Supplementary material to "Penalized likelihood estimation of a trivariate additive probit model"

Supplementary Material A: Examples of covariate effects

In what follows subscript m is omitted to avoid cluttering the notation.

Non-linear effects For continuous variables the smooth functions are represented using the regression spline approach popularized by Eilers & Marx (1996). In particular, for each continuous variable $z_{\nu i}$ we use equation (2) and employ low rank thin plate regression splines (Wood, 2006) which are numerically stable and have convenient mathematical properties, although other spline definitions (including B-splines and cubic regression splines) and corresponding penalties are supported in our implementation. To enforce smoothness, a conventional integrated square second derivative penalty is typically employed, that is $\mathbf{S}_\nu = \int \mathbf{d}_\nu(z_\nu)\mathbf{d}_\nu(z_\nu)^\top dz_\nu$, where the j^{th} element of $\mathbf{d}_\nu(z_\nu)$ is equal to $\partial^2 b_{\nu,j}(z_\nu)/\partial z_\nu^2$ and integration is over the range of z_ν. The formulae used to compute the basis functions and penalties for many spline definitions are provided in Ruppert et al. (2003) and Wood (2006).

Linear and Random Effects In general, no penalty is assigned to the parametric part of the model. That is, when $\mathbf{v}_{mi}$ is composed of binary and categorical variables, the entries in the penalty matrix that correspond to these variables are equal to zero. However, if the coefficients of, for instance, some factor variables in the model are weakly or not identified by the data then some penalization on the effects of these variables may be required. This

can be achieved by employing, for instance, a Ridge-type penalty (which is made up of a smoothing parameter and an identity penalty matrix). This is equivalent to the assumption that the coefficients of the factor variable are i.i.d. normal random effects with unknown variance (e.g., Ruppert et al., 2003; Wood, 2006).

Spatial Effects To allow the probabilities of the responses to co-vary smoothly across, say, the regions of a country we can include in the model a variable that can exploit the spatial dependence of observations in neighbouring areas. Also, spatially adjacent regions are more likely to share similar effects. When a geographic area is divided into discrete contiguous geographic units, spatial information can be modelled via a Markov random field smoother. In this case, the spatial regional effects can be represented as $s_\nu(z_{\nu i}) = \mathbf{L}_\nu(z_{\nu i})\boldsymbol{\alpha}_\nu$, where $\boldsymbol{\alpha}_\nu = (\alpha_{\nu,1}, \ldots, \alpha_{\nu,\mathfrak{R}})^\top$ denotes the vector of spatial effects, $\mathfrak{R}$ is the total number of regions and $\mathbf{L}_\nu(z_{\nu i})$ is a set of area labels. The $[i, \mathfrak{r}]^{\text{th}}$ entry of the corresponding design matrix, that links observation i with the corresponding spatial effect, is equal to 1 if the observation belongs to region $\mathfrak{r}$ and 0 otherwise, $\forall \mathfrak{r} = 1, \ldots, \mathfrak{R}$. Following the assumption that spatially adjacent regions share similar effects, we form the smoothing penalty based on the neighbourhood structure of the geographic units

$$\mathbf{S}_\nu[\mathfrak{r}, \mathfrak{q}] = \begin{cases} -1 & \text{if} \quad \mathfrak{r} \neq \mathfrak{q} \wedge \mathfrak{r} \text{ and } \mathfrak{q} \text{ are adjacent neighbors} \\ 0 & \text{if} \quad \mathfrak{r} \neq \mathfrak{q} \wedge \mathfrak{r} \text{ and } \mathfrak{q} \text{ are not adjacent neighbors} \\ K_\mathfrak{r} & \text{if} \quad \mathfrak{r} = \mathfrak{q} \wedge \mathfrak{r} \sim \mathfrak{q} \end{cases},$$

where $K_\mathfrak{r}$ is the number of neighbours for region $\mathfrak{r}$. In a stochastic interpretation, this penalty is equivalent to the assumption that $\boldsymbol{\alpha}_\nu$ follows a Gaussian Markov random field (e.g., Rue & Held, 2005).

Supplementary Material B: Proof of Lemma 1

Proof. For convenience we ignore index $\tilde{k}$ and term $\mathcal{Y}_{i\tilde{k}}$. By definition,

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{y}_i;\boldsymbol{\delta}) &= \mathbb{P}(-\tilde{y}_{1i}y^*_{1i}\leq 0,\ldots,-\tilde{y}_{Mi}y^*_{Mi}\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}(\eta_{1i}+\varepsilon_{1i})\leq 0,\ldots,-\tilde{y}_{Mi}(\eta_{Mi}+\varepsilon_{Mi})\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}\eta_{1i}-\tilde{y}_{1i}\varepsilon_{1i}\leq 0,\ldots,-\tilde{y}_{Mi}\eta_{Mi}-\tilde{y}_{Mi}\varepsilon_{Mi}\leq 0)\\
&= \mathbb{P}(-\tilde{y}_{1i}\varepsilon_{1i}\leq \tilde{y}_{1i}\eta_{1i},\ldots,-\tilde{y}_{Mi}\varepsilon_{Mi}\leq \tilde{y}_{Mi}\eta_{Mi})\\
&= \Phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i;\mathbf{0},\boldsymbol{\Sigma})\\
&= \int_{-\infty}^{\tilde{y}_{Mi}\eta_{Mi}}\cdots\int_{-\infty}^{\tilde{y}_{1i}\eta_{1i}}\phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\Sigma})\prod_{\tilde{c}=1}^{M}dl_{\tilde{c},i}. \qquad (1)
\end{aligned}$$

Since $\tilde{y}_{mi}$ is either equal to -1 or 1, it follows that $\boldsymbol{\mathcal{B}}_i=\boldsymbol{\mathcal{B}}_i^{-1}$ and $|\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i|=|\boldsymbol{\Sigma}|$. In addition, the pdf of a multivariate normal vector $-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\Sigma}$ can be re-expressed as the pdf of a multivariate normal vector $\boldsymbol{\varepsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i$, that is

$$\begin{aligned}
\phi_{M,-\boldsymbol{\mathcal{B}}_i\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(-\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i)^\top(\boldsymbol{\Sigma})^{-1}(-\boldsymbol{\mathcal{B}}_i\boldsymbol{l}_i)\right\}\\
&= |2\pi(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{l}_i^\top(\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)^{-1}\boldsymbol{l}_i\right\}\\
&= \phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i).
\end{aligned}$$

Therefore, equation (1) can be written as

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{y}_i;\boldsymbol{\delta}) &= \int_{-\infty}^{\tilde{y}_{Mi}\eta_{Mi}}\cdots\int_{-\infty}^{\tilde{y}_{1i}\eta_{1i}}\phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{l}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)\prod_{\tilde{c}=1}^{M}dl_{\tilde{c},i}\\
&= \Phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{\mathcal{B}}_i\boldsymbol{\eta}_i;\mathbf{0},\boldsymbol{\mathcal{B}}_i\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_i)\\
&= \Phi_{M,\boldsymbol{\varepsilon}_i}(\boldsymbol{w}_i;\mathbf{0},\boldsymbol{\Upsilon}_i),
\end{aligned}$$

where

$$\boldsymbol{\Upsilon}_i \;=\; \begin{pmatrix} 1 & \vartheta_{12} & \dots & \vartheta_{1M} \\ \vartheta_{12} & 1 & \dots & \vartheta_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \vartheta_{1M} & \vartheta_{2M} & \dots & 1 \end{pmatrix}.$$

Note that the above derivation applies to all $\tilde{k}$s, thus the likelihood $\mathcal{L}_{i\tilde{k}}$ is equal to

$$\mathcal{L}_{i\tilde{k}}(\boldsymbol{y}_i;\boldsymbol{\delta}) \;=\; \{\Phi_{M,\boldsymbol{\varepsilon}_i}((\boldsymbol{w}_i)_{\tilde{k}};\mathbf{0},(\boldsymbol{\Upsilon}_i)_{\tilde{k}})\}^{\mathcal{Y}_{i\tilde{k}}},$$

as required. □

Supplementary Material C: Computation of trivariate normal integrals

First we describe the algorithm implemented in `pmnorm()` and then the method by Trinh & Genz (2015).

C.1: Numerical computation of multivariate normal integrals

In general, accurate approximations can be employed via `ghkvec()` in `bayesm` (Rossi, 2015), `pCopula()` in `copula` (Marius Hofert & Yan, 2015) and `pmnorm()` in `mnormt` (Azzalini, 2014), all implemented in the `R` environment. We adopted the latter approach as it was found to be more efficient than the former ones. In what follows we describe in detail the numerical method used in `pmnorm()` for the evaluation of multivariate normal integrals.

Introduction

Let $(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}}) = (\mathcal{E}_1, \mathcal{J}_1) \times (\mathcal{E}_2, \mathcal{J}_2) \times \ldots \times (\mathcal{E}_M, \mathcal{J}_M)$ be a M-dimensional rectangle. Then the problem is to find

$$\Phi_M(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}}) \;=\; \frac{1}{\sqrt{|\boldsymbol{\Upsilon}_i|(2\pi)^M}} \int_{\mathcal{E}_1}^{\mathcal{J}_1} \ldots \int_{\mathcal{E}_M}^{\mathcal{J}_M} \exp\left(-\frac{1}{2}\boldsymbol{l}_i^\top \boldsymbol{\Upsilon}_i^{-1} \boldsymbol{l}_i\right) d\boldsymbol{l}_i, \tag{2}$$

where $|\boldsymbol{\Upsilon}_i|$ denotes the determinant of $\boldsymbol{\Upsilon}_i$. Since we are interested in the value of the distribution function $\Phi_M(\boldsymbol{w}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}_i)$, we have that $\boldsymbol{\mathcal{E}} = (-\infty, \ldots, -\infty)$; this reduces the number of variables in the problem and makes the evaluation of Φ_M simpler (see next section for more details). In addition, the upper bound $\boldsymbol{\mathcal{J}}$ is equal to $(w_{1,i}, \ldots, w_{M,i})$. For $M = 1$ and $M = 2$, a reliable way to calculate the distribution function is via `pnorm()` in `stats` (Team & contributors worldwide, 2015) and `pbinorm()` in `VGAM` (Yee, 2015). Here, we assume $M > 2$ and we describe Genz's approach for computing Φ_M which uses numerical integration software based on sub-region adaptive methods. A problem, however, that arises

with these methods is that they assume finite integration limits. Because infinite limits are used in our case, we need to handle them: we apply a sequence of transformations to turn the problem into a form that allows for efficient computation of Φ_M. Note that even if $\boldsymbol{\mathcal{E}}$ is finite, the transformations are also applied in order to make the numerical computation of the integral easy. The set of transformations that are employed are described in the next section.

Genz's method

The basic idea of this method is to transform the original domain of integration $(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}})$ to $[0,1]^M = [0,1] \times [0,1] \times \ldots \times [0,1]$. First we will keep the domain of integration general and assume that both $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{J}}$ are finite. Then we move onto our case where $\mathcal{E}_m = -\infty$ and $\mathcal{J}_m = w_{m,i}$, $\forall m$. Genz's method can be employed using the following sequence of three transformations.

(T.1) We begin by employing the Cholesky decomposition transformation $\boldsymbol{l}_i = \mathbf{C}_i^* \boldsymbol{a}_i$, where $\mathbf{C}_i^*$ denotes the Cholesky factor of the covariance matrix $\boldsymbol{\Upsilon}_i$, such that $\mathbf{C}_i^*$ is a lower triangular matrix and $\boldsymbol{\Upsilon}_i = \mathbf{C}_i^* \mathbf{C}_i^{*\top}$. Vector $\boldsymbol{a}_i$ consists of univariate standard normal random variables that are independent of each other. Applying this transformation to

