Bayesian Gaussian Copula Factor Models for Mixed Data (Supplement)

1 Conditional independence

Assume $F(Y_1, Y_2, Y_3)$ has a Gaussian copula with correlation matrix $\boldsymbol{C}$, that Y_3 is discrete, and that $r_{12} = 0$. Let $(Z_1, Z_2, Z_3) \sim N(\mathbf{0}, \boldsymbol{C})$ and $\mathcal{B}_c = (F_3(c-1),\ F_3(c)]$ for c in the domain of Y_3 and define $g_j(z_3) = \Phi\left(F_j(y_j) - c_{j3}z_3)/(1-c_{j3}^2)^{1/2}\right)$. It is straightforward to show that

$$Pr(Y_1 \leq y_1 | Y_3 = c) Pr(Y_2 \leq y_2 | Y_3 = c) = E(g_1(z_3))E(g_2(z_3)) \tag{1.1}$$

$$Pr(Y_1 \leq y_1,\ Y_2 \leq y_2 \mid Y_3 = c) = E(g_1(z_3)g_2(z_3)) \tag{1.2}$$

where the expectations are with respect to $\pi(z_3|y_3 = c) = TN(0, 1, F_3(c-1), F_3(c))$ and (1.2) holds because $\pi(z_1, z_2|z_3) = \pi(z_1|z_3)\pi(z_2|z_3)$ when $r_{12} = 0$. Since g_1, g_2 are monotone it is well known that $E(g_1(z_3)g_2(z_3)) \neq E(g_1(z_3))E(g_2(z_3))$ (and Y_1, Y_2 are dependent given Y_3) unless one or both functions are a.s. constant, which occurs only if one or both of Y_1, Y_2 are marginally independent of Y_3 ($c_{13}c_{23} = 0$). This result extends to conditioning on one discrete variable and any number of continuous variables since conditioning on a continuous variable $Y_4 = y_4$ implies that $Pr(z_4 = \Phi^{-1}(F(y_4)) = 1$, and $\pi(z_3|y_3, z_4)$ is again univariate truncated normal (with a different mean and variance).

2 Posterior Predictive Conditional Distributions

To sample from conditional posterior predictive distributions such as $\pi(y_1^* \mid \boldsymbol{y}^*_{(-1)} = \boldsymbol{x}, \boldsymbol{Y})$ we could sample from $\pi(\boldsymbol{y}^*|\boldsymbol{Y})$ and discard draws where $y_j^* \neq x_j$ for any $2 \leq j \leq p$. This approach can be wasteful computationally since even in moderate dimensions most samples will be discarded. Instead we might prefer to estimate this distribution directly. We can write $Pr(y_1^* \leq y \mid \boldsymbol{y}^*_{(-1)} = \boldsymbol{x},\ \boldsymbol{Y})$ as

$$\int\limits_{\mathcal{C}} \int\limits_{\mathbb{R}^{p-1}} \left(\int\limits_{-\infty}^{\hat{F}_1(y)} \pi(z_1^* |\ \boldsymbol{z}^*_{(-1)}, \boldsymbol{C})\, dz_1^* \right) \pi(\boldsymbol{z}^*_{(-1)}|\ \boldsymbol{y}^*_{(-1)} = \boldsymbol{x}, \boldsymbol{C}) \pi(\boldsymbol{C} \mid \boldsymbol{Y})\, d\boldsymbol{z}^*_{(-1)}\, d\boldsymbol{C} \tag{2.1}$$

Assume that $y_2, \ldots y_p$ are discrete, or that the empirical cdfs are used for $\hat{F}_j$ (if y_j is continuous and $\hat{F}_j$ is a smooth estimator then $z_j^* = \Phi^{-1}(\hat{F}(x_j))$ is fixed in the following). Then $\pi(\boldsymbol{z}^*_{(-1)} \mid \boldsymbol{y}^*_{(-1)} = \boldsymbol{x}, \boldsymbol{C})$ is the $(p-1)$-dimensional truncated normal distribution $N(\boldsymbol{0}, \boldsymbol{C}_{(-1)})$ where $\boldsymbol{C}_{(-1)}$ is obtained by dropping the first row and column of $\boldsymbol{C}$, restricted to the set $\mathcal{B}_x = \{\boldsymbol{z}^*_{(-1)};\ \Phi^{-1}(\hat{F}_j(x_j^-)) < z_j^* \leq \Phi^{-1}(\hat{F}_j(x_j))\ \forall\ 2 \leq j \leq p\}$ (where $F(x^-)$ is the lower limit of F at x). To estimate (2.1) from MCMC output we need to draw from this distribution (at least) once for every sample of $\boldsymbol{C}$. For a general $\boldsymbol{C}$ this is prohibitive unless p is very small, but our factor-analytic representation allows us to efficiently draw from $\pi(\boldsymbol{z}^*_{(-1)} \mid \boldsymbol{y}_{(-1)} = \boldsymbol{x}, \boldsymbol{C})$ by sampling $(p-1)$ *univariate* truncated normals: Let $\tilde{\boldsymbol{\Lambda}}_{(-1)}$ be $\tilde{\boldsymbol{\Lambda}}$ with the first row removed and $\boldsymbol{U}_{(-1)}$ be $\boldsymbol{U}$ with the first row and column removed. Since $\boldsymbol{C}_{(-1)} = \tilde{\boldsymbol{\Lambda}}_{(-1)}\tilde{\boldsymbol{\Lambda}}'_{(-1)} + \boldsymbol{U}_{(-1)}$ we have

