisfied
Output: marginal parameters $\left(\{\boldsymbol{\omega}^{(j)}\}_{j=1:p}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}\right)$ and copula parameters $\boldsymbol{\Upsilon}$

$\mathbb{E}_{q_G(\widetilde{\boldsymbol{z}})}[-\ln q_{\mathrm{G}}(\widetilde{\boldsymbol{z}})]$ and assume $\nabla_{\boldsymbol{\mu}}\mathbb{E}_{q_G(\widetilde{\boldsymbol{z}})}[-\ln q_{\mathrm{G}}(\widetilde{\boldsymbol{z}})] = 0$ and $\nabla_{\boldsymbol{C}}\mathbb{E}_{q_G(\widetilde{\boldsymbol{z}})}[-\ln q_{\mathrm{G}}(\widetilde{\boldsymbol{z}})] = \mathrm{diag}[1/C_{jj}]_{j=1:p}$. This creates an inconsistency as we only take finite samples in approximating $\mathbb{E}_{q_G(\widetilde{\boldsymbol{z}})}[\nabla_{\widetilde{\boldsymbol{z}}}\ell_s(\widetilde{\boldsymbol{z}})]$, and perhaps surprisingly, this also results in an increase of the gradient variance and the sensitivity to the learning rates. Our method is inherently more stable, as the difference between the gradients, $\nabla_{\widetilde{\boldsymbol{z}}}[\ell_s(h(\widetilde{\boldsymbol{z}})) - q_{\mathrm{G}}(\widetilde{\boldsymbol{z}})], \forall \widetilde{\boldsymbol{z}}$, tends to zero when the convergent point is approached. In contrast, the gradients in previous method diffuses with a constant variance even around the global maximum. This phenomenon is illustrated in Section 6.2.

The alternative log derivative approach are also applicable to VGC inference and other types of copulas, see Paisley et al. (2012); Mnih and Gregor (2014); Rezende et al. (2014) for references. We leave this exploration open for future investigation.

5.2 Update the BP Weights

Under a given computational budget, we prefer a higher degree k, as there is no over-fitting issue in this variational density approximation task. Given k, the basis functions are completely known, depending only on index r. The only parameter left to be optimized in the Bernstein polynomials is the mixture weights. Therefore, this construction is relatively simpler than Gaussian mixture proposals (Gershman et al., 2012; Nguyen and Bonilla, 2014). Assuming permissibility of interchange of integration and differentiation holds, we have $\nabla_{\boldsymbol{\omega}^{(j)}}\mathcal{L} = \mathbb{E}_{q_G(\widetilde{\boldsymbol{z}})}\left[\nabla_{\boldsymbol{\omega}^{(j)}}\ell_s(\widetilde{\boldsymbol{z}}, h, \boldsymbol{y})\right]$, with the stochastic gradients

$$\nabla_{\boldsymbol{\omega}^{(j)}}\ell_s(\widetilde{\boldsymbol{z}}, h, \boldsymbol{y}) = \nabla_{\boldsymbol{\omega}^{(j)}}\ln p(\boldsymbol{y}, h(\widetilde{\boldsymbol{z}})) + \nabla_{\boldsymbol{\omega}^{(j)}}\ln h_j'(\widetilde{z}_j)$$

$$= \frac{\partial \ln p(\boldsymbol{y}, \boldsymbol{x})}{\partial x_j}\left[\frac{\partial h_j(\widetilde{z}_j)}{\partial \omega_{r,k}^{(j)}}\right]_{r=1:k} + \left[\frac{\partial \ln h_j'(\widetilde{z}_j)}{\partial \omega_{r,k}^{(j)}}\right]_{r=1:k},$$

where

$$\frac{\partial h_j(\widetilde{z}_j)}{\partial \omega_{r,k}^{(j)}} = \frac{\partial \Psi^{-1}[B(u_j; k, \boldsymbol{\omega}^{(j)})]}{\partial \omega_{r,k}^{(j)}} = \frac{I_{u_j}(r, k - r + 1)}{\psi(h_j(\widetilde{z}_j))},$$

$$\partial \ln h_j'(\widetilde{z}_j)/\partial \omega_{r,k}^{(j)} = \beta(u_j; r, k - r + 1)/b(u_j; k, \boldsymbol{\omega}^{(j)})$$
$$- \frac{\psi'(h_j(\widetilde{z}_j))}{\{\psi(h_j(\widetilde{z}_j))\}^2}I_{u_j}(r, k - r + 1).$$