equation (2) leads to

$$\begin{aligned}
\Phi_M(\boldsymbol{\mathcal{E}},\boldsymbol{\mathcal{J}}) &= 1/\sqrt{|\boldsymbol{\Upsilon}_i|(2\pi)^M}\int_{\mathcal{E}'_1}^{\mathcal{J}'_1}\int_{\mathcal{E}'_2(a_1)}^{\mathcal{J}'_2(a_1)}\dots\int_{\mathcal{E}'_M(a_1,\dots,a_{M-1})}^{\mathcal{J}'_M(a_1,\dots,a_{M-1})}\exp\left(-\frac{1}{2}(\mathbf{C}_i^*\boldsymbol{a}_i)^\top\right.\\
&\left.(\mathbf{C}_i^*\mathbf{C}_i^{*\top})^{*-1}(\mathbf{C}_i^*\boldsymbol{a}_i)\right)|\mathbf{C}_i^*|d\boldsymbol{a}_i = 1/\left(|\boldsymbol{\Upsilon}_i^{\frac{1}{2}}|(2\pi)^{\frac{M}{2}}\right)\int_{\mathcal{E}'_1}^{\mathcal{J}'_1}\int_{\mathcal{E}'_2(a_1)}^{\mathcal{J}'_2(a_1)}\dots\\
&\int_{\mathcal{E}'_M(a_1,\dots,a_{M-1})}^{\mathcal{J}'_M(a_1,\dots,a_{M-1})}\exp\left(-\frac{1}{2}\boldsymbol{a}_i^\top\mathbf{C}_i^{*\top}\mathbf{C}_i^{*-\top}\mathbf{C}_i^{*-1}\mathbf{C}_i^*\boldsymbol{a}_i\right)|\boldsymbol{\Upsilon}_i^{\frac{1}{2}}|d\boldsymbol{a}_i = \frac{1}{(2\pi)^{\frac{M}{2}}}\\
&\int_{\mathcal{E}'_1}^{\mathcal{J}'_1}\int_{\mathcal{E}'_2(a_1)}^{\mathcal{J}'_2(a_1)}\dots\int_{\mathcal{E}'_M(a_1,\dots,a_{M-1})}^{\mathcal{J}'_M(a_1,\dots,a_{M-1})}\exp\left(-\frac{1}{2}\boldsymbol{a}_i^\top\boldsymbol{a}_i\right)d\boldsymbol{a}_i = \int_{\mathcal{E}'_1}^{\mathcal{J}'_1}\frac{1}{\sqrt{2\pi}}e^{-\frac{a_1^2}{2}}\\
&\int_{\mathcal{E}'_2(a_1)}^{\mathcal{J}'_2(a_1)}\frac{1}{\sqrt{2\pi}}e^{-\frac{a_2^2}{2}}\dots\int_{\mathcal{E}'_M(a_1,\dots,a_{M-1})}^{\mathcal{J}'_M(a_1,\dots,a_{M-1})}\frac{1}{\sqrt{2\pi}}e^{-\frac{a_M^2}{2}}da_M\dots da_1 =\\
&= \int_{\mathcal{E}'_1}^{\mathcal{J}'_1}\phi(a_1)\int_{\mathcal{E}'_2(a_1)}^{\mathcal{J}'_2(a_1)}\phi(a_2)\dots\int_{\mathcal{E}'_M(a_1,\dots,a_{M-1})}^{\mathcal{J}'_M(a_1,\dots,a_{M-1})}\phi(a_M)da_M\dots da_1, \qquad (3)
\end{aligned}$$

where the limits $\mathcal{E}'_m(a_1,\dots,a_{M-1})$ and $\mathcal{J}'_m(a_1,\dots,a_{M-1})$ come from inequality $\boldsymbol{\mathcal{E}} \leq \boldsymbol{l}_i = \mathbf{C}_i^*\boldsymbol{a}_i \leq \boldsymbol{\mathcal{J}}$. Specifically, for $m=1$

$$\mathcal{E}'_1 = \mathcal{E}_1 \leq a_1 \leq \mathcal{J}_1 = \mathcal{J}'_1,$$

while for $m = 2,\dots,M$

$$\mathcal{E}'_m = \frac{\left(\mathcal{E}_m - \sum_{h=1}^{m-1}c_{mh,i}^*a_h\right)}{c_{mm,i}^*} \leq a_m \leq \frac{\left(\mathcal{J}_m - \sum_{h=1}^{m-1}c_{mh,i}^*a_h\right)}{c_{mm,i}^*} = \mathcal{J}'_m,$$

where $\mathcal{E}'_m = \mathcal{E}'_m(a_1,\dots,a_{M-1})$ and $\mathcal{J}'_m = \mathcal{J}'_m(a_1,\dots,a_{M-1})$. The elements $c_{mh,i}^*$ and $c_{mm,i}^*$ refer to the components of the lower triangular matrix $\mathbf{C}_i^*$, that is

$$\mathbf{C}_i^* = \begin{pmatrix} 1 & 0 & \dots & 0 \\ c_{21,i}^* & c_{22,i}^* & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ c_{M1,i}^* & c_{M2,i}^* & \dots & c_{MM,i}^* \end{pmatrix}.$$

The element $c^*_{11,i}$ is equal to 1 because of the following relation: $c^*_{11,i} = \sqrt{r_{11,i}}$ where $r_{11,i}$ refers to the (1, 1) diagonal element of $\boldsymbol{\Upsilon}_i$. Since $r_{11,i} = 1$, it follows that $c^*_{11,i} = \sqrt{1} = 1$.

(T.2) Next we transform the a_m's by using $a_m = \Phi^{-1}(\mathcal{Z}_m)$, where $\Phi(a_m)$ is the standard univariate normal distribution. Therefore equation (3) becomes

$$\begin{aligned}\Phi_M(\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{J}}) &= \int_{\mathcal{S}_1}^{\mathcal{T}_1} \int_{\mathcal{S}_2(\mathcal{Z}_1)}^{\mathcal{T}_2(\mathcal{Z}_1)} \dots \int_{\mathcal{S}_M(\mathcal{Z}_1,\dots,\mathcal{Z}_{M-1})}^{\mathcal{T}_M(\mathcal{Z}_1,\dots,\mathcal{Z}_{M-1})} \phi(a_1)\phi(a_2)\dots\phi(a_M)\frac{d\mathcal{Z}_M \dots d\mathcal{Z}_1}{\phi(a_1)\dots\phi(a_M)} \\ &= \int_{\mathcal{S}_1}^{\mathcal{T}_1} \int_{\mathcal{S}_2(\mathcal{Z}_1)}^{\mathcal{T}_2(\mathcal{Z}_1)} \dots \int_{\mathcal{S}_M(\mathcal{Z}_1,\dots,\mathcal{Z}_{M-1})}^{\mathcal{T}_M(\mathcal{Z}_1,\dots,\mathcal{Z}_{M-1})} d\mathcal{Z}_M \dots d\mathcal{Z}_1, \qquad (4)\end{aligned}$$

where the limits for $m = 1$ can be defined as

$$\mathcal{S}_1 = \Phi(\mathcal{E}_1) \le \mathcal{Z}_1 \le \Phi(\mathcal{J}_1) = \mathcal{T}_1,$$

while for $m = 2, \dots, M$

$$\mathcal{S}_M = \Phi\left(\frac{\mathcal{E}_m - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{Z}_h)}{c^*_{mm,i}}\right) \le \mathcal{Z}_m \le \Phi\left(\frac{\mathcal{J}_m - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{Z}_h)}{c^*_{mm,i}}\right) = \mathcal{T}_M,$$

where $\mathcal{S}_M$ and $\mathcal{T}_M$ refer to $\mathcal{S}_M(\mathcal{Z}_1, \dots, \mathcal{Z}_{M-1})$ and $\mathcal{T}_M(\mathcal{Z}_1, \dots, \mathcal{Z}_{M-1})$.

(T.3) Even though (4) is much simpler than (3), the integration region is more complicated. To overcome this, Genz (1992) suggested the transformation $\mathcal{Z}_m = \mathcal{S}_m + \omega_m(\mathcal{T}_m - \mathcal{S}_m)$, which standardizes this region, that is $0 \le \omega_m \le 1$, $\forall m$. In addition,

$$\frac{d\mathcal{Z}_m}{d\omega_m} = \mathcal{T}_m - \mathcal{S}_m \implies d\mathcal{Z}_m = (\mathcal{T}_m - \mathcal{S}_m)d\omega_m.$$

Therefore, (4) can be expressed as

$$\begin{aligned}\Phi_M(\boldsymbol{\mathcal{J}}) &= \int_0^1 \int_0^1 \dots \int_0^1 (\mathcal{T}_1 - \mathcal{S}_1)(\mathcal{T}_2 - \mathcal{S}_2)\dots(\mathcal{T}_M - \mathcal{S}_M)d\omega_M \dots d\omega_1 \\ &= (\mathcal{T}_1 - \mathcal{S}_1)\int_0^1 (\mathcal{T}_2 - \mathcal{S}_2)\dots\int_0^1 (\mathcal{T}_M - \mathcal{S}_M)d\boldsymbol{\omega},\end{aligned}$$

where $\mathcal{S}_m = \Phi\left(\left(\mathcal{E}_m - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{S}_h + \omega_h(\mathcal{T}_h - \mathcal{S}_h))\right)/c^*_{mm,i}\right)$ and $\mathcal{T}_m = \Phi\left(\left(\mathcal{J}_m - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{S}_h + \omega_h(\mathcal{T}_h - \mathcal{S}_h))\right)/c^*_{mm,i}\right)$. Since both $\mathcal{S}_m$ and $\mathcal{T}_m$ do not depend on ω_m, the innermost integral is equal to 1 and the number of integration variables can be reduced to $M-1$. Therefore, standard numerical integration methods can be applied for the computation of

$$\Phi_M(\boldsymbol{\mathcal{J}}) = \int_0^1 \dots \int_0^1 \tilde{f}(\omega_1, \dots, \omega_{M-1}) d\boldsymbol{\omega},$$

for $\tilde{f}(\omega_1, \dots, \omega_{M-1}) = (\mathcal{T}_1 - \mathcal{S}_1)(\mathcal{T}_2(\omega_1) - \mathcal{S}_2(\omega_1)) \dots (\mathcal{T}_M(\omega_1, \dots, \omega_{M-1}) - \mathcal{S}_M(\omega_1 \dots, \omega_{M-1}))$.

Computation of $\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i)$ using Genz's method

Since we are interested in the computation of the multivariate normal distribution function $\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i)$, $\mathcal{E}_m = -\infty$ and $\mathcal{J}_m = w_{m,i}$, $\forall m$. Therefore,

$$\mathcal{S}_m = \Phi\left(\frac{-\infty - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{S}_h + \omega_h(\mathcal{T}_h - \mathcal{S}_h))}{c^*_{mm,i}}\right) \to 0,$$

and

$$\begin{aligned} \mathcal{T}_m &= \Phi\left(\frac{w_{m,i} - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\mathcal{S}_h + \omega_h(\mathcal{T}_h - \mathcal{S}_h))}{c^*_{mm,i}}\right) \\ &= \Phi\left(\frac{w_{m,i} - \sum_{h=1}^{m-1} c^*_{mh,i}\Phi^{-1}(\omega_h \mathcal{T}_h))}{c^*_{mm,i}}\right), \end{aligned}$$

since $\mathcal{S}_h = 0$, for all h. It follows that

$$\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i) = \int_0^1 \dots \int_0^1 \mathcal{T}_1\mathcal{T}_2(\omega_1) \dots \mathcal{T}_M(\omega_1, \dots, \omega_{M-1}) d\boldsymbol{\omega}. \quad (5)$$

Once we get the transformed expression (5), the sub-region adaptive method is applied (see next section) and thus the cumulative distribution function Φ_M is obtained.

The algorithm

The algorithm that is used in the subroutine `sadmvn()` in `Fortran-77` for the numerical computation of the multivariate normal distribution function $\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i)$ is based on subdivisions of $[0, 1]$, where each sub-region is used to provide a better approximation to $\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i)$. As previously described, we set $\mathcal{S}_m = 0$ to avoid wasteful evaluation of Φ.

The basic algorithm can be described as follows. Suppose that $\tilde{\epsilon}$ denotes the global absolute error and $\bar{\mathrm{N}}_{\max}$ is the maximum number of sub-regions. The algorithm starts with region $\tilde{\mathrm{R}}_{11} = [0, 1]^M$. At the $\bar{\mathrm{N}}^{\text{th}}$ step, $[0, 1]$ is partitioned into $\bar{\mathrm{N}}$ sub-regions $\tilde{\mathrm{R}}_{\bar{\mathrm{N}}1}, \ldots \tilde{\mathrm{R}}_{\bar{\mathrm{N}}\bar{\mathrm{N}}}$ and in each sub-region we get estimates $\tilde{\mathrm{I}}_{\bar{\mathrm{N}}1}, \ldots, \tilde{\mathrm{I}}_{\bar{\mathrm{N}}\bar{\mathrm{N}}}$ of the corresponding integrals by applying quadrature rules. Moreover, we obtain absolute error estimates $\tilde{E}_{\bar{\mathrm{N}}1}, \ldots, \tilde{E}_{\bar{\mathrm{N}}\bar{\mathrm{N}}}$. If $\tilde{E}_{\bar{\mathrm{N}}1} + \ldots + \tilde{E}_{\bar{\mathrm{N}}\bar{\mathrm{N}}} < \tilde{\epsilon} = 10^{-6}$ or $\bar{\mathrm{N}} \geq \bar{\mathrm{N}}_{\max} = 2000 \times M$ then the algorithm stops. Otherwise a new subdivision has to be determined and the above procedure is repeated. Further details about the algorithm can be found in Genz (1991), Genz (1992), Genz & Kass (1997) and Genz & Bretz (2002).

C.2: Bivariate conditioning approximation for trivariate normal integrals

This section describes the bivariate conditioning algorithm applied for the evaluation of trivariate normal integrals, which is based on the work by Trinh & Genz (2015). As described in Section 3, the computation of triple integrals is required only for the evaluation of $\mathbb{P}(y_{1i} = 1, y_{2i} = 1, y_{3i} = 1)$; the remaining probabilities can be evaluated via univariate and bivariate normal integrals. Thus, the aim is to provide accurate and low computational cost methods for approximating the triple integrals

$$\Phi_3\left(\eta_{1i}, \eta_{2i}, \eta_{3i}; \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|(2\pi)^3}} \int_{-\infty}^{\eta_{1i}} \int_{-\infty}^{\eta_{2i}} \int_{-\infty}^{\eta_{3i}} \exp\left(-\frac{1}{2}\boldsymbol{l}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{l}_i\right) d\boldsymbol{l}_i,$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Upsilon}_i)_1$.

