Appendix to "Sparsity Inducing Prior Distributions for Correlation Matrices of Longitudinal Data"

Published in the Journal of Computational and Graphical Statistics

J. T. Gaskins[*], M. J. Daniels[†], and B. H. Marcus [‡]

1. Standard Errors for Risk Estimates

Tables A.1 and A.2 contain the Monte Carlo standard errors (MCSE) of the risk estimates obtained in the simulation study of Section 5. The estimated risks are available in Tables 1 and 3 of the article.

2. Approximating the DIC

We describe in this appendix the details involved in approximating the DIC term used for model comparison of the CTQ I data, and in particular, the estimation of the integral in (10). First, we introduce notation. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{R})$ be the set of parameters, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\mathbf{R}})$ the set of the posterior estimates, and $\boldsymbol{\theta}_g$ the value of $\boldsymbol{\theta}$ at the g-th iteration of the Markov chain ($g = 1, \ldots, G$). The function $I_i(\mathbf{Y}) = I\{Q_{it} Y_t \geq 0\, \forall t\}$ indicates whether $\mathbf{Y}$ is a set of latent variables whose signs

[*] *jeremy.gaskins@louisville.edu*; Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202

[†] *mjdaniels@austin.utexas.edu*; Division of Statistics & Scientific Computation, University of Texas, Austin, TX 78712

[‡] *bmarcus@ucsd.edu*; Department of Family and Preventive Medicine, UC San Diego, San Diego, CA 92093

R	N	Loss Fcn	MCSE of Risk Estimates by Prior Shrinkage	Selection (2,1)	Selection (1,1)	flat-**R**	flat-$\mathbf{\Pi}$	Triangular	Naive Shrink
A	20	1	0.024	0.022	0.027	0.032	0.028	0.026	0.032
A	50	1	0.011	0.0088	0.0093	0.015	0.013	0.013	0.014
A	200	1	0.0020	0.0013	0.0013	0.0026	0.0024	0.0024	0.0025
B	20	1	0.0090	0.0085	0.0086	0.020	0.023	0.022	0.011
B	50	1	0.0043	0.0041	0.0041	0.0091	0.0099	0.0098	0.0040
B	200	1	0.0009	0.0008	0.0009	0.0025	0.0026	0.0026	0.0010
C	20	1	0.033	0.032	0.040	0.041	0.038	0.030	0.028
C	50	1	0.011	0.014	0.015	0.015	0.013	0.012	0.014
C	200	1	0.0025	0.0026	0.0027	0.0028	0.0026	0.0026	0.0025
D	20	1	0.026	0.028	0.031	0.039	0.033	0.029	0.035
D	50	1	0.011	0.012	0.013	0.015	0.013	0.013	0.013
D	200	1	0.0022	0.0028	0.0029	0.0026	0.0024	0.0024	0.0026
A	20	2	0.011	0.010	0.011	0.018	0.016	0.016	0.016
A	50	2	0.0057	0.0033	0.0036	0.0089	0.0083	0.0083	0.0080
A	200	2	0.0016	0.0004	0.0004	0.0020	0.0020	0.0020	0.0020
B	20	2	0.0066	0.0047	0.0051	0.016	0.018	0.017	0.0079
B	50	2	0.0039	0.0033	0.0033	0.0081	0.0088	0.0087	0.0035
B	200	2	0.0008	0.0008	0.0008	0.0024	0.0025	0.0025	0.0010
C	20	2	0.017	0.015	0.019	0.025	0.023	0.018	0.016
C	50	2	0.0062	0.0077	0.0081	0.0086	0.0079	0.0074	0.0080
C	200	2	0.0017	0.0019	0.0019	0.0019	0.0019	0.0018	0.0018
D	20	2	0.013	0.013	0.014	0.020	0.019	0.017	0.019
D	50	2	0.0071	0.0076	0.0076	0.0096	0.0092	0.0087	0.0090
D	200	2	0.0015	0.0020	0.0021	0.0018	0.0017	0.0017	0.0020

Table A.1: Monte Carlo standard errors (MCSE) of risk estimates for simulation study with dimension $J = 6$. The estimated risks are contained in Table 1 of the article. Correlation matrices: A autoregressive structure; B independence; C non-zero decaying; D sparse. Loss functions: $L_1(\hat{\mathbf{R}}, \mathbf{R}) = \mathrm{tr}(\hat{\mathbf{R}}\,\mathbf{R}^{-1}) - \log|\hat{\mathbf{R}}\,\mathbf{R}^{-1}| - p$; $L_2(\hat{\mathbf{\Pi}}, \mathbf{\Pi}) = \sum_{i<j}(\hat{\pi}_{ij} - \pi_{ij})^2$.