The gradients w.r.t $\boldsymbol{\omega}^{(j)}$ turn into expectation straightforwardly, to enable stochastic optimization of the ELBO. To satisfy the constraints of $\boldsymbol{\omega}^{(j)}$ on the probability simplex, we apply the gradient projection operation $\mathcal{P}$ introduced in Duchi et al. (2008) with complexity $\mathcal{O}(k\log k)$. The above derivations related to BPs together with those in Section 5.1 are all analytic and model-independent. The only two model-specific terms are $\ln p(\boldsymbol{y}, \boldsymbol{x})$ and $\partial \ln p(\boldsymbol{y}, \boldsymbol{x})/\partial \boldsymbol{x}$. The stochastic optimization algorithm is summarized in Algorithm 1, with little computational overhead added relative to stochastic VG. The stability and efficiency of the stochastic optimization algorithm can be further improved by embedding adaptive subroutines (Duchi et al., 2011) and considering second-order optimization method (Fan et al., 2015).

6 Experiments

We use Gaussian copulas with fixed/free-form margins as automated *inference engines* for posterior approximation in generic hierarchical Bayesian models. We evaluate the peculiarities reproduced in the univariate margins and the posterior dependence captured broadly across latent variables. This is done by comparing VGC methods to the ground truth and other baseline methods such as MCMC,

MFVB, and VG (see Supplementary Material for detailed derivations). Matlab code for VGC is available from the GitHub repository: `https://github.com/shaobohan/VariationalGaussianCopula`

6.1 Flexible Margins

We first assess the marginal approximation accuracy of our BP-based constructions in Section 4.2, i.e., $h(\cdot) = \Psi^{-1}(B(\Phi(\widetilde{z}); k, \boldsymbol{\omega}))$ via 1-d variational optimization, where $\widetilde{z} \sim \mathcal{N}(0,1)$ in VIT-BP, and $\widetilde{z} \sim \mathcal{N}(\mu, \sigma^2)$ in VGC-BP. For fixed BP order k, the shape of $q(x)$ is adjusted solely by updating $\boldsymbol{\omega}$, according to the variational rule. In VGC-BP, the additional marginal parameters $\{\mu, \sigma^2\}$ also contribute in changing location and dispersion of $q(x)$. Examining Figure 1, VGC-BP produces more accurate densities than VIT-BP under the same order k. Hereafter, the predefined $\Psi(\cdot)$ for real variables, positive real variable, and truncated [0,1] variables are chosen to be the CDF of $\mathcal{N}(0,1)$, $\text{Exp}(1)$ and $\text{Beta}(2,2)$, respectively.
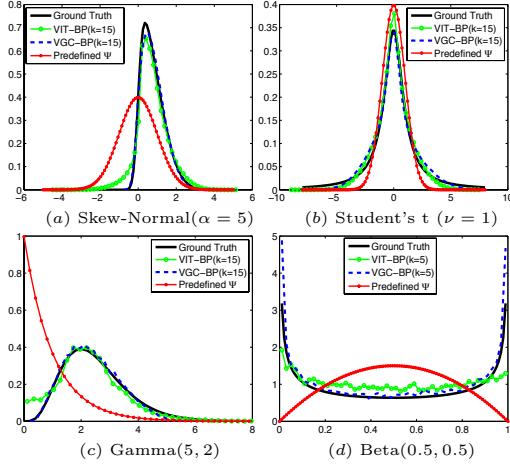

(a) Skew-Normal($\alpha = 5$) (b) Student's t ($\nu = 1$)

(c) Gamma(5, 2) (d) Beta(0.5, 0.5)

Figure 1: Marginal Adaptation: VIT-BP v.s. VGC-BP

6.2 Bivariate Log-Normal

The bivariate log-normal PDF $p(x_1, x_2)$ (Aitchison and Brown, 1957) is given by

$$p(x_1, x_2) = \exp\left(-\zeta/2\right)/[2\pi x_1 x_2 \sigma_1 \sigma_2 \sqrt{1-\rho^2}],$$
$$\zeta = \frac{1}{1-\rho^2}\left[\alpha_1^2(x_1) - 2\rho\alpha_1(x_1)\alpha_2(x_2) + \alpha_2^2(x_2)\right],$$

where $\alpha_i(x_i) = (\ln x_i - \mu_i)/\sigma_i$, $i = 1, 2$, $-1 < \rho < 1$.

