Variational Gaussian Copula Inference Supplementary Material

Shaobo Han **Xuejun Liao** **David B. Dunson**[†] **Lawrence Carin**

Department of ECE, Department of Statistical Science[†], Duke University, Durham, NC 27708, USA

A: KL Additive Decomposition

Letting the variational proposal in Sklar's representation be $q_{\mathrm{VC}}(\boldsymbol{x}) = c(\boldsymbol{u}) \prod_{j=1}^{p} f_j(x_j)$, and the true posterior be $p(\boldsymbol{x}|\boldsymbol{y}) = c^\star(\boldsymbol{v}) \prod_j f_j^\star(x_j)$, where $\boldsymbol{u} = F(\boldsymbol{x}) = [F_1(x_1), \ldots, F_p(x_p)]$, $\boldsymbol{v} = F^\star(\boldsymbol{x}) = [F_1^\star(x_1), \ldots, F_p^\star(x_p)]$. The KL divergence decomposes into additive terms,

$$
\begin{aligned}
&\mathrm{KL}\{q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})\} = \int q(\boldsymbol{x}) \left(\log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{y})} \right) d\boldsymbol{x} \\
&= \int c[F(\boldsymbol{x})] \prod_j f_j(x_j) \left(\log \frac{c[F(\boldsymbol{x})] \prod_j f_j(x_j)}{c^\star[F^\star(\boldsymbol{x})] \prod_j f_j^\star(x_j)} \right) d\boldsymbol{x} \\
&= \int c[F(\boldsymbol{x})] \left(\log \frac{c[F(\boldsymbol{x})]}{c^\star[F^\star(\boldsymbol{x})]} \right) \prod_j dF_j(x_j) \\
&+ \int c[F(\boldsymbol{x})] \prod_j f_j(x_j) \left(\log \frac{\prod_j f_j(x_j)}{\prod_j f_j^\star(x_j)} \right) \prod_j dx_j. \qquad (1)
\end{aligned}
$$

The first term in (1)

$$
\begin{aligned}
&\int c[F(\boldsymbol{x})] \left(\log \frac{c[F(\boldsymbol{x})]}{c^\star[F^\star(\boldsymbol{x})]} \right) \prod_j dF_j(x_j) \\
&= \int c(\boldsymbol{u}) \left(\log \frac{c(\boldsymbol{u})}{c^\star(F^\star(F^{-1}(\boldsymbol{u})))} \right) d\boldsymbol{u} \\
&= \mathrm{KL}\{c(\boldsymbol{u})||c^\star[F^\star(F^{-1}(\boldsymbol{u}))]\},
\end{aligned}
$$

The second term in (1)

$$
\begin{aligned}
&\int c[F(\boldsymbol{x})] \prod_j f_j(x_j) \left(\log \frac{\prod_j f_j(x_j)}{\prod_j f_j^\star(x_j)} \right) \prod_j dx_j \\
&= \sum_j \int c[F(\boldsymbol{x})] \prod_j f_j(x_j) \left(\log \frac{f_j(x_j)}{f_j^\star(x_j)} \right) \prod_j dx_j \\
&= \sum_j \int f_j(x_j) \left(\log \frac{f_j(x_j)}{f_j^\star(x_j)} \right) dx_j \text{ (Marginal Closed Property)} \\
&= \sum_j \mathrm{KL}\{f_j(x_j)||f_j^\star(x_j)\},
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\mathrm{KL}\{q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})\} &= \mathrm{KL}\{c(\boldsymbol{u})||c^\star[F^\star(F^{-1}(\boldsymbol{u}))]\} \\
&+ \sum_j \mathrm{KL}\{f_j(x_j)||f_j^\star(x_j)\} \qquad (2)
\end{aligned}
$$

B: Model-Specific Derivations

B1: Skew Normal Distribution

1. $\ln p(x) \propto \ln \phi(x) + \ln \Phi(\alpha x)$ and $\partial \ln p(x)/\partial x = -x + \alpha\phi(\alpha x)/\Phi(\alpha x)$, α is the shape parameter

2. $\Psi(x)$ is predefined as CDF of $\mathcal{N}(0, 1)$

B2: Student's t Distribution

1. $\ln p(x) \propto -(\nu + 1)/2 \ln(1 + x^2/\nu)$ and $\partial \ln p(x)/\partial x = -(\nu + 1)x/(\nu + x^2)$, $\nu > 0$ is the degrees of freedom

2. $\Psi(x)$ is predefined as CDF of $\mathcal{N}(0, 1)$

B3: Gamma Distribution

1. $\ln p(x) \propto (\alpha - 1)\ln x - \beta x$ and $\partial \ln p(x)/\partial x = (\alpha - 1)/x - \beta$, α is the shape parameter, β is the rate parameter

2. $\Psi(x)$ is predefined as CDF of $\mathrm{Exp}(1)$

B4: Beta Distribution

1. $\ln p(x) \propto (a - 1)\ln x + (b - 1)\ln(1 - x)$ and $\partial \ln p(x)/\partial x = (a - 1)/x - (b - 1)/(1 - x)$, both a, $b > 0$