$\mathbf{R}$	N	Loss Fcn	MCSE of Risk Estimates by Prior Shrinkage	Selection (2,1)	Selection (1,1)	flat-$\mathbf{R}$	flat-$\mathbf{\Pi}$	Triangular	Naive Shrink
C	50	1	0.018	0.021	0.025	0.034	0.025	0.023	0.029
C	200	1	0.0037	0.0043	0.0044	0.0056	0.0050	0.0050	0.0054
D′	50	1	0.014	0.015	0.016	0.031	0.023	0.021	0.025
D′	200	1	0.0036	0.0043	0.0045	0.0057	0.0051	0.0051	0.0053
C	50	2	0.011	0.012	0.014	0.020	0.017	0.016	0.018
C	200	2	0.0028	0.0033	0.0034	0.0044	0.0042	0.0042	0.0043
D′	50	2	0.009	0.010	0.010	0.018	0.017	0.016	0.016
D′	200	2	0.0026	0.0032	0.0034	0.0042	0.0040	0.0040	0.0041

Table A.2: Monte Carlo standard errors (MCSE) of risk estimates for simulation study with dimension $J = 10$. The estimated risks are contained in Table 3 of the article. Correlation matrices: C non-zero decaying; D' sparse. Loss functions: $L_1(\hat{\mathbf{R}}, \mathbf{R}) = \text{tr}(\hat{\mathbf{R}}\,\mathbf{R}^{-1}) - \log|\hat{\mathbf{R}}\,\mathbf{R}^{-1}| - p$; $L_2(\hat{\mathbf{\Pi}}, \mathbf{\Pi}) = \sum_{i<j}(\hat{\pi}_{ij} - \pi_{ij})^2$.

agree with the data $\mathbf{Q}_i$. Define $p_i(\boldsymbol{\theta})$ to be the probability of observing $\mathbf{Q}_i$ under the parameters $\boldsymbol{\theta}$. As in (10), this is

$$p_i(\boldsymbol{\theta}) = p_i(\boldsymbol{\beta}, \mathbf{R}) = \int_{(-\infty,\infty)^J} I_i(\mathbf{y})\phi(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y},$$

where $\phi(\cdot|\boldsymbol{\theta})$ is the multivariate normal density with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{R}$ when $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{R})$. Hence, $\text{loglik}(\boldsymbol{\theta}|\mathbf{Q}_i) = \log p_i(\boldsymbol{\theta})$. As previously noted, this integral in intractable.

Using the definitions of the deviance and complexity parameter (equations (8) and (9)) and the new notation, we can write the DIC as the sum of the contributions DIC_i for each patient,

$$\text{DIC} = \sum_i \text{DIC}_i = \sum_i \left[2\log p_i(\hat{\boldsymbol{\theta}}) - 4\text{E}\{\log p_i(\boldsymbol{\theta})\}\right].$$

As observations $\mathbf{Q}_i$ are independent, it suffices to consider the per patient contribution DIC_i. Note that the expectation in the final term is with respect to the posterior distribution of the parameters $\boldsymbol{\theta}$ and will be estimated by its average over the values from the posterior sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$. Additionally, we will have to approximate the intractable $p_i(\boldsymbol{\theta})$ with some estimate $\hat{p}_i(\boldsymbol{\theta})$. So,

$$\widehat{\text{DIC}}_i = 2\log\hat{p}_i(\hat{\boldsymbol{\theta}}) - 4G^{-1}\sum_{g=1}^{G}\log\hat{p}_i(\boldsymbol{\theta}_g).$$

Note that calculation of the DIC will require $(G+1)$ estimates of the integral $p_i(\cdot)$ for each $i = 1, \ldots, N$.