We construct a bivariate Gaussian with (i) Log-normal margins (VGC-LN) and (ii) BP-based margins (VGC-BP). We set $\mu_1 = \mu_2 = 0.1$ and $\sigma_1 = \sigma_2 = 0.5$, $\rho = 0.4$ or -0.4 (first and second row in Figure 2). Both VGC-LN and VGC-BP methods presume the correct form of

the underlying copula (bivariate Gaussian) and learn the copula parameters ρ. VGC-LN further assumes exactly the true form of the univariate margins (log-normal) while VGC-BP is without any particular assumptions on parametric form of margins. Figure 2 shows that VGC-BP find as accurate joint posteriors as VGC-LN, even though the former assumes less knowledge about the true margins.
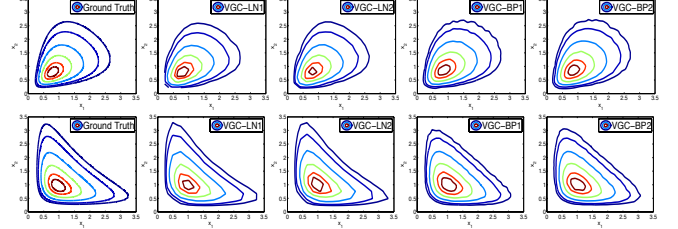

Figure 2: Approximate Posteriors via VGC methods

In updating $(\boldsymbol{\mu}, \boldsymbol{C})$, VGC-LN1 and VGC-BP1 follow the scheme in (Titsias and Lázaro-Gredilla, 2014) and neglect the stochasticity in the entropy term; while VGC-LN2 and VGC-BP2 are based on our scheme in (10). Under the same learning rates, we define the relative mean square error (RMSE) of the copula parameter as $R(\rho) = \frac{(\hat{\rho}-\rho)^2}{\rho^2}$; both VGC-LN and VGC-BP results in Figure 3 consistently show that our method leads to less noisy gradients and converges faster.
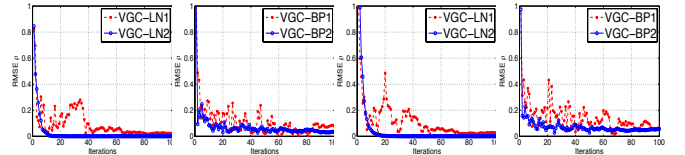

Figure 3: RMSE(ρ) of VGC-LN and VGC-BP v.s. Iterations; Left two: $\rho = 0.4$; Right two: $\rho = -0.4$

6.3 Horseshoe Shrinkage

The horseshoe distribution (Carvalho et al., 2010) can be represented in equivalent conjugate hierarchies (Neville et al., 2014) $y|\tau \sim \mathcal{N}(0, \tau)$, $\tau|\lambda \sim \text{InvGa}(0.5, \lambda)$, $\lambda \sim \text{InvGa}(0.5, 1)$. Here we assume $y = 0.01$ is the (single) observation. Denoting $\boldsymbol{x} = (x_1, x_2) = (\tau, \gamma = 1/\lambda)$, we implemented the VGC-BP algorithm ($k = 10$) and VGC-LN algorithms (deterministic implementations[3] are available in this special case). We compared them with two baselines: (i) Gibbs sampler (1×10^6 samples), and (ii) MFVB. From Figure 4, it is noted that the VGC methods with full correlation matrix (VGC-LN-full, VGC-BP-full) are able to preserve the posterior dependence and alleviate

---

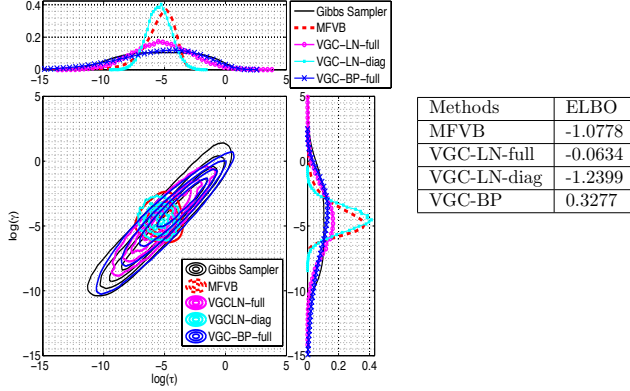[3]For gradient updates, we use a quasi-Newton strategy implemented in Schmidt (2012).