The algorithm is based on the Cholesky decomposition of the correlation matrix $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^\top$, where $\mathbf{C}$ is a lower triangular matrix - the decomposition always exists as $\mathbf{\Sigma}$ is symmetric and positive-definite because of the restrictions imposed on the correlation parameters. Based on this, we have that $\boldsymbol{l}_i^\top \mathbf{\Sigma}^{-1} \boldsymbol{l}_i = \boldsymbol{l}_i^\top \mathbf{C}^{-\top} \mathbf{C}^{-1} \boldsymbol{l}_i$ and by using transformation $\boldsymbol{l}_i = \mathbf{C}\mathbf{a}_i$ we get $\boldsymbol{l}_i^\top \mathbf{\Sigma}^{-1} \boldsymbol{l}_i = \mathbf{a}_i^\top \mathbf{a}_i$ with $d\boldsymbol{l}_i = |\mathbf{C}| d\mathbf{a}_i = \sqrt{|\mathbf{\Sigma}|} d\mathbf{a}_i$. The integrals are transformed according to $-\infty \leq \mathbf{C}\mathbf{a}_i \leq \boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i = (\eta_{1i}, \eta_{2i}, \eta_{3i})^\top$. Specifically, the limits can be determined as follows

$$
\begin{aligned}
-\infty &\leq \mathrm{a}_{1i} \leq \frac{\eta_{1i}}{c_{11}} = \frac{\eta_{1i}}{\sqrt{\sigma_{11}}} = \eta'_{1i} \\
-\infty &\leq \mathrm{a}_{2i} \leq \frac{\eta_{2i} - c_{21}\mathrm{a}_{1i}}{c_{22}} = \frac{\eta_{2i} - c_{21}\mathrm{a}_{1i}}{\sqrt{\sigma_{22}}} = \eta'_{2i} \\
-\infty &\leq \mathrm{a}_{2i} \leq \frac{\eta_{3i} - c_{31}\mathrm{a}_{1i} - c_{32}\mathrm{a}_{2i}}{c_{33}} = \frac{\eta_{3i} - c_{31}\mathrm{a}_{1i} - c_{32}\mathrm{a}_{2i}}{\sqrt{\sigma_{33}}} = \eta'_{3i}.
\end{aligned}
$$

The a_{zi} values, $\forall z = 1, 2$, cannot be computed directly, so they are approximated using their truncated expected values: $\tilde{\mu}_{\mathrm{a}_{zi}} = \mathbb{E}(-\infty, \eta'_{zi}) = \left(\phi(-\infty) - \phi(\eta'_{zi})\right) / \left(\Phi(\eta'_{zi}) - \Phi(-\infty)\right) = -\phi(\eta'_{zi})/\Phi(\eta'_{zi})$. The basic idea of this replacement is that these values are the average values that the a_{zi}s would have if we simulated a_{zi}s with values taken from truncated univariate distributions. In order to improve the accuracy of the results the authors apply variable re-ordering. They specify that these orderings do not change the value of the probabilities as long as the integration limits and corresponding rows and columns of $\mathbf{\Sigma}$ are also permuted. Specifically, sorting the variables so that those with the shortest integration interval widths are the outer integration variables reduces the overall variation of the integrand and thus makes the numerical integration problem easier.

The algorithm is structured as follows.

Step 1 First, we need to select the outermost integration variable. This can be done by choosing the variable ς so that

$$
\varsigma = \mathrm{argmin}_{1 \leq \varsigma \leq 3} \left\{ \Phi\left(\frac{\eta_{\varsigma i}}{\sqrt{\sigma_{\varsigma\varsigma}}}\right) - \Phi(-\infty) \right\} = \mathrm{argmin}_{1 \leq \varsigma \leq 3} \left\{ \Phi\left(\frac{\eta_{\varsigma i}}{\sqrt{\sigma_{\varsigma\varsigma}}}\right) \right\}.
$$

The rows and columns of $\mathbf{\Sigma}$ as well as the integration limits for variables 1 and ς are interchanged. The elements in the first column of $\mathbf{C}$ are computed as follows: $c_{11} = \sqrt{\sigma_{11}}$, $c_{21} = \sigma_{21}/c_{11}$ and $c_{31} = \sigma_{31}/c_{11}$, where $\sigma_{..}$ denotes the $(\cdot,\cdot)^{\text{th}}$ element of $\mathbf{\Sigma}$. Then, we set $\hat{\eta}_{1i} = \eta'_{1i}$ and $\tilde{\mu}_{\text{a}_{1i}} = -\phi(\hat{\eta}_{1i})/\Phi(\hat{\eta}_{1i})$.

Step 2 Next, ς is chosen such that

$$\begin{aligned}\varsigma &= \text{argmin}_{2\leq\varsigma\leq 3}\left\{\Phi\left(\frac{\eta_{\varsigma i} - c_{\varsigma 1}\tilde{\mu}_{\text{a}_{1i}}}{\sqrt{\sigma_{\varsigma\varsigma} - c_{\varsigma 1}^2}}\right) - \Phi(-\infty)\right\} \\ &= \text{argmin}_{2\leq\varsigma\leq 3}\left\{\frac{\eta_{\varsigma i} - c_{\varsigma 1}\tilde{\mu}_{\text{a}_{1i}}}{\sqrt{\sigma_{\varsigma\varsigma} - c_{\varsigma 1}^2}}\right\}.\end{aligned}$$

The rows and columns of $\mathbf{\Sigma}$, the integration limits, and c_{12} and $c_{\varsigma 2}$ are interchanged. The elements in the second column of $\mathbf{C}$ are computed as follows: $c_{22} = \sqrt{\sigma_{22} - c_{21}^2}$, $c_{32} = (\sigma_{32} - c_{21}c_{31})/c_{22}$. Then we let $\hat{\eta}_{2i} = (\eta_{2i} - c_{21}\tilde{\mu}_{\text{a}_{1i}})/c_{22}$ and $\tilde{\mu}_{\text{a}_{2i}} = -\phi(\hat{\eta}_{2i})/\Phi(\hat{\eta}_{2i})$.

Step 3 At this step, we calculate the $(3,3)^{\text{th}}$ element of $\mathbf{C}$ as $c_{33} = \sqrt{\sigma_{33} - c_{31}^2 - c_{32}^2}$ and we set $\hat{\eta}_{3i} = (\eta_{3i} - c_{31}\tilde{\mu}_{\text{a}_{1i}} - c_{32}\tilde{\mu}_{\text{a}_{2i}})/c_{33}$.

Step 4 Based on the resulting $\mathbf{C}$ matrix, we can determine $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{D}}$ using the relation $\tilde{\mathbf{L}}\tilde{\mathbf{D}}\tilde{\mathbf{L}}^\top = \mathbf{C}\mathbf{C}^\top$, where $\tilde{\mathbf{D}} = \mathbf{C}_{\tilde{\mathbf{D}}}\mathbf{C}_{\tilde{\mathbf{D}}}^\top$, $\tilde{\mathbf{L}} = \mathbf{C}\mathbf{C}_{\tilde{\mathbf{D}}}^{-1}$ and $\mathbf{C}_{\tilde{\mathbf{D}}}$ denotes the block diagonal matrix of $\mathbf{C}$.

Step 5 Once we obtain $\mathbf{C}$, $\tilde{\mathbf{L}}$, $\tilde{\mathbf{D}}$ and the new upper integration limits, say $\tilde{\eta}_{1i}$, $\tilde{\eta}_{2i}$ and $\tilde{\eta}_{3i}$, the next step is the computation of the bivariate normal approximation. In particular, based on a similar transformation that has been discussed above, we obtain the updated upper integration limits as follows

$$\tilde{\eta}_{1i} = \frac{\hat{\eta}_{1i}}{\sqrt{\tilde{d}_{11}}}, \tilde{\eta}_{2i} = \frac{\hat{\eta}_{2i}}{\sqrt{\tilde{d}_{22}}}, \tilde{\eta}_{3i} = \frac{\hat{\eta}_{3i} - \tilde{g}_3}{\sqrt{\tilde{d}_{33}}},$$

where $\tilde{g}_3 = \tilde{l}_{31}\tilde{e}_1 + \tilde{l}_{32}\tilde{e}_2$, $\tilde{e}_1 = \bar{\mu}_1\sqrt{\tilde{d}_{11}}$, $\tilde{e}_2 = \bar{\mu}_2\sqrt{\tilde{d}_{22}}$, $\bar{\mu}_1 = 1/\mathcal{F}\left\{-\rho\phi(\tilde{\eta}_{2i})\Phi\left((\tilde{\eta}_{1i} - \rho\tilde{\eta}_{2i})/\tilde{q}\right) - \phi(\tilde{\eta}_{1i})\Phi\left((\tilde{\eta}_{2i} - \rho\tilde{\eta}_{1i})/\tilde{q}\right)\right\}$, $\bar{\mu}_2 = 1/\mathcal{F}\left\{-\rho\phi(\tilde{\eta}_{1i})\Phi\left((\tilde{\eta}_{2i} - \rho\tilde{\eta}_{1i})/\tilde{q}\right)\right.$ $\left.-\phi(\tilde{\eta}_{2i})\Phi\left((\tilde{\eta}_{1i} - \rho\tilde{\eta}_{2i})/\tilde{q}\right)\right\}$, $\rho = \tilde{d}_{12}/\sqrt{\tilde{d}_{11}\tilde{d}_{22}}$, $\tilde{q} = \sqrt{1-\rho}$, $\mathcal{F} = \Phi_2\left(\tilde{\eta}_{1i}, \tilde{\eta}_{2i}; \mathbf{\Omega}\right)$ and $\mathbf{\Omega}$

is a 2×2 correlation matrix with 1s in the diagonals and ρ in the off-diagonals. The elements $\tilde{d}_{..}$ and $\tilde{l}_{..}$ correspond o the $(\cdot, \cdot)^{\text{th}}$ entry of $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{L}}$, respectively.

Step 6 Based on Trinh & Genz (2015), the bivariate normal approximation for trivariate normal probabilities can be written as follows

$$\Phi_3(\eta_{1i}, \eta_{2i}, \eta_{3i}; \boldsymbol{\Sigma}) \approx \Phi_2\left(\tilde{\eta}_{1i}, \tilde{\eta}_{2i}; \boldsymbol{\Omega}\right) \Phi(\tilde{\eta}_{3i}).$$

Supplementary Material D: Restrictions on the correlation parameters

D.1: The eigenvalue method

Constraints on the correlation parameters have been discussed in the previous literature: a proof on this was first given by Stanley & Wang (1969), while novel geometric proofs were provided by Glass & Collins (1970) and Leung & Lam (1975). Based on the property of positive-definiteness of correlation matrices, Hubert (1972) also provided a proof for the bounds. Here, the restriction is imposed using the eigenvalue method (Rousseeuw & Molenberghs, 1993). The method assumes that a positive-definite correlation matrix $(\boldsymbol{\Upsilon}_i)_{\tilde{k}}$ is expressed as $(\boldsymbol{\Upsilon}_i)_{\tilde{k}} = \bar{\mathbf{P}}\bar{\mathbf{D}}\bar{\mathbf{P}}^\top$, $\forall\tilde{k}$, where $\bar{\mathbf{D}}$ is a diagonal matrix containing the eigenvalues of $(\boldsymbol{\Upsilon}_i)_{\tilde{k}}$ and $\bar{\mathbf{P}}$ is an orthogonal matrix of corresponding eigenvectors. In this context, when $(\boldsymbol{\Upsilon}_i)_{\tilde{k}}$ is not positive-definite, some eigenvalues are negative and typically not large in absolute sense. According to Rousseeuw & Molenberghs (1993), a common approach for transforming a non-positive-definite matrix into a positive-definite one is to replace the negative eigenvalues by their absolute values; that is re-express $(\boldsymbol{\Upsilon}_i)_{\tilde{k}}$ as $(\boldsymbol{\Upsilon}_i)'_{\tilde{k}} = \bar{\mathbf{P}}\mathbf{D}'\bar{\mathbf{P}}^\top$ where $\mathbf{D}'$ now contains positive eigenvalues. The diagonal elements of $(\boldsymbol{\Upsilon}_i)'_{\tilde{k}}$ will not necessarily be equal to 1. Therefore, we transform $(\boldsymbol{\Upsilon}_i)'_{\tilde{k}}$ to $(\tilde{\boldsymbol{\Upsilon}}_i)_{\tilde{k}} = \tilde{\mathbf{D}}'(\boldsymbol{\Upsilon}_i)'_{\tilde{k}}\tilde{\mathbf{D}}'^\top$ where $\tilde{\mathbf{D}}'$ is the diagonal matrix with elements equal to $1/\sqrt{r'_{mm,i}}$ with $r'_{mm,i}$ denoting the diagonal element of $(\boldsymbol{\Upsilon}_i)'_{\tilde{k}}$, $\forall\tilde{k}, m$.

D.2: Geometric proof of the restriction on a correlation matrix

Geometric proofs of the restriction on a correlation matrix were first provided by Glass & Collins (1970) and Leung & Lam (1975). In what follows, we discuss a proof and show that the restriction on the values ϑ_{12} can assume when ϑ_{13} and ϑ_{23} are fixed can be displayed through spherical triangles.