To evaluate this integral we use importance sampling (Robert and Casella, 2004, Section 3.3). We take as our sampling density $t(\cdot|\boldsymbol{\xi})$, the multivariate t-distribution with 5 degrees of freedom, location parameter $\mathbf{X}_i\hat{\boldsymbol{\beta}}$, and scale matrix $k\hat{\mathbf{R}}$ for some constant $k > 1$. We define $\boldsymbol{\xi} = (\hat{\boldsymbol{\beta}}, k\hat{\mathbf{R}})$ to be the set of parameters for the sampling distribution. We choose the t-distribution so that $t(\cdot|\boldsymbol{\xi})$ will have heavier tails than $\phi(\cdot|\boldsymbol{\theta}_g)$ for $g = 1, \dots, G$ and $\hat{\boldsymbol{\theta}}$. This also motivates the choice to use a scale matrix that is an inflated version of $\hat{\mathbf{R}}$. Note that we can write $p_i(\boldsymbol{\theta})$ as

$$p_i(\boldsymbol{\theta}) = \int_{(-\infty,\infty)^J} I_i(\mathbf{y})\phi(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y} = \int_{(-\infty,\infty)^J} I_i(\mathbf{z}) \frac{\phi(\mathbf{z}|\boldsymbol{\theta})}{t(\mathbf{z}|\boldsymbol{\xi})} t(\mathbf{z}|\boldsymbol{\xi})\, d\mathbf{z}.$$

To estimate this, independently draw $\mathbf{Z}_1, \dots, \mathbf{Z}_{H_i}$ from $t(\cdot|\boldsymbol{\xi})$, and

$$\hat{p}_i(\boldsymbol{\theta}) = H_i^{-1} \sum_{h=1}^{H_i} I_i(\mathbf{Z}_h) \frac{\phi(\mathbf{Z}_h|\boldsymbol{\theta})}{t(\mathbf{Z}_h|\boldsymbol{\xi})}$$

is an unbiased and consistent estimator of $p_i(\boldsymbol{\theta})$.

Evaluation of $\widehat{\text{DIC}}_i$ involves, for each $g = 1, \dots, G$, simulating a dataset $\mathcal{Z} = \{\mathbf{Z}_h\}_{h=1}^{H_i}$ and calculating $\hat{p}_i(\boldsymbol{\theta}_g)$, followed by drawing a final $\mathcal{Z}$ to estimate $\hat{p}_i(\hat{\boldsymbol{\theta}})$. Drawing $G + 1$ independent datasets turns out to be computationally slow. Instead, we will draw a single sample $\mathcal{Z}$ to use to calculate all $\hat{p}_i(\boldsymbol{\theta}_1), \dots, \hat{p}_i(\boldsymbol{\theta}_G), \hat{p}_i(\hat{\boldsymbol{\theta}})$. It is clear that these estimates remain unbiased and consistent. What remains is to consider what effect this will have on the variability of the individual contributions to the DIC, $\widehat{\text{DIC}}_i$.

First we derive the variance of $\widehat{\text{DIC}}_i$ in the situation where we draw a new dataset $\mathcal{Z}$ for each $\hat{p}_i(\boldsymbol{\theta}_g)$. In this case,

$$\text{Var}\{\widehat{\text{DIC}}_i\} = 4\text{Var}\{\log \hat{p}_i(\hat{\boldsymbol{\theta}})\} + 16G^{-2} \sum_g \text{Var}\{\log \hat{p}_i(\boldsymbol{\theta}_g)\}, \tag{A.1}$$

where the expectation (in the variance) is with respect to the sampling distribution of $\mathcal{Z}$. For any $\boldsymbol{\theta}_g$ or $\hat{\boldsymbol{\theta}}$, this variance is

$$\begin{aligned} \text{Var}\{\log \hat{p}_i(\boldsymbol{\theta})\} &\approx p_i(\boldsymbol{\theta})^{-2}\text{Var}\{\hat{p}_i(\boldsymbol{\theta})\} = p_i(\boldsymbol{\theta})^{-2} H_i^{-1} \text{Var}\left\{ I_i(\mathbf{Z}_1) \frac{\phi(\mathbf{Z}_1|\boldsymbol{\theta})}{t(\mathbf{Z}_1|\boldsymbol{\xi})} \right\} \\ &= p_i(\boldsymbol{\theta})^{-2} H_i^{-1} \left[\text{E}\left\{ I_i(\mathbf{Z}_1) \frac{\phi(\mathbf{Z}_1|\boldsymbol{\theta})^2}{t(\mathbf{Z}_1|\boldsymbol{\xi})^2} \right\} - p_i(\boldsymbol{\theta})^2 \right], \end{aligned}$$