Figure 4: (Left Panel) Approximated Posteriors (Shown in Log Space for Visualization Purpose); (Right Panel) comparison of ELBO of different variational methods

Figure 5: Univariate Margins and Pairwise Posteriors

the under-estimation of the posterior variance. VGC-LN-full lead to higher ELBO than MFVB, and the gain is lost with factorized assumption $\Upsilon = I$ (VGC-LN-diag) in which case the Gaussian copula reduces to the independence copula. The restriction of parametric margins is relaxed in VGC-BP. With refinement of the mixture weights, VGC-BP leads to higher ELBO than VGC-LN. Since the Gaussian copula admits neither lower nor upper tail dependence, the posterior dependence it is able to preserve can be restrictive. It is a future research topic to explore other copula families that allow more complex posterior dependencies in variational copula inference.

6.4 Poisson Log-Linear Regression

We consider the tropical rain forest dataset (Møller and Waagepetersen, 2007), a point pattern giving the locations of 3605 trees accompanied by covariate data giving the elevation. Resampling the data into a grid of 50×50m (u_i locates the i-th grid), the number of trees y_i per unit area is modeled as, $y_i \sim \text{Poisson}(\mu_i)$, $i = 1, \ldots, n$, $\log(\mu_i) = \beta_0 + \beta_1 u_i + \beta_2 u_i^2$, $\beta_0 \sim N(0, \tau)$, $\beta_1 \sim N(0, \tau)$, $\beta_2 \sim N(0, \tau)$, $\tau \sim \text{Ga}(1, 1)$. We denote $\boldsymbol{x} = (\beta_0, \beta_1, \beta_2, \tau)$, and choosing $\Psi^{-1}(\cdot)$ to be the CDF of $\mathcal{N}(0, 1)$ or $\text{Exp}(1)$ accordingly. The implementation of VGC-BP leads to highly accurate marginal and pairwise posteriors (See Figure 5), as compared to the MCMC sampler (1×10^6 runs) implemented in JAGS [4] as reference solutions.

Interestingly, for non-conjugate models with unknown exact joint posteriors, VGC still provides a Sklar's representation of the approximated posterior, including an analytical Gaussian copula, and a number of univariate margins (summarized as univariate histograms if not in closed-form). For further uses such as calculating sample quantiles, simulating samples from
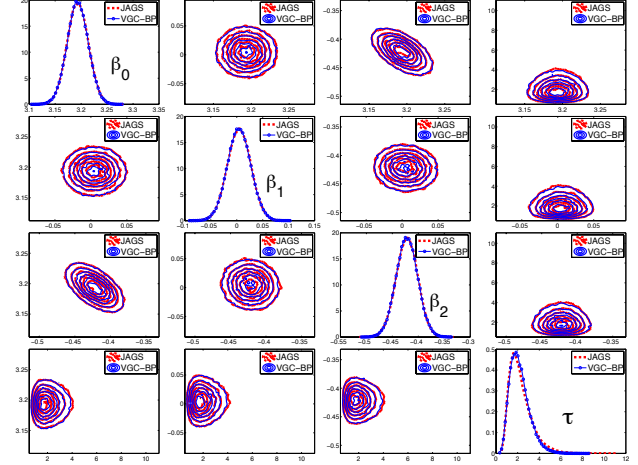
$q_{\text{VGC}}(\boldsymbol{x})$ is independent and faster, as compared to MCMC. The obtained posterior approximation could possibly improve the efficiency of Metropolis-Hastings (MH) samplers by replacing the MCMC prerun as a reasonable proposal (Schmidl et al., 2013).

The proposed method is an automated approach of approximating full posteriors. It is readily applicable to a broad scope of latent Gaussian models with non-conjugate likelihoods. Compared with the integrated nested Laplace approximation (INLA) (Rue et al., 2009) and integrated non-factorized variational inference (Han et al., 2013), our approach does not need to discretize the space for non-Gaussian variables and thus does not suffer from the limits on the number of hyperparameters.

7 Discussions

This article proposes a unified variational copula inference framework. In VGC, we have focused on Gaussian copula family for simplicity, however, other more flexible forms such as Gaussian mixture copula can be considered as well. To avoid the difficulty of specifying marginals for hidden variables, a nonparametric procedure based on Bernstein polynomials indirectly induces highly flexible univariate margins. Tran et al. (2015) and Kucukelbir et al. (2015) could potentially benefit from our flexible margins, while our approach is likely to benefit from the vine copula decomposition (Tran et al., 2015) to allow richer or more complex dependencies and the automatic differentiation techniques applied in Kucukelbir et al. (2015).