Suppose that the n observations of the error term ε_{1i} are coordinates on the n-orthogonal

axes of an n-dimensional space. Thus, the observations of ε_{1i} may be considered as corresponding to a vector, $\tilde{\boldsymbol{\varepsilon}}_1$, in the n-space. Similarly, two vectors corresponding to the n observations on ε_{2i} and ε_{3i} may be established in the n-space. By using the well known result, that the Pearson's coefficient is equivalent to the cosine of the angle between two vectors (Anderson et al., 1958, p.49-50), we re-express ϑ_{zk} as

$$\vartheta_{zk} \quad = \quad \cos(\boldsymbol{\varphi}_{zk}), \tag{6}$$

where $\boldsymbol{\varphi}_{zk}$ denotes the angle that separates $\tilde{\boldsymbol{\varepsilon}}_z$ and $\tilde{\boldsymbol{\varepsilon}}_k$. Now, consider vectors $\tilde{\boldsymbol{\varepsilon}}_1$, $\tilde{\boldsymbol{\varepsilon}}_2$ and $\tilde{\boldsymbol{\varepsilon}}_3$ in a three-dimensional subspace of the n-dimensional space. Let the angles separating $\tilde{\boldsymbol{\varepsilon}}_1$ and $\tilde{\boldsymbol{\varepsilon}}_2$, $\tilde{\boldsymbol{\varepsilon}}_1$ and $\tilde{\boldsymbol{\varepsilon}}_3$, and $\tilde{\boldsymbol{\varepsilon}}_2$ and $\tilde{\boldsymbol{\varepsilon}}_3$ be fixed at $\boldsymbol{\varphi}_{12}$, $\boldsymbol{\varphi}_{13}$ and $\boldsymbol{\varphi}_{23}$, respectively. Then, $\tilde{\boldsymbol{\varepsilon}}_1$, $\tilde{\boldsymbol{\varepsilon}}_2$ and $\tilde{\boldsymbol{\varepsilon}}_3$ form a spherical triangle on the surface of a sphere of radius equal to one, centred at the origin $\boldsymbol{O} = (0, 0, 0)$ with vertices $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ (Figure 1). Planes $\mathcal{P}_2$ and $\mathcal{P}_3$, $\mathcal{P}_1$ and $\mathcal{P}_3$, and $\mathcal{P}_1$ and $\mathcal{P}_2$ form the dihedral angles $\angle\mathcal{CAB}$, $\angle\mathcal{CBA}$ and $\angle\mathcal{ACB}$ respectively. Suppose that $\angle\mathcal{CAB} = \mathfrak{a}$, $\angle\mathcal{CBA} = \mathfrak{b}$ and $\angle\mathcal{ACB} = \mathfrak{c}$ and assume that angles $\boldsymbol{\varphi}_{12}$, $\boldsymbol{\varphi}_{13}$, $\boldsymbol{\varphi}_{23}$, $\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{c}$ are between 0 and π radians. By using the spherical law of cosines for angles, we have the following three equations

$$\cos\boldsymbol{\varphi}_{12} \quad = \quad \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23} + \sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23}\cos\mathfrak{c}, \tag{7}$$

$$\cos\boldsymbol{\varphi}_{13} \quad = \quad \cos\boldsymbol{\varphi}_{12}\cos\boldsymbol{\varphi}_{23} + \sin\boldsymbol{\varphi}_{12}\sin\boldsymbol{\varphi}_{23}\cos\mathfrak{b}, \tag{8}$$

$$\cos\boldsymbol{\varphi}_{23} \quad = \quad \cos\boldsymbol{\varphi}_{12}\cos\boldsymbol{\varphi}_{13} + \sin\boldsymbol{\varphi}_{12}\sin\boldsymbol{\varphi}_{13}\cos\mathfrak{a}. \tag{9}$$

Solving (7), (8) and (9) with respect to $\mathfrak{c}$, $\mathfrak{b}$ and $\mathfrak{a}$, respectively, it can be shown that the correlation parameters are restricted to a specific range. For instance, by solving equation (7) for $\cos\mathfrak{c}$ we have that $\cos\mathfrak{c} = (\cos\boldsymbol{\varphi}_{12} - \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23})/\sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23}$. Since $\cos\mathfrak{c} \in (-1, 1)$ it follows that $-1 < (\cos\boldsymbol{\varphi}_{12} - \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23})/\sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23} < 1$ which implies that

$-\sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23} < \cos\boldsymbol{\varphi}_{12} - \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23} < \sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23}$ and therefore

$$\cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23} - \sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23} < \cos\boldsymbol{\varphi}_{12} < \cos\boldsymbol{\varphi}_{13}\cos\boldsymbol{\varphi}_{23} + \sin\boldsymbol{\varphi}_{13}\sin\boldsymbol{\varphi}_{23}. \tag{10}$$

Then, by using equation (6) and the trigonometric identity $\cos^2(\boldsymbol{\varphi}_{zk}) + \sin^2(\boldsymbol{\varphi}_{zk}) = 1 \implies \sin(\boldsymbol{\varphi}_{zk}) = \sqrt{1-\vartheta_{zk}^2}$, $\forall z = 1, 2, k = 3$, it follows that inequality (10) becomes

$$\vartheta_{13}\vartheta_{23} - \sqrt{1-\vartheta_{13}^2}\sqrt{1-\vartheta_{23}^2} < \vartheta_{12} < \vartheta_{13}\vartheta_{23} + \sqrt{1-\vartheta_{13}^2}\sqrt{1-\vartheta_{23}^2},$$

which is equal to expression (5) in Section 3. The interval for ϑ_{13} and ϑ_{23} is obtained by solving (8) and (9) respectively.

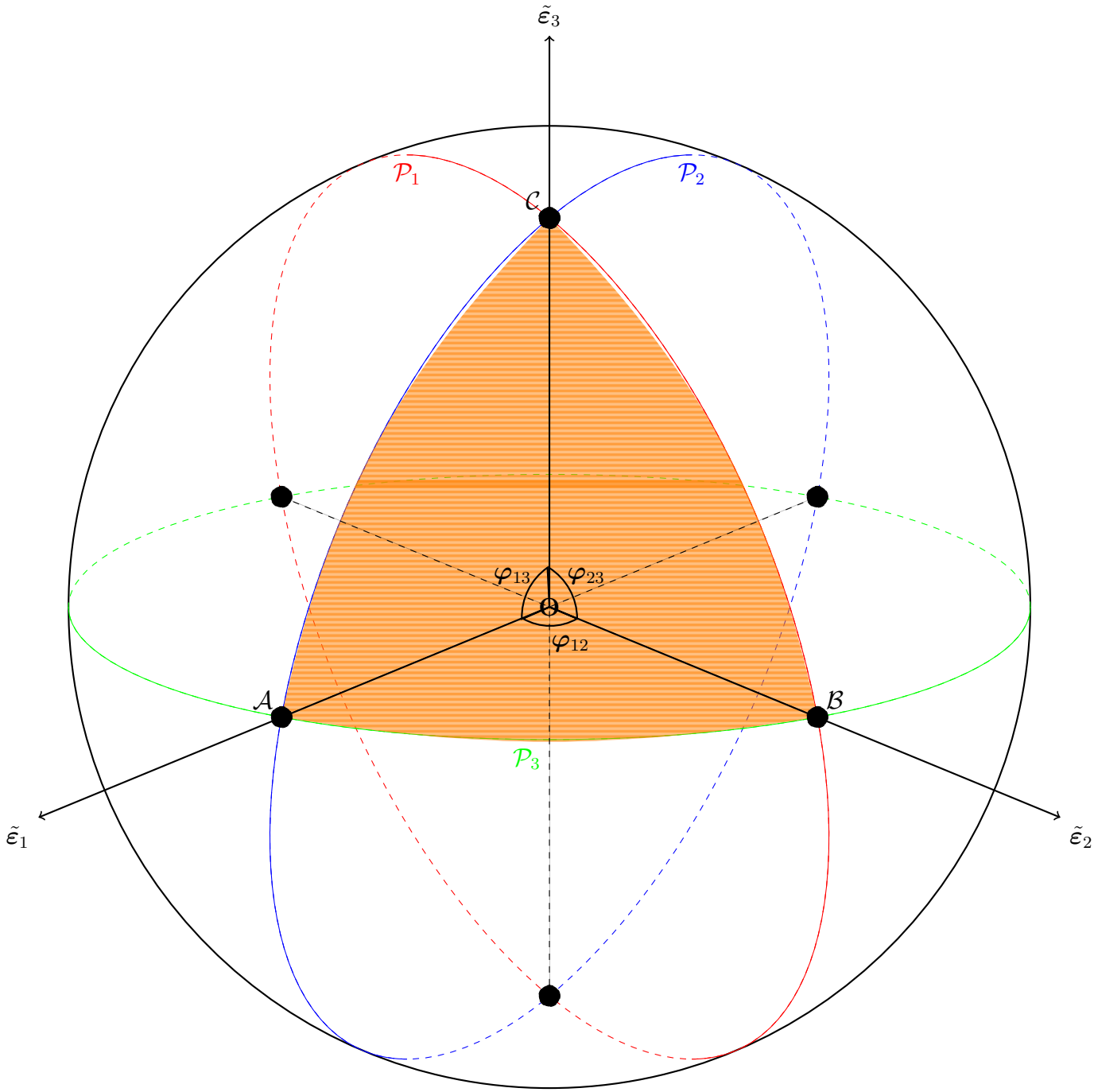

Figure 1: Spherical representation of intercorrelations among the error terms $\tilde{\varepsilon}_1$, $\tilde{\varepsilon}_2$ and $\tilde{\varepsilon}_3$.

Supplementary Material E: Trust-region and line-search methods

As presented in Section 3.1, each iteration $\varkappa$ of the trust-region algorithm solves the sub-problem

$$\begin{aligned} \min_{\boldsymbol{s}} \mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]}) &:= -\left\{ \ell_p\left(\boldsymbol{\delta}^{[\varkappa]}\right) + \boldsymbol{s}^\top \boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa]}) + \frac{1}{2}\boldsymbol{s}^\top \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\varkappa]})\boldsymbol{s} \right\} \\ &\quad \text{subject to } \|\boldsymbol{s}\| \leq \boldsymbol{\Delta}^{[\varkappa]}, \\ \boldsymbol{\delta}^{[\varkappa+1]} &= \arg\min_{\boldsymbol{s}} \mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]}) + \boldsymbol{\delta}^{[\varkappa]}. \end{aligned}$$

Line-search methods compute $\boldsymbol{s}^{[\varkappa]}$ by minimising the unconstrained problem given in the first line. The current solution $\boldsymbol{\delta}^{[\varkappa+1]}$ is then updated by scaling the step $\boldsymbol{s}^{[\varkappa]}$ by a factor $\boldsymbol{\tau}^{[\varkappa]}$ that approximately minimizes $-\ell_p(\boldsymbol{\delta})$ along the line that passes through $\boldsymbol{\delta}^{[\varkappa]}$ in the direction of $\boldsymbol{s}^{[\varkappa]}$, $\boldsymbol{\delta}^{[\varkappa+1]} = \boldsymbol{\delta}^{[\varkappa]} + \boldsymbol{\tau}^{[\varkappa]}\boldsymbol{s}^{[\varkappa]}$. If the function is non-convex then the optimizer may search far away from $\boldsymbol{\delta}^{[\varkappa]}$ but still choose $\boldsymbol{\delta}^{[\varkappa+1]}$ to be close to $\boldsymbol{\delta}^{[\varkappa]}$. In some cases the function will be evaluated so far away from $\boldsymbol{\delta}^{[\varkappa]}$ that it will not be finite and the algorithm will fail. On the contrary, trust-region methods use a maximum distance for the move from $\boldsymbol{\delta}^{[\varkappa]}$ to $\boldsymbol{\delta}^{[\varkappa+1]}$ based on a region $\mathcal{R}^{[\varkappa]}$ around the current iterate $\boldsymbol{\delta}^{[\varkappa]}$ in which the algorithm 'trusts' that model function $\mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]})$ behaves like objective function $\ell_p(\boldsymbol{\delta})$. Current iteration $\boldsymbol{\delta}^{[\varkappa]}$ is updated with $\boldsymbol{s}^{[\varkappa]}$ if this step produces an improvement over the objective function $\ell_p(\boldsymbol{\delta})$, $\boldsymbol{\delta}^{[\varkappa+1]} = \boldsymbol{\delta}^{[\varkappa]} + \boldsymbol{s}^{[\varkappa]}$. Since points outside $\mathcal{R}^{[\varkappa]}$ are not considered, the algorithm never runs too far from the current iteration. The trust-region is shrunken if the proposed point in the region is not better than the current point, in which case the new problem is solved with smaller region. If the quadratic model is a good representation of the original objective function, then trial point $\boldsymbol{\delta}^{[\varkappa+1]}$ becomes the new iterate and the trust-region is enlarged, i.e. the iteration is successful. The trust-region algorithm is summarised in Algorithm 1. A detailed description of trust-region and line search techniques can be found in Nocedal & Wright

(2006, Chap. 3, 4).