where the approximation in the first line is due to the delta method. This quantity can be consistently estimated by

$$\widehat{\text{Var}}\{\log \hat{p}_i(\boldsymbol{\theta})\} \;=\; \hat{p}_i(\boldsymbol{\theta})^{-2} H_i^{-1} \left[H_i^{-1} \sum_{h=1}^{H_i} I_i(\mathbf{Z}_h) \frac{\phi(\mathbf{Z}_h|\boldsymbol{\theta})^2}{t(\mathbf{Z}_h|\boldsymbol{\xi})^2} - \hat{p}_i(\boldsymbol{\theta})^2 \right]. \tag{A.2}$$

To calculate the variance of $\widehat{\text{DIC}}_i$ under our sampling scheme with a single sample $\mathcal{Z}$, note

$$\begin{aligned} \text{Var}\{\widehat{\text{DIC}}_i\} \;=\;& 4\text{Var}\{\log \hat{p}_i(\hat{\boldsymbol{\theta}})\} + 16G^{-2} \sum_g \text{Var}\{\log \hat{p}_i(\boldsymbol{\theta}_g)\} \\ & +16G^{-2} \sum_g \sum_{g' \neq g} \text{Cov}\{\log \hat{p}_i(\boldsymbol{\theta}_g), \log \hat{p}_i(\boldsymbol{\theta}_{g'})\} \\ & -16G^{-1} \sum_g \text{Cov}\{\log \hat{p}_i(\boldsymbol{\theta}_g), \log \hat{p}_i(\hat{\boldsymbol{\theta}})\}. \end{aligned} \tag{A.3}$$

The quantities on the second and third lines of (A.3) represent the additional terms due to sharing the dataset $\mathcal{Z}$ across calculations of $\hat{p}_i(\cdot)$. Define COV_i to be the sum of these covariance terms. We may write COV_i as

$$\text{COV}_i = 16G^{-2} \sum_g \sum_{g' \neq g} \left[\text{Cov}\{\log \hat{p}_i(\boldsymbol{\theta}_g), \log \hat{p}_i(\boldsymbol{\theta}_{g'})\} - \frac{G}{G-1} \text{Cov}\{\log \hat{p}_i(\boldsymbol{\theta}_g), \log \hat{p}_i(\hat{\boldsymbol{\theta}})\} \right].$$

From the delta method, we have $\text{Cov}\{\log \hat{p}_i(\boldsymbol{\theta}_g), \log \hat{p}_i(\boldsymbol{\theta})\} \approx \text{Cov}\{p_i(\boldsymbol{\theta}_g)^{-1}\hat{p}_i(\boldsymbol{\theta}_g), p_i(\boldsymbol{\theta})^{-1}\hat{p}_i(\boldsymbol{\theta})\}$, and so