2. $\Psi(x)$ is predefined as CDF of $\mathrm{Beta}(2, 2)$

B5: Bivariate Log-Normal

1. $\ln p(x_1, x_2) \propto -\ln x_1 - \ln x_2 - \zeta/2$ and

$$\frac{\partial \ln f(x_1, x_2)}{\partial x_1} = -\frac{1}{x_1} - \frac{\alpha_1(x_1) - \rho\alpha_2(x_2)}{(1-\rho^2)x_1\sigma_1}$$
$$\frac{\partial \ln f(x_1, x_2)}{\partial x_2} = -\frac{1}{x_2} - \frac{\alpha_2(x_2) - \rho\alpha_1(x_1)}{(1-\rho^2)x_2\sigma_2}$$

2. $\Psi(x)$ is predefined as CDF of Exp(1)

C. Derivations in the Horseshoe Shrinkage Model

The equilvalent hierarchical model is

$$y|\tau \sim \mathcal{N}(0, \tau), \quad \tau|\gamma \sim \text{InvGa}(0.5, \gamma), \quad \gamma \sim \text{Ga}(0.5, 1)$$

C1: Gibbs Sampler

The full conditional posterior distributions are

$$p(\tau|y, \gamma) = \text{InvGa}\left(1, y^2/2 + \gamma\right), \quad p(\gamma|\tau) = \text{Ga}\left(1, \tau^{-1} + 1\right)$$

C2: Mean-field Variational Bayes

The ELBO under MFVB is

$$\mathcal{L}_{\text{MFVB}}[q_{\text{VB}}(\tau, \gamma)] = \mathbb{E}_{q(\tau)q(\gamma)}[\ln p(y, \tau, \gamma)] + H_1[q(\tau; \alpha_1, \beta_1)] + H_2[q(\gamma; \alpha_2, \beta_2)]$$

where

$$\mathbb{E}_{q(\tau)q(\gamma)}[\ln p(y, \tau, \gamma)] = -0.5\ln(2\pi) - 2\ln\Gamma(0.5) - 2\langle\ln\tau\rangle - y^2\left\langle\tau^{-1}\right\rangle/2 - \langle\gamma\rangle\left\langle\tau^{-1}\right\rangle - \langle\gamma\rangle$$
$$H_1[q(\tau; \alpha_1, \beta_1)] = \alpha_1 + \ln\beta_1 + \ln[\Gamma(\alpha_1)] - (1+\alpha_1)\psi(\alpha_1)$$
$$H_2[q(\gamma; \alpha_2, \beta_2)] = \alpha_2 - \ln\beta_2 + \ln[\Gamma(\alpha_2)] + (1-\alpha_2)\psi(\alpha_2)$$

The variational distribution

$$q(\tau) = \mathcal{IG}(\tau; \alpha_1, \beta_1) = \mathcal{IG}\left(\tau; 1, y^2/2 + \langle\gamma\rangle\right),$$
$$q(\gamma) = \mathcal{G}(\gamma; \alpha_2, \beta_2) = \mathcal{G}\left(\gamma; 1, \left\langle\tau^{-1}\right\rangle + 1\right)$$

where

$$\langle\ln\tau\rangle = \ln\beta_1 - \psi(\alpha_1) = \ln\left(y^2/2 + \langle\gamma\rangle\right) - \psi(1),$$
$$\left\langle\tau^{-1}\right\rangle = \frac{\alpha_1}{\beta_1} = \frac{1}{(y^2/2 + \langle\gamma\rangle)}, \quad \langle\gamma\rangle = \frac{\alpha_2}{\beta_2} = \frac{1}{\langle\tau^{-1}\rangle + 1}$$

C3: Deterministic VGC-LN

Denoting $\boldsymbol{x} = (x_1, x_2) = (\tau, \gamma)$, we construct a variational Gaussian copula proposal with (1) a bivariate Gaussian copula, and (2) fixed-form margin for both $x_1 = \tau \in (0, \infty)$ and $x_2 = \gamma \in (0, \infty)$; we employ $f_j(x_j; \mu_j, \sigma_{jj}^2) = \mathcal{LN}(x_j; \mu_j, \sigma_{jj}^2)$,$x_j = h_j(\widetilde{z}_j) = \exp(\widetilde{z}_j) = g(z_j) = \exp(\sigma_{jj}z_j + \mu_j)$, $j = 1, 2$. The ELBO of VGC-LN is

$$\mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \boldsymbol{C}) = c_1 - \mu_1 + \mu_2 - \frac{y^2\exp\left(-\mu_1 + \frac{C_{11}^2}{2}\right)}{2} - \ell_0 - \exp\left(\mu_2 + \frac{C_{21}^2 + C_{22}^2}{2}\right) + \ln|\boldsymbol{C}|$$
$$\ell_0 = \exp\left((\mu_2 - \mu_1) + \frac{C_{11}^2 - 2C_{11}C_{21} + C_{21}^2 + C_{22}^2}{2}\right)$$