Algorithm 1 (Trust-Region Algorithm)

Require:
$\boldsymbol{\Delta}_{\max} > 0$, $\boldsymbol{\delta}^{[0]}$, $\boldsymbol{s}^{[0]}$, $\boldsymbol{\Delta}^{[0]} \in (0, \boldsymbol{\Delta}_{\max})$
Ensure:
$\|\boldsymbol{s}^{[\varkappa+1]}\| \geq 1.490116 \times 10^{-8}$ or $\varkappa \leq 100$

for $\varkappa = 0, 1, 2, \ldots$ **do**
$\boldsymbol{s}^{[\varkappa+1]} := \arg\min_{\boldsymbol{s}} \mathcal{Q}_p(\boldsymbol{\delta}^{[\varkappa]})$, subject to $\|\boldsymbol{s}\| \leq \boldsymbol{\Delta}^{[\varkappa]}$
$\tilde{r}^{[\varkappa+1]} = \left\{\ell_p(\boldsymbol{\delta}^{[\varkappa]}) - \ell_p\left(\boldsymbol{\delta}^{[\varkappa]} + \boldsymbol{s}^{[\varkappa]}\right)\right\} / \left\{\ell_p\left(\boldsymbol{\delta}^{[\varkappa]}\right) - \mathcal{Q}_p\left(\boldsymbol{s}^{[k\varkappa+1]}\right)\right\}$
if $\tilde{r}^{[\varkappa]} < 1/4$ **then**
$\boldsymbol{\Delta}^{[\varkappa+1]} = \boldsymbol{\Delta}^{[\varkappa]}/4$
else if $\tilde{r}^{[\varkappa]} > 3/4$ and $\|\boldsymbol{s}^{[\varkappa]}\| = \boldsymbol{\Delta}^{[\varkappa]}$ **then**
$\boldsymbol{\Delta}^{[\varkappa+1]} = \min\left(2\boldsymbol{\Delta}^{[\varkappa]}, \boldsymbol{\Delta}_{\max}\right)$
else
$\boldsymbol{\Delta}^{[\varkappa+1]} = \boldsymbol{\Delta}^{[\varkappa]}$
end if
if $\tilde{r}^{[\varkappa]} \geq 1/4$ **then**
$\boldsymbol{\delta}^{[\varkappa+1]} = \boldsymbol{s}^{[\varkappa+1]} + \boldsymbol{\delta}^{[\varkappa]}$
else
$\boldsymbol{\delta}^{[\varkappa+1]} = \boldsymbol{\delta}^{[\varkappa]}$
end if
end for

Supplementary Material F: Proof of Propositions 1 and 2

The first-order derivatives of the log-likelihood function for a multivariate probit model are obtained as follows. First, we express the multivariate normal cdf Φ_M in terms of multivariate integrals. Then, by using conditional density distributions, we decompose ϕ_M into a product of two normal probability density functions (pdfs) and re-express Φ_M based on that decomposition. In doing so we proceed with the calculation of the two derivatives, where the derivative of Φ_M with respect to $\boldsymbol{\beta}_m$ is mainly based on a decomposition formula, while the derivative of Φ_M with respect to ϑ_{zk} has been derived by applying an idea by Plackett (1954).

The multivariate integrals

$$\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \;=\; \int_{-\infty}^{w_{M,i}} \dots \int_{-\infty}^{w_{1,i}} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \tag{11}$$

can be written in a more convenient form by using the conditional distribution of the normal multivariate distribution. This can be achieved by partitioning both $\boldsymbol{l}_i$ and $\boldsymbol{\Upsilon}_i^*$ such that

$$\boldsymbol{l}_i \;=\; (\boldsymbol{l}_{1,i}, \boldsymbol{l}_{2,i})^\top,$$

and

$$\boldsymbol{\Upsilon}_i^* = \left(\begin{array}{c|c} \boldsymbol{\Theta}_{11,i}^* & \boldsymbol{\Theta}_{12,i}^* \\ \hline \boldsymbol{\Theta}_{21,i}^* & \boldsymbol{\Theta}_{22,i}^* \end{array}\right) = \left(\begin{array}{cccc|ccc} 1 & r_{12,i}^* & \cdots & r_{1u,i}^* & r_{1,u+1,i}^* & \cdots & r_{1,M,i}^* \\ r_{21,i}^* & 1 & \cdots & r_{2u,i}^* & r_{2,u+1,i}^* & \cdots & r_{2,M,i}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ r_{u1,i}^* & r_{u2,i}^* & \cdots & 1 & r_{u,u+1,i}^* & \cdots & r_{u,M,i}^* \\ \hline r_{u+1,1,i}^* & r_{u+1,2,i}^* & \cdots & r_{u+1,u,i}^* & 1 & \cdots & r_{u+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{M1,i}^* & r_{M2,i}^* & \cdots & r_{Mu,i}^* & r_{M,u+1,i}^* & \cdots & 1 \end{array}\right), \quad (12)$$

respectively, where $\boldsymbol{l}_{1,i} = (l_{1,i}, l_{2,i}, \ldots, l_{u,i})^\top$, $\boldsymbol{l}_{2,i} = (l_{u+1,i}, l_{u+2,i}, \ldots, l_{M,i})^\top$, $u = 1, \ldots, M-1$, $r_{zk,i}^* = \tanh(\vartheta_{zk}^*)(2y_{zi} - 1)(2y_{ki} - 1)$, $\boldsymbol{\Theta}_{11,i}^*$ is a $u \times u$ matrix, $\boldsymbol{\Theta}_{22,i}^*$ is a $(M-u) \times (M-u)$ matrix and $\boldsymbol{\Theta}_{21,i}^* = \boldsymbol{\Theta}_{12,i}^{*\top}$. By using the chain rule for random variables and the partitioned vector $\boldsymbol{l}_i$ as well as the partitioned matrix $\boldsymbol{\Upsilon}_i^*$, the M-variate normal pdf $\phi_M(\boldsymbol{l}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}_i^*)$ can be expressed as the product of the conditional density function of $\boldsymbol{l}_{2,i}$ given $\boldsymbol{l}_{1,i}$ times the pdf of $\boldsymbol{l}_{1,i}$

$$\phi_M(\boldsymbol{l}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}_i^*) \;=\; \phi_{M-u}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i})\phi_u(\boldsymbol{l}_{1,i}), \quad (13)$$

where

$$\begin{aligned} \boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i} &\overset{iid}{\sim} \mathcal{N}_M(\mathbb{E}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}), \mathrm{Var}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i})) \\ &\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} + \boldsymbol{\Theta}_{21,i}^*\boldsymbol{\Theta}_{11}^{*-1}(\boldsymbol{l}_{1,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}), \boldsymbol{\Theta}_{22,i}^* - \boldsymbol{\Theta}_{21,i}^*\boldsymbol{\Theta}_{11,i}^{*-1}\boldsymbol{\Theta}_{12,i}^*), \end{aligned} \quad (14)$$

and

$$\begin{aligned} \boldsymbol{l}_{1,i} &\overset{iid}{\sim} \mathcal{N}_M(\mathbb{E}(\boldsymbol{l}_{1,i}), \mathrm{Var}(\boldsymbol{l}_{1,i})) \\ &\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}_{11,i}^*). \end{aligned} \quad (15)$$

$\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}$ and $\boldsymbol{\mu}_{\boldsymbol{l}_{2,i}}$ stand for the mean of $\boldsymbol{l}_{1,i}$ and $\boldsymbol{l}_{2,i}$ respectively. It follows that the integrals (11) can be rewritten as

$$\begin{aligned}\Phi_M(\boldsymbol{w}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) &= \int_{-\infty}^{w_{M,i}} \dots \int_{-\infty}^{w_{1,i}} \phi_{M-u}(\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}; \boldsymbol{M}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}}) \phi_u(\boldsymbol{l}_{1,i}; \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}}, \boldsymbol{\Theta}_{11,i}^*) \\ &\quad \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}, \end{aligned} \tag{16}$$

where $\boldsymbol{M}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}} = \boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} + \boldsymbol{\Theta}_{21,i}^* \boldsymbol{\Theta}_{11}^{*-1}(\boldsymbol{l}_{1,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{1,i}})$, $\boldsymbol{\Theta}_i^{*\boldsymbol{l}_{2,i}|\boldsymbol{l}_{1,i}} = \boldsymbol{\Theta}_{22,i}^* - \boldsymbol{\Theta}_{21,i}^* \boldsymbol{\Theta}_{11,i}^{*-1} \boldsymbol{\Theta}_{12,i}^*$, $\boldsymbol{\mu}_{\boldsymbol{l}_{1,i}} = \boldsymbol{\mu}_{\boldsymbol{l}_{2,i}} = \mathbf{0}$ and $\boldsymbol{\Theta}_{11,i}^*$ denotes the $u \times u$ sub-matrix of $\boldsymbol{\Upsilon}_i^*$.

F.1: Proof of Proposition 2

Proof. Consider formula (13) and let $u = 1$, such that

$$\phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) = \phi_{M-1}(\boldsymbol{l}_{2,i}|l_{1,i}) \phi(l_{1,i}).$$

By re-ordering matrix (12) we obtain

$$\boldsymbol{\Upsilon}_i^{*m} = \begin{pmatrix} \overbrace{\Theta_{11,i}^{*m}}^{1\times 1} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*m}}^{1\times(M-1)} \\ \underbrace{\boldsymbol{\Theta}_{21,i}^{*m}}_{(M-1)\times 1} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*m}}_{(M-1)\times(M-1)} \end{pmatrix},$$

$\forall m$, where $\Theta_{11,i}^{*m}$, $\boldsymbol{\Theta}_{12,i}^{*m}$, $\boldsymbol{\Theta}_{21,i}^{*m}$ and $\boldsymbol{\Theta}_{22,i}^{*m}$ are defined in Proposition 2, while the full matrix $\boldsymbol{\Upsilon}_i^{*m}$ can be found in Supplementary Material G. Then the multivariate normal cdf (16) becomes

$$\begin{aligned}\Phi_M(\boldsymbol{w}_i; 0, \boldsymbol{\Upsilon}_i^{*m}) &= \int_{-\infty}^{w_{M,i}} \dots \int_{-\infty}^{w_{m,i}} \dots \int_{-\infty}^{w_{1,i}} \phi_M(\boldsymbol{l}_i; 0, \boldsymbol{\Upsilon}_i^{*m}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \\ &= \int_{\bar{\mathbf{C}}_i} \phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) \phi(l_{m,i}; \mu_{l_{m,i}}, \Theta_{11,i}^{*m}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}, \end{aligned} \tag{17}$$

for $\bar{\mathbf{C}}_i = \bar{C}_{1i} \times \bar{C}_{2i} \times \ldots \times \bar{C}_{Mi}$, where $\bar{C}_{mi}$ is the interval $[w_{m,i}, +\infty)$ if $y_{mi} = 1$ and the interval $(-\infty, w_{m,i}]$ if $y_{mi} = 0$. Vector $\boldsymbol{l}_{-m,i} = (l_{1,i}, \ldots, l_{m-1,i}, l_{m+1,i}, \ldots, l_{M,i})^\top$, where $l_{m,i}$ refers to the m^{th} element of vector $\boldsymbol{l}_i$. $\boldsymbol{M}_i^{*m}$ and $\boldsymbol{\Theta}_i^{*m}$, respectively, denote the mean and the variance of $\boldsymbol{l}_{-m,i}|l_{m,i}$, while $\mu_{l_{m,i}}$ and $\Theta_{11,i}^{*m}$ denote the mean and variance of $l_{m,i}$. Applying the properties of the conditional multivariate normal distribution, it follows that $\mathbb{E}(l_{m,i}) = \mu_{l_{m,i}} = 0$ and $\mathbb{E}(\boldsymbol{l}_{-m,i}) = \boldsymbol{\mu}_{\boldsymbol{l}_{-m,i}} = \mathbf{0}$. (Note that $\mathbb{E}(\boldsymbol{l}_{-m,i}|l_{m,i}) \neq \mathbf{0}$.) Hence, the distribution of $\boldsymbol{l}_{-m,i}|l_{m,i}$ and $l_{m,i}$ is equal to

$$\begin{aligned}
\boldsymbol{l}_{-m,i}|l_{m,i} &\overset{iid}{\sim} \mathcal{N}_M\left(\mathbb{E}(\boldsymbol{l}_{-m,i}|l_{m,i}), \mathrm{Var}(\boldsymbol{l}_{-m,i}|l_{m,i})\right) \\
&\overset{iid}{\sim} \mathcal{N}_M\left(\boldsymbol{\mu}_{\boldsymbol{l}_{-m,i}} + \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11}^{*m}\right)^{-1}(l_{m,i} - \mu_{l_{m,i}}), \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11,i}^{*m}\right)^{-1}\boldsymbol{\Theta}_{12,i}^{*m}\right) \\
&\overset{iid}{\sim} \mathcal{N}_M\left(\boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11}^{*m}\right)^{-1} l_{m,i}, \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\left(\Theta_{11,i}^{*m}\right)^{-1}\boldsymbol{\Theta}_{12,i}^{*m}\right),
\end{aligned}$$

and

$$\begin{aligned}
l_{m,i} &\overset{iid}{\sim} \mathcal{N}(\mathbb{E}(l_{m,i}), \mathrm{Var}(l_{m,i})) \\
&\overset{iid}{\sim} \mathcal{N}(\mu_{l_{m,i}}, \Theta_{11,i}^{*m}) \\
&\overset{iid}{\sim} \mathcal{N}(0, \Theta_{11,i}^{*m}),
\end{aligned}$$

respectively, where the sub-matrix $\Theta_{11,i}^{*m}$ in this case is equal to 1, $\forall m, i$. It follows that (17) becomes

$$\begin{aligned}
\Phi_M(\boldsymbol{w}_i; 0, \boldsymbol{\Upsilon}_i^{*m}) &= \int_{\bar{\mathbf{C}}_i} \phi(l_{m,i}; 0, 1)\phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \\
&= \int_{-\infty}^{w_{m,i}} \phi(l_{m,i}; 0, 1) \left\{ \int_{\bar{\mathbf{C}}_{i,-m}} \phi_{M-1}(\boldsymbol{l}_{-m,i}|l_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) d\boldsymbol{l}_{-m,i} \right\} dl_{m,i} \\
&= \int_{-\infty}^{w_{m,i}} \phi(l_{m,i}; 0, 1)\Phi_{M-1}(\boldsymbol{w}_{-m,i}|l_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) dl_{m,i}, \qquad (18)
\end{aligned}$$