$$\begin{aligned} \text{COV}_i \;&\approx 16G^{-2} \sum_g \sum_{g' \neq g} \left[\text{Cov}\left\{ \frac{\hat{p}_i(\boldsymbol{\theta}_g)}{p_i(\boldsymbol{\theta}_g)}, \frac{\hat{p}_i(\boldsymbol{\theta}_{g'})}{p_i(\boldsymbol{\theta}_{g'})} \right\} - \frac{G}{G-1} \text{Cov}\left\{ \frac{\hat{p}_i(\boldsymbol{\theta}_g)}{p_i(\boldsymbol{\theta}_g)}, \frac{\hat{p}_i(\hat{\boldsymbol{\theta}})}{p_i(\hat{\boldsymbol{\theta}})} \right\} \right] \\ =\;& 16G^{-2} \sum_g \sum_{g' \neq g} \text{Cov}\left\{ \frac{\hat{p}_i(\boldsymbol{\theta}_g)}{p_i(\boldsymbol{\theta}_g)}, \frac{\hat{p}_i(\boldsymbol{\theta}_{g'})}{p_i(\boldsymbol{\theta}_{g'})} - \frac{G}{G-1} \frac{\hat{p}_i(\hat{\boldsymbol{\theta}})}{p_i(\hat{\boldsymbol{\theta}})} \right\} \\ =\;& 16G^{-2} H_i^{-1} \sum_g \sum_{g' \neq g} \text{Cov}\left\{ I_i(\mathbf{Z}_1) \frac{\phi(\mathbf{Z}_1|\boldsymbol{\theta}_g)}{p_i(\boldsymbol{\theta}_g) t(\mathbf{Z}_1|\boldsymbol{\xi})}, \right. \\ & \left. I_i(\mathbf{Z}_1) \left(\frac{\phi(\mathbf{Z}_1|\boldsymbol{\theta}_{g'})}{p_i(\boldsymbol{\theta}_{g'}) t(\mathbf{Z}_1|\boldsymbol{\xi})} - \frac{G}{G-1} \frac{\phi(\mathbf{Z}_1|\hat{\boldsymbol{\theta}})}{p_i(\hat{\boldsymbol{\theta}}) t(\mathbf{Z}_1|\boldsymbol{\xi})} \right) \right\} \\ =\;& 16G^{-2} H_i^{-1} \sum_g \sum_{g' \neq g} \left[\frac{G}{G-1} \right. \\ & \left. + \text{E}\left\{ I_i(\mathbf{Z}_1) \frac{\phi(\mathbf{Z}_1|\boldsymbol{\theta}_g)}{p_i(\boldsymbol{\theta}_g) t(\mathbf{Z}_1|\boldsymbol{\xi})} \left(\frac{\phi(\mathbf{Z}_1|\boldsymbol{\theta}_{g'})}{p_i(\boldsymbol{\theta}_{g'}) t(\mathbf{Z}_1|\boldsymbol{\xi})} - \frac{G}{G-1} \frac{\phi(\mathbf{Z}_1|\hat{\boldsymbol{\theta}})}{p_i(\hat{\boldsymbol{\theta}}) t(\mathbf{Z}_1|\boldsymbol{\xi})} \right) \right\} \right]. \end{aligned} \tag{A.4}$$

As long as this quantity (A.4) is small (relative to the independence variance estimator (A.1)), we may save computational time by only sampling one dataset $\mathcal{Z}$ without sacrificing precision.

In the situation of the analysis of the CTQ data, estimation of COV_i with a representative sample of observations showed COV_i to be small relative to the independence variance estimator (A.1). In fact, this term is often negative, indicating that using the common dataset $\mathcal{Z}$ may improve estimation for some observations i. In the representative sample we considered, the addition of the COV_i term tended to lead to changes in the standard error of $\widehat{\text{DIC}_i}$ ranging from a decrease of 5% to an increase of 25%. As it is computationally infeasible to compute COV_i for all observations (a nested loop over g' inside a loop over g), we estimate the standard error of the DIC estimate using the independence estimator obtained from (A.1) and (A.2).

The Dev, p_D, and DIC estimates in Table 4 are computed in this way. The importance sampling size H_i is chosen by first drawing 200,000 values of $\mathbf{Z}_h \sim t(\cdot|\boldsymbol{\xi})$, where variance scaling factor k is 1.5^2. This choice of k is made with consideration to the dimension J of $\mathbf{Z}$ (increasing J should correspond to increasing k), how far from the origin $\boldsymbol{\mu}$ tends to be, how likely $I_i(\mathbf{Z})$ is to be one, among other considerations; ultimately, trial-and-error with small choices of H_i led us conclude that $k = 1.5^2$ works reasonably well. Having drawn 200,000 values of $\mathbf{Z}_h$, if $\sum_h I_i(\mathbf{Z}_h) \geq 2000$ (i.e., at least 2000 of the $\mathbf{Z}_h$'s have signs appropriate for a latent variable of $\mathbf{Q}_i$), then H_i =200,000 and this set $\{\mathbf{Z}_h\}_{h=1}^{200{,}000}$ is the importance sample $\mathcal{Z}$. If not, we continue to draw additional sets of 200,000 $\mathbf{Z}_h$'s to append to the dataset until $\sum_h I_i(\mathbf{Z}_h) \geq 2000$. This implies that we have larger samples for those patients i with small values of $p_i(\boldsymbol{\theta})$. This helps to control the variance of DIC_i since the term is preceded by $p_i(\boldsymbol{\theta})^{-2}$ (see equation (A.2)). With this scheme, we estimate the standard errors for our DIC estimates in Table 4 to be around 0.5.