$$\begin{aligned} \pi(\boldsymbol{z}^*_{(-1)} \mid \boldsymbol{y}^*_{(-1)} = \boldsymbol{x}, \tilde{\boldsymbol{\Lambda}}_{(-1)}) &\propto N(\boldsymbol{z}^*_{(-1)};\ \boldsymbol{0}, \tilde{\boldsymbol{\Lambda}}_{(-1)}\tilde{\boldsymbol{\Lambda}}'_{(-1)} + \boldsymbol{U}_{(-1)}) \boldsymbol{1}\big((z^*_{(-1)} \in \mathcal{B}_x)\big) \\ &\propto \int_{\mathbb{R}^k} \prod_{j=2}^{p} \Big(TN(\tilde{\boldsymbol{\lambda}}_j \boldsymbol{\eta}, u_j, a_j, b_j) \Big)\, N(\boldsymbol{\eta}; \boldsymbol{0}, \boldsymbol{I})\, d\boldsymbol{\eta} \end{aligned}$$

where $a_j = \Phi^{-1}(\hat{F}_j(x_j^-))$, $b_j = \Phi^{-1}(\hat{F}_j(x_j))$ and $\boldsymbol{\eta}$ is an auxiliary variable. Therefore we can approximate (2.1) as follows:

1. Draw $\tilde{\boldsymbol{\Lambda}}$ via the PX-Gibbs sampler, and draw $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{I})$
2. Draw $z_j^* \sim TN(\tilde{\boldsymbol{\lambda}}_j \boldsymbol{\eta}, u_j, a_j, b_j)$ for $2 \leq j \leq p$
3. For each distinct value of $\boldsymbol{y}_1$ set $\tilde{F}^{(t)}(y_i) = \int_{-\infty}^{\hat{F}_1(y_i)} N(z_1^*; m, v)\, dz_1^*$ where

$$\begin{aligned} m &= \tilde{\boldsymbol{\lambda}}_1 \tilde{\boldsymbol{\Lambda}}'_{(-1)} [\tilde{\boldsymbol{\Lambda}}_{(-1)} \tilde{\boldsymbol{\Lambda}}'_{(-1)} + \boldsymbol{U}_{(-1)}]^{-1} \boldsymbol{z}^*_{(-1)} \\ v &= 1 - \tilde{\boldsymbol{\lambda}}_1 \tilde{\boldsymbol{\Lambda}}'_{(-1)} [\tilde{\boldsymbol{\Lambda}}_{(-1)} \tilde{\boldsymbol{\Lambda}}'_{(-1)} + \boldsymbol{U}_{(-1)}]^{-1} \tilde{\boldsymbol{\Lambda}}_{(-1)} \tilde{\boldsymbol{\lambda}}'_1 \end{aligned} \tag{2.2}$$

where again the matrix inverses in (2.2) can be computed efficiently as in (**??**). This procedure provides estimates of the conditional cdf at the observed data points. For a discrete response we can then directly compute conditional probabilities, odds ratios, and so on. When y_1 is continuous these can be interpolated to give a histogram estimate of $\pi(y_1 | \boldsymbol{y}_{(-1)} = \boldsymbol{x})$ with support on the range of the observed data. A number of modifications to this approach are possible; for example, to condition on a subset of $\boldsymbol{y}_{(-1)}$ we simply drop the irrelevant rows of $\boldsymbol{\Lambda}_{(-1)}$ and only perform step 3 for the j^{th} variable if we are conditioning on y_j.

This is a natural extension of factor regression models which posit a Gaussian factor model for $(y_i, \boldsymbol{x}_i')'$, implying a linear regression model for $\pi(y_i \mid \boldsymbol{x}_i)$ (Carvalho et al., 2008; West, 2003). These are especially useful when $p > n$ as a model-based form of reduced rank regression (automatically selecting batches of correlated predictors by loading them highly on the same factor), or when there is missing data in $\boldsymbol{X}$. Here we have a flexible joint model which accommodates any ordered response or covariates while retaining the computational simplicity of factor regression models.

References

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(484):1438–1456.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. *Bayesian statistics*, 7(2003):723–732.