where $\boldsymbol{w}_{-m,i} = (w_{1,i}, w_{2,i}, \ldots, w_{m-1,i}, w_{m+1,i}, \ldots, w_{M,i})^\top$ and $\bar{\mathbf{C}}_{i,-m} \in \{\bar{\mathbf{C}}_i\} \setminus \bar{C}_{mi}$. According to the properties of the conditional multivariate normal distribution, it follows that the expected value of $\boldsymbol{w}_{-m,i}|l_{m,i}$ is equal to $\boldsymbol{M}_i^{*m} = \boldsymbol{\Theta}_{21,i}^{*m} l_{m,i}$ while its variance-covariance matrix is expressed as $\boldsymbol{\Theta}_i^{*m} = \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\boldsymbol{\Theta}_{12,i}^{*m}$. By using the chain rule as well as the fundamental theorem of calculus, it follows that the derivative of (18) with respect to $\boldsymbol{\beta}_m$ is equal to

$$\begin{aligned}
\frac{\partial \Phi_M(\boldsymbol{w}_i; 0, \boldsymbol{\Upsilon}_i^{*m})}{\partial \boldsymbol{\beta}_m} &= \frac{\partial \Phi_M(\boldsymbol{w}_i; 0, \boldsymbol{\Theta}_i^{*m})}{\partial w_{m,i}} \frac{\partial w_{m,i}}{\partial \boldsymbol{\beta}_m} \\
&= \frac{\partial}{\partial w_{m,i}} \left\{ \int_{-\infty}^{w_{m,i}} \phi(l_{m,i}; 0, 1) \Phi_{M-1}(\boldsymbol{w}_{-m,i}|l_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) dl_{m,i} \right\} \left(\frac{\partial w_{m,i}}{\partial \boldsymbol{\beta}_m} \right) \\
&= \phi(w_{m,i}; 0, 1) \Phi_{M-1}(\boldsymbol{w}_{-m,i}|w_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m}) \left(\frac{\partial w_{m,i}}{\partial \boldsymbol{\beta}_m} \right).
\end{aligned}$$

Since $w_{m,i} = (2y_{mi} - 1)\mathbf{x}_{mi}^\top \boldsymbol{\beta}_m$, the derivative of $w_{m,i}$ with respect to $\boldsymbol{\beta}_m$ is equal to

$$\frac{\partial w_{m,i}}{\partial \boldsymbol{\beta}_m} = (2y_{mi} - 1)\mathbf{x}_{mi}^\top,$$

and thus

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*m})}{\partial \boldsymbol{\beta}_m} = \phi(w_{m,i}; 0, 1) \Phi_{M-1}(\boldsymbol{w}_{-m,i}|w_{m,i}; \boldsymbol{M}_i^{*m}, \boldsymbol{\Theta}_i^{*m})(2y_{mi} - 1)\mathbf{x}_{mi}^\top,$$

for $\boldsymbol{M}_i^{*m} = \boldsymbol{\Theta}_{21,i}^{*m} w_{m,i}$ and $\boldsymbol{\Theta}_i^{*m} = \boldsymbol{\Theta}_{22,i}^{*m} - \boldsymbol{\Theta}_{21,i}^{*m}\boldsymbol{\Theta}_{12,i}^{*m}$, as required. □

F.2: Proof of Proposition 3

Proof. If we differentiate equation (11) with respect to the correlation coefficient ϑ_{zk}^*, we get the following

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk}^*} = \frac{\partial}{\partial \vartheta_{zk}^*} \left\{ \int_{-\infty}^{w_{M,i}} \cdots \int_{-\infty}^{w_{1,i}} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i} \right\},$$

and by using the chain rule

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk,i}^*} &= \frac{\partial}{\partial r_{zk}^*}\left\{\int_{\bar{\mathbf{C}}_i} \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}\right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*} \\ &= \left\{\int_{\bar{\mathbf{C}}_i} \frac{\partial \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial r_{zk,i}^*} \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}\right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \end{aligned} \tag{19}$$

where $r_{zk,i}^*$ and region $\bar{\mathbf{C}}_i$ have been defined previously. By using the following differential equation derived by Plackett (1954)

$$\frac{\partial \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial r_{zk,i}^*} = \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}},$$

equation (19) becomes

$$\begin{aligned}\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk,i}^*} &= \left\{\int_{\bar{\mathbf{C}}_i} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}} \prod_{\tilde{c}=1}^{M} dl_{\tilde{c},i}\right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*} \\ &= \left\{\int_{\bar{\mathbf{C}}_{-zk,i}} \left[\int_{\bar{\mathbf{C}}_{zk,i}} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}} d\boldsymbol{l}_{zk,i}\right] d\boldsymbol{l}_{-zk,i}\right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \end{aligned} \tag{20}$$

where $\bar{\mathbf{C}}_{-zk,i} \in \bar{\mathbf{C}}_i \setminus \{\bar{C}_{zi}, \bar{C}_{ki}\}$, $\bar{\mathbf{C}}_{zk,i} = \bar{C}_{zi} \times \bar{C}_{ki}$, $\boldsymbol{l}_{zk,i} = (l_{z,i}, l_{k,i})^\top$ and $\boldsymbol{l}_{-zk,i} = (l_{1,i}, \ldots, l_{k-1,i}, l_{k+1,i}, \ldots, l_{z-1,i}, l_{z+1,i}, l_{M,i})^\top$. According to the fundamental theorem of calculus, the integral inside the brackets is equal to

$$\begin{aligned}\int_{\bar{\mathbf{C}}_{zk,i}} \frac{\partial^2 \phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial l_{z,i} \partial l_{k,i}} d\boldsymbol{l}_{zk,i} &= \frac{\partial^2}{\partial l_{z,i} \partial l_{k,i}}\left\{\int_{\bar{\mathbf{C}}_{zk,i}} \phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) dl_{k,i} dl_{z,i}\right\} \\ &= \frac{\partial^2}{\partial l_{z,i} \partial l_{k,i}}\left\{\int_{\bar{C}_{zi}} \int_{\bar{C}_{ki}} \phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) dl_{k,i} dl_{z,i}\right\} \\ &= \phi_M(l_{1,i}, \ldots, l_{z-1,i}, w_{z,i}, l_{z+1,i}, \ldots, l_{k-1,i}, w_{k,i}, l_{k+1,i}, \ldots, l_{M,i}; \mathbf{0}, \\ &\quad \boldsymbol{\Upsilon}_i^*). \end{aligned}$$

Therefore, (20) can be expressed as

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*)}{\partial \vartheta_{zk,i}^*} = \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_M(l_{1,i}, \ldots, l_{z-1,i}, w_{z,i}, l_{z+1,i}, \ldots, l_{k-1,i}, w_{k,i}, l_{k+1,i}, \ldots, l_{M,i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^*) d\boldsymbol{l}_{-zk,i} \right\} \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}.$$

The last expression can be written in a more convenient form by using the conditional distributions of the normal multivariate distribution. This can be done by imposing the special case $u = 2$ in equation (13), that is

$$\phi_M(\boldsymbol{l}_i; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk}) = \phi_{M-2}(\boldsymbol{l}_{2,i} | \boldsymbol{l}_{1,i}) \phi_2(\boldsymbol{l}_{1,i}), \tag{21}$$

where $\boldsymbol{l}_{2,i}$ and $\boldsymbol{l}_{1,i}$ correspond to $\boldsymbol{l}_{-zk,i}$ and $\boldsymbol{w}_{zk,i}$, respectively, with $\boldsymbol{w}_{zk,i} = (w_{z,i}, w_{k,i})^\top$. Re-ordering matrix (12), we obtain

$$\boldsymbol{\Upsilon}_i^{*zk} = \begin{pmatrix} \overbrace{\boldsymbol{\Theta}_{11,i}^{*zk}}^{2\times 2} & \overbrace{\boldsymbol{\Theta}_{12,i}^{*zk}}^{2\times(M-2)} \\ \underbrace{\boldsymbol{\Theta}_{21,i}^{*zk}}_{(M-2)\times 2} & \underbrace{\boldsymbol{\Theta}_{22,i}^{*zk}}_{(M-2)\times(M-2)} \end{pmatrix}, \tag{22}$$

$\forall z = 1, \ldots, M-1, k = z+1, \ldots M$, where the sub-matrices $\boldsymbol{\Theta}_{11,i}^{*zk}$, $\boldsymbol{\Theta}_{12,i}^{*zk}$, $\boldsymbol{\Theta}_{21,i}^{*zk}$ and $\boldsymbol{\Theta}_{22,i}^{*zk}$ are defined in Proposition 3, while the full matrix $\boldsymbol{\Upsilon}_i^{*zk}$ can be found in Supplementary Material G. By using both (21) and (22), we have that

$$\frac{\partial \Phi_M(\boldsymbol{w_i}; \mathbf{0}, \boldsymbol{\Upsilon}_i^{*zk})}{\partial \vartheta_{zk,i}^*} = \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_{M-2}(\boldsymbol{l}_{-zk,i} | \boldsymbol{w}_{zk,i}; \boldsymbol{M}_i^{*-zk}, \boldsymbol{\Theta}_i^{*-zk}) \phi_2(\boldsymbol{w}_{zk,i}; \boldsymbol{\mu}_{\boldsymbol{w}_{zk,i}}, \boldsymbol{\Theta}_{11,i}^{*zk}) d\boldsymbol{l}_{-zk,i} \right\} \times \frac{\partial r_{zk,i}^*}{\partial \vartheta_{zk}^*}, \tag{23}$$

where $\boldsymbol{M}_i^{*-zk}$ and $\boldsymbol{\Theta}_i^{*-zk}$ refer to the mean and variance-covariance matrix of $\boldsymbol{l}_{-zk,i} | \boldsymbol{w}_{zk,i}$, while $\boldsymbol{\mu}_{\boldsymbol{w}_{zk,i}}$ and $\boldsymbol{\Theta}_{11,i}^{*zk}$ denote the mean and variance-covariance of $\boldsymbol{w}_{zk,i}$. By using the properties of the conditional multivariate normal distribution, it follows that $\mathbb{E}(\boldsymbol{w}_{zk,i}) =$

$\boldsymbol{\mu}_{\boldsymbol{w}_{zk,i}} = 0$ and $\mathbb{E}(\boldsymbol{l}_{-zk,i}) = \boldsymbol{\mu}_{\boldsymbol{l}_{-zk,i}} = 0$. (Note that $\mathbb{E}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}) \neq 0$.) Hence, according to (14) and (15)

$$\begin{aligned}
\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i} &\overset{iid}{\sim} \mathcal{N}_M(\mathbb{E}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}), \mathrm{Var}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i})) \\
&\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\mu}_{\boldsymbol{l}_{-zk,i}} + \boldsymbol{\Theta}^{*zk}_{21,i}\left(\boldsymbol{\Theta}^{*zk}_{11}\right)^{-1}(\boldsymbol{w}_{zk,i} - \boldsymbol{\mu}_{\boldsymbol{l}_{zk,i}}), \boldsymbol{\Theta}^{*zk}_{22,i} - \boldsymbol{\Theta}^{*zk}_{21,i}\left(\boldsymbol{\Theta}^{*zk}_{11,i}\right)^{-1}\boldsymbol{\Theta}^{*zk}_{12,i}) \\
&\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{\Theta}^{*zk}_{21,i}\left(\boldsymbol{\Theta}^{*zk}_{11,i}\right)^{-1}\boldsymbol{w}_{zk,i}, \boldsymbol{\Theta}^{*zk}_{22,i} - \boldsymbol{\Theta}^{*zk}_{21,i}\left(\boldsymbol{\Theta}^{zk}_{11,i}\right)^{-1}\boldsymbol{\Theta}^{*zk}_{12,i}),
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{l}_{zk,i} &\overset{iid}{\sim} \mathcal{N}_M(\mathbb{E}(\boldsymbol{w}_{zk,i}), \mathrm{Var}(\boldsymbol{w}_{zk,i})) \\
&\overset{iid}{\sim} \mathcal{N}_M(\mu_{\boldsymbol{w}_{zk,i}}, \boldsymbol{\Theta}^{*zk}_{11,i}) \\
&\overset{iid}{\sim} \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{\Theta}^{*zk}_{11,i}),
\end{aligned}$$

where the sub-matrix $\boldsymbol{\Theta}^{*zk}_{11,i}$ is a 2×2 diagonal matrix with unit variances and correlations equal to $r^*_{zk,i}$. For simplicity, we will denote this matrix as $\boldsymbol{\Theta}^{*zk}_i$. Consequently, equation (23) can be expressed as

$$\begin{aligned}
\frac{\partial \Phi_M(\boldsymbol{w}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^{*zk}_i)}{\partial \vartheta^*_{zk}} &= \left\{ \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_2(\boldsymbol{w}_{zk,i}; \boldsymbol{0}, \boldsymbol{\Theta}^{*zk}_i) \phi_{M-2}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}; \boldsymbol{M}^{*-zk}_i, \boldsymbol{\Theta}^{*-zk}_i) d\boldsymbol{l}_{-zk,i} \right\} \times \\
&\quad \frac{\partial r^*_{zk,i}}{\partial \vartheta^*_{zk}}.
\end{aligned}$$