Acknowledgments

---

[4] http://mcmc-jags.sourceforge.net/

References

J. Aitchison and J. A. Brown. *The lognormal distribution with special reference to its uses in economics.* Cambridge Univ. Press, 1957.

D. Barber and C. M. Bishop. Ensemble learning in Bayesian neural networks. *Neural networks and machine learning*, 168:215–238, 1998.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

E. Challis and D. Barber. Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research (JMLR)*, 14(1):2239–2286, 2013.

L. Devroye. *Non-uniform random variate generation.* New York: Springer-Verlag., 1986.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, pages 272–279, 2008.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 12:2121–2159, 2011.

K. Fan, Z. Wang, J. Beck, J. Kwok, and K. Heller. Fast second-order stochastic backpropagation for variational inference. In *arXiv:1509.02866*, 2015.

E. Ferkingstad and H. Rue. Improving the INLA approach for approximate bayesian inference for latent Gaussian models. *arXiv:1503.07307*, 2015.

S. J. Gershman, M. D. Hoffman, and D. M. Blei. Nonparametric variational inference. In *International Conference on Machine Learning (ICML)*, 2012.

S. Han, X. Liao, and L. Carin. Integrated non-factorized variational inference. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2481–2489, 2013.

M. D. Hoffman and D. M. Blei. Structured stochastic variational inference. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research (JMLR)*, 11:3235–3268, 2010.

T. S. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning Graphical Models*, 1998.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

A. Kucukelbir, R. Ranganath, A. Gelman, and D. M. Blei. Automatic variational inference in Stan. In *arXiv:1506.03431*, 2015.

J. Li, D. J. Nott, Y. Fan, and S. A. Sisson. Extending approximate Bayesian computation methods to high dimensions via Gaussian copula. *arXiv:1504.04093*, 2015.

C. Liu, D. B. Rubin, and Y. Wu. Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, 85(4):755–770, 1998.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research (JMLR)*, 10:2295–2328, 2009.

J. S. Liu and Y. Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association (JASA)*, 94(448):1264–1274, 1999.

D. Lopez-Paz, J. M. Hernandez-Lobato, and Z. Ghahramani. Gaussian process vine copulas for multivariate dependence. In *International Conference on Machine Learning (ICML)*, pages 10–18, 2013.

A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning (ICML)*, pages 1791–1799, 2014.

J. Møller and R. P. Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

R. B. Nelsen. *An introduction to copulas.* Springer Science & Business Media, 2007.

S. E. Neville, J. T. Ormerod, and M. Wand. Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, 8:1113–1151, 2014.

T. V. Nguyen and E. V. Bonilla. Automated variational inference for Gaussian process models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1404–1412, 2014.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

J. W. Paisley, D. M. Blei, and M. I. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning (ICML)*, 2012.

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

Y. Qi and T. S. Jaakkola. Parameter expanded variational Bayesian methods. In *Advances in Neural In-*

formation Processing Systems (NIPS), pages 1097–1104, 2006.

R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, pages 1278–1286, 2014.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2): 319–392, 2009.

L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. *Advances in Neural Information Processing Systems (NIPS)*, pages 486–492, 1996.

D. Schmidl, C. Czado, S. Hug, and F. J. Theis. A vine-copula based adaptive MCMC sampler for efficient inference of dynamical systems. *Bayesian Analysis*, 8(1):1–22, 2013.

M. Schmidt. minfunc: unconstrained differentiable multivariate optimization in matlab. 2012. URL http://www.di.ens.fr/mschmidt/Software/minFunc.html.

A. Sklar. *Fonctions de Répartition à n Dimensions Et Leurs Marges*. Publ. Inst. Statist. Univ. Paris 8, 1959.

M. S. Smith. Bayesian approaches to copula modelling. *Bayesian Theory and Applications*, page 336, 2013.

P. X. Song. Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

A. Talhouk, A. Doucet, and K. Murphy. Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics*, 21(3):739–757, 2012.

M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning (ICML)*, pages 1971–1979, 2014.

D. Tran, D. M. Blei, and E. M. Airoldi. Variational inference with copula augmentation. In *arXiv:1506.03159*, 2015.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

F. L. Wauthier and M. I. Jordan. Heavy-tailed process priors for selective shrinkage. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2406–2414, 2010.