where $c_0 = -0.5\ln(2\pi) - 2\ln\Gamma(0.5)$, $c_1 = c_0 + \ln(2\pi e)$. The gradients are

$$\frac{\partial\mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \boldsymbol{C})}{\partial\mu_1} = -1 + \frac{y^2}{2}\exp\left(\frac{C_{11}^2}{2} - \mu_1\right) + \ell_0$$
$$\frac{\partial\mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \boldsymbol{C})}{\partial\mu_2} = 1 - \ell_0 - \exp\left(\mu_2 + \frac{C_{21}^2 + C_{22}^2}{2}\right)$$
$$\frac{\partial\mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \boldsymbol{C})}{\partial C_{11}} = -\frac{y^2}{2}C_{11}\exp\left(\frac{C_{11}^2}{2} - \mu_1\right) - (C_{11} - C_{21})\ell_0 + \frac{1}{C_{11}}$$
$$\frac{\partial\mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \boldsymbol{C})}{\partial C_{21}} = (C_{11} - C_{21})\ell_0 - C_{21}\exp\left(\mu_2 + \frac{C_{21}^2 + C_{22}^2}{2}\right)$$
$$\frac{\partial\mathcal{L}_{\text{VGC}}(\boldsymbol{\mu}, \boldsymbol{C})}{\partial C_{22}} = -C_{22}\ell_0 - C_{22}\exp\left(\mu_2 + \frac{C_{21}^2 + C_{22}^2}{2}\right) + \frac{1}{C_{22}}$$

C4: Stochastic VGC-LN

The stochastic part of the ELBO is,

$$\ell_s(\widetilde{\boldsymbol{z}}) = c_0 + \widetilde{z}_2 - \widetilde{z}_1 - \frac{y^2\exp(-\widetilde{z}_1)}{2} - \exp(\widetilde{z}_2 - \widetilde{z}_1) - \exp(\widetilde{z}_2)$$

and

$$\nabla_{\widetilde{z}_1}\ell_s(\widetilde{\boldsymbol{z}}) = -1 + \frac{y^2\exp(-\widetilde{z}_1)}{2} + \exp(\widetilde{z}_2 - \widetilde{z}_1)$$
$$\nabla_{\widetilde{z}_2}\ell_s(\widetilde{\boldsymbol{z}}) = 1 - \exp(\widetilde{z}_2 - \widetilde{z}_1) - \exp(\widetilde{z}_2)$$

C5: Stochastic VGC-BP

1. $\ln p(y, x_1, x_2) = c_0 - 2\ln x_1 - y^2/(2x_1) - x_2/x_1 - x_2$,

$$\partial\ln p(y, x_1, x_2)/\partial x_1 = -2/x_1 + y^2/(2x_1^2) + x_2/x_1^2,$$
$$\partial\ln p(y, x_1, x_2)/\partial x_2 = -1/x_1 - 1$$

2. $\Psi(x)$ is predefined as CDF of Exp(0.01).

D. Derivations in Poisson Log Linear Regression

For $i = 1, \ldots, n$, the hierarchical model is

$$y_i \sim \text{Poisson}(\mu_i), \quad \log(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,$$
$$\beta_0 \sim N(0,\tau), \quad \beta_1 \sim N(0,\tau), \quad \beta_2 \sim N(0,\tau), \quad \tau \sim \text{Ga}(1,1)$$

The log likelihood and prior,

$$\ln p(\boldsymbol{y}, \boldsymbol{\beta}, \tau) = \sum_{i=1}^{n} \ln p(y_i|\boldsymbol{\beta}) + \ln \mathcal{N}(\beta_0; 0, \tau) + \ln \mathcal{N}(\beta_1; 0, \tau) + \ln \mathcal{N}(\beta_2; 0, \tau) + \ln \text{Ga}(\tau; 1, 1)$$

where $\ln p(y_i|\boldsymbol{\beta}) = y_i \ln \mu_i - \mu_i - \ln y_i!$, and $\mu_i = \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2)$.

The derivatives are

$$\frac{\partial \ln p(\boldsymbol{y}, \boldsymbol{\beta}, \tau)}{\partial \beta_0} = \left[\sum_{i=1}^{n} (y_i - \mu_i)\right] - \tau^{-1}\beta_0$$

$$\frac{\partial \ln p(\boldsymbol{y}, \boldsymbol{\beta}, \tau)}{\partial \beta_1} = \left[\sum_{i=1}^{n} x_i (y_i - \mu_i)\right] - \tau^{-1}\beta_1$$

$$\frac{\partial \ln p(\boldsymbol{y}, \boldsymbol{\beta}, \tau)}{\partial \beta_2} = \left[\sum_{i=1}^{n} x_i^2 (y_i - \mu_i)\right] - \tau^{-1}\beta_2$$

$$\frac{\partial \ln p(\boldsymbol{y}, \boldsymbol{\beta}, \tau)}{\partial \tau} = -\frac{3}{2\tau} + \frac{\beta_0^2 + \beta_1^2 + \beta_2^2}{2\tau^2} + \frac{a_0 - 1}{\tau} - b_0$$