Because only the term $\phi_{M-2}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}; \boldsymbol{M}^{*-zk}_i, \boldsymbol{\Theta}^{*-zk}_i)$ depends on $\boldsymbol{l}_{-zk,i}$, it follows that

$$\begin{aligned}
\frac{\partial \Phi_M(\boldsymbol{w}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^{*zk}_i)}{\partial \vartheta^*_{zk}} &= \left\{ \phi_2(\boldsymbol{w}_{zk,i}; \boldsymbol{0}, \boldsymbol{\Theta}^{*zk}_i) \int_{\bar{\mathbf{C}}_{-zk,i}} \phi_{M-2}(\boldsymbol{l}_{-zk,i}|\boldsymbol{w}_{zk,i}; \boldsymbol{M}^{*-zk}_i, \boldsymbol{\Theta}^{*-zk}_i) d\boldsymbol{l}_{-zk,i} \right\} \times \\
&\quad \frac{\partial r^*_{zk,i}}{\partial \vartheta^*_{zk}} \\
&= \left\{ \phi_2(\boldsymbol{w}_{zk,i}; \boldsymbol{0}, \boldsymbol{\Theta}^{*zk}_i) \Phi_{M-2}(\boldsymbol{w}_{-zk,i}|\boldsymbol{w}_{zk,i}; \boldsymbol{M}^{*-zk}_i, \boldsymbol{\Theta}^{*-zk}_i) \right\} \frac{\partial r^*_{zk,i}}{\partial \vartheta^*_{zk}}, \qquad (24)
\end{aligned}$$

where the last term comes from basic results of the multivariate normal distribution function. In addition, $\boldsymbol{w}_{-zk,i} = (w_{1,i}, w_{2,i}, \ldots, w_{z-1,i}, w_{z+1,i}, \ldots, w_{k-1,i}, w_{k+1,i}, \ldots w_{M,i})^\top$, while the partial derivative $\partial r^*_{zk,i}/\partial \vartheta^*_{zk}$ is equal to

$$\begin{aligned}
\frac{\partial r^*_{zk,i}}{\partial \vartheta^*_{zk}} &= \frac{\partial}{\partial \vartheta^*_{zk}} \{\tanh(\vartheta^*_{zk})(2y_{z,i}-1)(2y_{k,i}-1)\} \\
&= (2y_{z,i}-1)(2y_{k,i}-1)\frac{\partial}{\partial \vartheta^*_{zk}} \{\tanh(\vartheta^*_{zk})\} \\
&= (2y_{z,i}-1)(2y_{k,i}-1)\mathrm{sech}^2(\vartheta^*_{zk}) \\
&= (2y_{z,i}-1)(2y_{k,i}-1)\frac{1}{\cosh^2(\vartheta^*_{zk})} \\
&= (2y_{z,i}-1)(2y_{k,i}-1)\frac{1}{\left(\frac{\exp(\vartheta^*_{zk})+\exp(-\vartheta^*_{zk})}{2}\right)^2} \\
&= (2y_{z,i}-1)(2y_{k,i}-1)\frac{4}{\{\exp(\vartheta^*_{zk})+\exp(-\vartheta^*_{zk})\}^2} \\
&= (2y_{z,i}-1)(2y_{k,i}-1)\frac{4e^{2\vartheta^*_{zk}}}{\left\{e^{2\vartheta^*_{zk}}+1\right\}^2},
\end{aligned}$$

by using definitions and properties of the hyperbolic functions. Therefore, (24) becomes

$$\begin{aligned}
\frac{\partial \Phi_M(\boldsymbol{w}_i; \boldsymbol{0}, \boldsymbol{\Upsilon}^{*zk}_i)}{\partial \vartheta_{*zk}} &= \phi_2(\boldsymbol{w}_{zk,i}; \boldsymbol{0}, \boldsymbol{\Theta}^{*zk}_i)\Phi_{M-2}(\boldsymbol{w}_{-zk,i}|\boldsymbol{w}_{zk,i}; \boldsymbol{M}^{*-zk}_i, \boldsymbol{\Theta}^{*-zk}_i)(2y_{z,i}-1)\times \\
&\quad (2y_{k,i}-1)\frac{4e^{2\vartheta^*_{zk}}}{\left\{e^{2\vartheta^*_{zk}}+1\right\}^2},
\end{aligned}$$

for $\boldsymbol{w}_{zk,i} = (w_{z,i}, w_{k,i})^\top$, $\boldsymbol{w}_{-zk,i} = (w_{1,i}, w_{2,i}, \ldots, w_{z-1,i}, w_{z+1,i}, \ldots, w_{k-1,i}, w_{k+1,i}, \ldots w_{M,i})^\top$, $\boldsymbol{\Theta}^{*zk}_i = \boldsymbol{\Theta}^{*zk}_{11,i}$, $\boldsymbol{M}^{*-zk}_i = \boldsymbol{\Theta}^{*zk}_{21,i}\left(\boldsymbol{\Theta}^{*zk}_{11,i}\right)^{-1}\boldsymbol{w}_{zk,i}$ and $\boldsymbol{\Theta}^{*-zk}_i = \boldsymbol{\Theta}^{*zk}_{22,i} - \boldsymbol{\Theta}^{*zk}_{21,i}\left(\boldsymbol{\Theta}^{*zk}_{11,i}\right)^{-1}\boldsymbol{\Theta}^{*zk}_{12,i}$, as required. □

Supplementary Material G: Correlation matrices $\boldsymbol{\Upsilon}_i^{*m}$ and $\boldsymbol{\Upsilon}_i^{*zk}$

For $m = 1$, matrix $\boldsymbol{\Upsilon}_i^{*m}$ is equal to

$$\boldsymbol{\Upsilon}_i^{*1} = \left(\begin{array}{c|c} \Theta_{11,i}^{*1} & \boldsymbol{\Theta}_{12,i}^{*1} \\ \hline \boldsymbol{\Theta}_{21,i}^{*1} & \boldsymbol{\Theta}_{22,i}^{*1} \end{array}\right) = \left(\begin{array}{c|cccccc} 1 & r_{12,i}^* & r_{13,i}^* & \cdots & r_{1,M-1,i}^* & r_{1M,i}^* \\ \hline r_{12,i}^* & 1 & r_{23,i}^* & \cdots & r_{2,M-1,i}^* & r_{2M,i}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{1,M-1,i}^* & r_{2,M-1,i} & r_{3,M-1,i}^* & \cdots & 1 & r_{M-1,M,i}^* \\ r_{1M,i}^* & r_{2M,i}^* & r_{3M,i}^* & \cdots & r_{M-1,M,i}^* & 1 \end{array}\right),$$

while for $m \geq 2$

$$\begin{aligned} \boldsymbol{\Upsilon}_i^{*m} &= \left(\begin{array}{c|c} \Theta_{11,i}^{*m} & \boldsymbol{\Theta}_{12,i}^{*m} \\ \hline \boldsymbol{\Theta}_{21,i}^{*m} & \boldsymbol{\Theta}_{22,i}^{*m} \end{array}\right) \\ &= \left(\begin{array}{c|ccccccc} 1 & r_{m1,i}^* & r_{m2,i}^* & \cdots & r_{m,m-1,i}^* & r_{m,m+1,i}^* & \cdots & r_{mM,i}^* \\ \hline r_{m1,i}^* & 1 & r_{12,i}^* & \cdots & r_{1,m-1,i}^* & r_{1,m+1,i}^* & \cdots & r_{1M,i}^* \\ r_{m2,i}^* & r_{12,i}^* & 1 & \cdots & r_{2,m-1,i}^* & r_{2,m+1,i}^* & \cdots & r_{2M,i}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ r_{m,m-1,i}^* & r_{1,m-1,i} & r_{2,m-1,i}^* & \cdots & 1 & r_{m-1,m+1,i}^* & \cdots & r_{m-1,m,i}^* \\ r_{m,m+1,i}^* & r_{1,m+1,i}^* & r_{2,m+1,i}^* & \cdots & r_{m-1,m+1,i}^* & 1 & \cdots & r_{m+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{mM,i}^* & r_{1M,i}^* & r_{2M,i}^* & \cdots & r_{m-1,M,i}^* & r_{m+1,M,i}^* & \cdots & 1 \end{array}\right). \end{aligned}$$

Matrix $\boldsymbol{\Upsilon}_i^{zk}$ is equal to

$$\boldsymbol{\Upsilon}_i^{*zk} = \left(\begin{array}{c|c} \boldsymbol{\Theta}_{11,i}^{*zk} & \boldsymbol{\Theta}_{12,i}^{*zk} \\ \hline \boldsymbol{\Theta}_{21,i}^{*zk} & \boldsymbol{\Theta}_{22,i}^{*zk} \end{array}\right) = \left(\begin{array}{cc:cccccccccc} 1 & r_{zk,i}^* & r_{z1,i}^* & r_{z2,i}^* & \cdots & r_{z,z-1,i}^* & r_{z,z+1,i}^* & \cdots & r_{z,k-1,i}^* & r_{z,k+1}^* & \cdots & r_{zM,i}^* \\ r_{zk,i}^* & 1 & r_{k1,i}^* & r_{k2,i}^* & \cdots & r_{k,z-1,i}^* & r_{k,z+1,i}^* & \cdots & r_{k,k-1,i}^* & r_{k,k+1,i}^* & \cdots & r_{kM,i}^* \\ \hdashline r_{z1,i}^* & r_{k1,i}^* & 1 & r_{12,i}^* & \cdots & r_{1,z-1,i}^* & r_{1,z+1,i}^* & \cdots & r_{1,k-1,i}^* & r_{1,k+1,i}^* & \cdots & r_{1M,i}^* \\ r_{z2,i}^* & r_{k2,i}^* & r_{12,i}^* & 1 & \cdots & r_{2,z-1,i}^* & r_{2,z+1,i}^* & \cdots & r_{2,k-1,i}^* & r_{2,k+1,i}^* & \cdots & r_{2M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{z,z-1,i}^* & r_{k,z-1,i}^* & r_{1,z-1,i}^* & r_{2,z-1,i}^* & \cdots & 1 & r_{z-1,z+1,i}^* & \cdots & r_{z-1,k-1,i}^* & r_{z-1,k+1,i}^* & \cdots & r_{z-1,M,i}^* \\ r_{z,z+1,i}^* & r_{k,z+1,i}^* & r_{1,z+1,i}^* & r_{2,z+1,i}^* & \cdots & r_{z-1,z+1,i}^* & 1 & \cdots & r_{z+1,k-1,i}^* & r_{z+1,k+1,i}^* & \cdots & r_{z+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ r_{z,k-1,i}^* & r_{k,k-1,i}^* & r_{1,k-1,i}^* & r_{2,k-1,i}^* & \cdots & r_{z-1,k-1,i}^* & r_{z+1,k-1,i}^* & \cdots & 1 & r_{k-1,k+1,i}^* & \cdots & r_{k-1,M,i}^* \\ r_{z,k+1,i}^* & r_{k,k+1,i}^* & r_{1,k+1,i}^* & r_{2,k+1,i}^* & \cdots & r_{z-1,k+1,i}^* & r_{z+1,k+1,i}^* & \cdots & r_{k-1,k+1,i}^* & 1 & \cdots & r_{k+1,M,i}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{z,M,i}^* & r_{k,M,i}^* & r_{1M,i}^* & r_{2M,i}^* & \cdots & r_{z-1,M,i}^* & r_{z+1,M,i}^* & \cdots & r_{k-1,M,i}^* & r_{k+1,M,i}^* & \cdots & 1 \end{array}\right).$$

Supplementary Material H: Multiple smoothing parameter estimation

There are several ways for estimating automatically multiple smoothing parameters (e.g., Wood, 2004; Radice et al., 2016; Marra et al., 2016). One way is to minimise a mean squared error criterion which can be shown to be equivalent to an approximate Akaike Information Criterion (AIC). In this work, we adopt this idea as well as a parametrisation of the smoothing criterion discussed by Marra et al. (2016) which makes estimation more stable and efficient.

Assume that, near the solution, the trust region method behaves as a classic unconstrained Newton-Raphson algorithm (Nocedal & Wright, 2006). Also, suppose that $\boldsymbol{\delta}^{[\varkappa+1]}$ is the 'true' parameter value, and thus $\boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa+1]}) = \mathbf{0}$. By using a Taylor expansion for $\boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa+1]})$ at $\boldsymbol{\delta}^{[\varkappa]}$ it follows that $\mathbf{0} = \boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa+1]}) \approx \boldsymbol{g}_p(\boldsymbol{\delta}^{[\varkappa]}) + \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\varkappa]})(\boldsymbol{\delta}^{[\varkappa+1]} - \boldsymbol{\delta}^{[\varkappa]})$. Solving for $\boldsymbol{\delta}^{[\varkappa+1]}$ yields, after some manipulation,

$$\boldsymbol{\delta}^{[\varkappa+1]} \;=\; \left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right)^{-1} \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\bar{\mathbf{z}}^{[\varkappa]}, \tag{25}$$

where $\boldsymbol{\mathcal{I}}^{[\varkappa]} = -\boldsymbol{\mathcal{H}}^{[\varkappa]}$ and $\bar{\mathbf{z}}^{[\varkappa]} = \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\boldsymbol{\delta}^{[\varkappa]} + \bar{\boldsymbol{\epsilon}}^{[\varkappa]}$ with $\bar{\boldsymbol{\epsilon}}^{[\varkappa]} = \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}^{-1}\boldsymbol{g}^{[\varkappa]}$. From standard likelihood theory $\bar{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\mathbf{z}} \sim \mathcal{N}(\boldsymbol{\mu}_{\bar{\mathbf{z}}}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix, $\boldsymbol{\mu}_{\bar{\mathbf{z}}} = \sqrt{\boldsymbol{\mathcal{I}}}\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_0$ is the true parameter vector. As shown below, representation (25) allows us to estimate the smoothing parameters based on a parametrization of $\bar{\mathbf{z}}$ that uses $\boldsymbol{g}$ and $\boldsymbol{\mathcal{H}}$ as a whole instead of the n components that makes them up. As argued by Marra et al. (2016), this is advantageous in estimation problems involving simultaneous systems of equations.

Now let $\hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}$ be the predicted value vector for $\bar{\mathbf{z}}$ defined as $\hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}} = \mathbf{C}_{\hat{\boldsymbol{\lambda}}}\bar{\mathbf{z}}$ where $\mathbf{C}_{\hat{\boldsymbol{\lambda}}} = \sqrt{\boldsymbol{\mathcal{I}}}\left(\boldsymbol{\mathcal{I}} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}}$, the influence matrix or hat matrix of the fitting problem which depends on the smoothing parameter vector. An appealing way of estimating $\boldsymbol{\lambda}$ is to minimise the distance between $\hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}$ and the truth $\boldsymbol{\mu}_{\bar{\mathbf{z}}}$. This can be achieved using

$$\mathbb{E}(\|\boldsymbol{\mu}_{\bar{\mathbf{z}}} - \hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}\|^2) \;=\; \mathbb{E}(\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\|^2) - \tilde{n} + 2\mathrm{tr}(\mathbf{C}_{\boldsymbol{\lambda}}), \tag{26}$$

where $\tilde{n} = 6n$ and $\text{tr}(\mathbf{C}_{\boldsymbol{\lambda}})$ is the number of estimated degrees of freedom (*edf*) of the penalized model which measures the flexibility of the fitted model. The overall *edf* is defined as the sum of the *edf* of the smooth functions. Note that the RHS of (26) depends on the smoothing parameter through $\mathbf{C}_{\boldsymbol{\lambda}}$, while $\bar{\mathbf{z}}$ is associated with the un-penalized part of the model. In practice, smoothing parameters are selected by minimizing an estimate of (26), that is

$$\mathcal{V}(\boldsymbol{\lambda}) = \|\widehat{\boldsymbol{\mu}_{\bar{\mathbf{z}}} - \hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}}\|^2 = \|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\|^2 - \tilde{n} + 2\text{tr}(\mathbf{C}_{\boldsymbol{\lambda}}),$$

which is approximately equivalent to the AIC, defined as $2\text{tr}(\mathbf{C}_{\boldsymbol{\lambda}}) - 2\ell(\hat{\boldsymbol{\delta}})$, where $-2\ell(\hat{\boldsymbol{\delta}})$ can be approximated as $\approx -2\ell(\boldsymbol{\delta}) - \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2$. Given $\boldsymbol{\delta}^{[\varkappa+1]}$, the estimation problem can be expressed as

$$\boldsymbol{\lambda}^{[\varkappa+1]} \;=\; \arg\min_{\boldsymbol{\lambda}} \mathcal{V}(\boldsymbol{\lambda}) := \|\bar{\mathbf{z}}^{[\varkappa+1]} - \mathbf{C}_{\boldsymbol{\lambda}}^{[\varkappa+1]}\bar{\mathbf{z}}^{[\varkappa+1]}\|^2 - \tilde{n} + 2\text{tr}(\mathbf{C}_{\boldsymbol{\lambda}}^{[\varkappa+1]}),$$

which is solved by using adapting the approach by Wood (2004) to the current context. This method implements a stable and efficient Newton method for estimating $\log(\boldsymbol{\lambda})$. Working with the logarithm of $\boldsymbol{\lambda}$ ensures that the smoothing parameter estimates are positive. The derivation of the above results is provided below.

Derivation of (25)

For notational convenience we denote $\boldsymbol{g}(\boldsymbol{\delta}^{[\cdot]})$ as $\boldsymbol{g}^{[\cdot]}$, $\boldsymbol{g}_p(\boldsymbol{\delta}^{[\cdot]})$ as $\boldsymbol{g}_p^{[\cdot]}$, $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}^{[\cdot]})$ as $\boldsymbol{\mathcal{H}}^{[\cdot]}$ and $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}^{[\cdot]})$ as $\boldsymbol{\mathcal{H}}_p^{[\cdot]}$.

By using the Taylor series expansion for $\boldsymbol{g}_p^{[\varkappa+1]}$ at $\boldsymbol{\delta}^{[\varkappa]}$ we have that $\mathbf{0} = \boldsymbol{g}_p^{[\varkappa+1]} \approx \boldsymbol{g}_p^{[\varkappa]} + \boldsymbol{\mathcal{H}}_p^{[\varkappa]}(\boldsymbol{\delta}^{[\varkappa+1]} - \boldsymbol{\delta}^{[\varkappa]})$, where $\boldsymbol{g}_p^{[\varkappa]} = \boldsymbol{g}^{[\varkappa]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}^{[\varkappa]}$ and $\boldsymbol{\mathcal{H}}_p^{[\varkappa]} = \boldsymbol{\mathcal{H}}^{[\varkappa]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}$. Suppose that $\boldsymbol{\mathcal{I}}^{[\varkappa]} = -\boldsymbol{\mathcal{H}}^{[\varkappa]}$; then we have that

$$\mathbf{0} \;=\; \boldsymbol{g}_p^{[\varkappa]} + \left(\boldsymbol{\delta}^{[\varkappa+1]} - \boldsymbol{\delta}^{[\varkappa]}\right)\left(-\boldsymbol{\mathcal{I}}^{[\varkappa]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right).$$

Re-arranging the above equation we get

$$\begin{aligned} \boldsymbol{g}_p^{[\varkappa]} &= \left(\boldsymbol{\delta}^{[\varkappa+1]} - \boldsymbol{\delta}^{[\varkappa]}\right)\left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right) \implies \\ \implies \boldsymbol{g}^{[\varkappa]} - \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}^{[\varkappa]} &= \boldsymbol{\delta}^{[\varkappa+1]}\left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right) - \boldsymbol{\delta}^{[\varkappa]}\boldsymbol{\mathcal{I}}^{[\varkappa]} - \boldsymbol{\delta}^{[\varkappa]}\tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}} \implies \\ \implies \boldsymbol{\delta}^{[\varkappa+1]}\left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right) &= \boldsymbol{g}^{[\varkappa]} + \boldsymbol{\delta}^{[\varkappa]}\boldsymbol{\mathcal{I}}^{[\varkappa]} \implies \\ \implies \boldsymbol{\delta}^{[\varkappa+1]} &= \left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\left(\sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\boldsymbol{\delta}^{[\varkappa]} + \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}^{-1}\boldsymbol{g}^{[\varkappa]}\right). \end{aligned}$$

Therefore, the parameter estimator can be expressed as

$$\boldsymbol{\delta}^{[\varkappa+1]} = \left(\boldsymbol{\mathcal{I}}^{[\varkappa]} + \tilde{\mathbf{S}}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\bar{\mathbf{z}}^{[\varkappa]},$$

where $\bar{\mathbf{z}}^{[\varkappa]} = \boldsymbol{\mu}_{\bar{\mathbf{z}}}^{[\varkappa]} + \bar{\boldsymbol{\epsilon}}^{[\varkappa]}$, $\boldsymbol{\mu}_{\bar{\mathbf{z}}}^{[\varkappa]} = \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}\boldsymbol{\delta}^{[\varkappa]}$ and $\bar{\boldsymbol{\epsilon}}^{[\varkappa]} = \sqrt{\boldsymbol{\mathcal{I}}^{[\varkappa]}}^{-1}\boldsymbol{g}^{[\varkappa]}$, as required.

Derivation of (26)

Based on the notation in Section 3.2, we have that

$$\begin{aligned} \mathbb{E}(\|\boldsymbol{\mu}_{\bar{\mathbf{z}}} - \hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}\|^2) &= \mathbb{E}\left(\left\|\left(\bar{\mathbf{z}} - \bar{\boldsymbol{\epsilon}}\right) - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2\right) \\ &= \mathbb{E}\left(\left\|\left(\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right) - \bar{\boldsymbol{\epsilon}}\right\|^2\right) \\ &= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2 + \bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}} - 2\left\|\left(\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right)\bar{\boldsymbol{\epsilon}}\right\|\right) \\ &= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2 + \bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}} - 2\left\|\left\{\boldsymbol{\mu}_{\bar{\mathbf{z}}} + \bar{\boldsymbol{\epsilon}} - \mathbf{C}_{\boldsymbol{\lambda}}(\boldsymbol{\mu}_{\bar{\mathbf{z}}} + \bar{\boldsymbol{\epsilon}})\right\}\bar{\boldsymbol{\epsilon}}\right\|\right) \\ &= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2 + \bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}} - 2\left\|\boldsymbol{\mu}_{\bar{\mathbf{z}}}\bar{\boldsymbol{\epsilon}} + \bar{\boldsymbol{\epsilon}}^2 - \mathbf{C}_{\boldsymbol{\lambda}}\boldsymbol{\mu}_{\bar{\mathbf{z}}}\bar{\boldsymbol{\epsilon}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\boldsymbol{\epsilon}}^2)\right\|\right) \\ &= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2\right) + \mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}}\right) - 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\boldsymbol{\mu}_{\bar{\mathbf{z}}}\right) - \\ &\quad 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}}\right) + 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\boldsymbol{\mu}_{\bar{\mathbf{z}}}\right) + 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\bar{\boldsymbol{\epsilon}}\right) \\ &= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}}\bar{\mathbf{z}}\right\|^2\right) - \mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\bar{\boldsymbol{\epsilon}}\right) - 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\boldsymbol{\mu}_{\bar{\mathbf{z}}}\right) + \\ &\quad 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\boldsymbol{\mu}_{\bar{\mathbf{z}}}\right) + 2\mathbb{E}\left(\bar{\boldsymbol{\epsilon}}^\top\mathbf{C}_{\boldsymbol{\lambda}}\bar{\boldsymbol{\epsilon}}\right). \end{aligned}$$

By using the following results (e.g., Wood, 2006, Section 1.8.5)

$$\begin{aligned}
\mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top \bar{\boldsymbol{\epsilon}}) &= \mathbb{E}\left(\sum_i \bar{\epsilon}_i^2\right) = \tilde{n} \cdot 1 = \tilde{n}, \text{ for } \tilde{n} = 6n, \\
\mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top \boldsymbol{\mu}_{\bar{\mathbf{z}}}) &= \mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top)\boldsymbol{\mu}_{\bar{\mathbf{z}}} = \mathbf{0}, \\
\mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top \mathbf{C}_{\boldsymbol{\lambda}} \bar{\boldsymbol{\epsilon}}) &= \mathbb{E}\left(\text{tr}(\bar{\boldsymbol{\epsilon}}^\top \mathbf{C}_{\boldsymbol{\lambda}} \bar{\boldsymbol{\epsilon}})\right), \text{ since a scalar is its own trace} \\
&= \text{tr}\left(\mathbf{C}_{\boldsymbol{\lambda}} \mathbb{E}(\bar{\boldsymbol{\epsilon}}^\top \bar{\boldsymbol{\epsilon}})\right) \\
&= \text{tr}\left(\mathbf{C}_{\boldsymbol{\lambda}} \mathbf{I}\right) \cdot 1 \\
&= \text{tr}\left(\mathbf{C}_{\boldsymbol{\lambda}}\right),
\end{aligned}$$

it follows that

$$\begin{aligned}
\mathbb{E}(\|\boldsymbol{\mu}_{\bar{\mathbf{z}}} - \hat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}}\|^2) &= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}} \bar{\mathbf{z}}\right\|^2\right) - \tilde{n} - 2 \cdot \mathbf{0} + 2 \cdot \mathbf{0} + 2\text{tr}\left(\mathbf{C}_{\hat{\boldsymbol{\lambda}}}\right) \\
&= \mathbb{E}\left(\left\|\bar{\mathbf{z}} - \mathbf{C}_{\boldsymbol{\lambda}} \bar{\mathbf{z}}\right\|^2\right) - \tilde{n} + 2\text{tr}\left(\mathbf{C}_{\boldsymbol{\lambda}}\right),
\end{aligned}$$

as required.

Equivalence of $\mathcal{V}(\boldsymbol{\lambda})$ and AIC

The AIC of a model can be defined as follows

$$\text{AIC} = 2Q - 2\ell(\hat{\boldsymbol{\delta}}),$$

where Q is the number of estimated parameters in the model.

Consider a Taylor expansion of $-2\ell(\hat{\boldsymbol{\delta}})$ about $\boldsymbol{\delta}$

$$\begin{aligned}
-2\ell(\hat{\boldsymbol{\delta}}) &\approx -2\ell(\boldsymbol{\delta}) + (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \nabla_{\boldsymbol{\delta}} \left\{-2\ell(\boldsymbol{\delta})\right\} + \frac{1}{2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \nabla\nabla_{\boldsymbol{\delta}} \left\{-2\ell(\boldsymbol{\delta})\right\} (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \\
&\approx -2\ell(\boldsymbol{\delta}) - 2(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{g} - (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{\mathcal{H}} (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}), \qquad (27)
\end{aligned}$$

where $\boldsymbol{g} := \boldsymbol{g}(\boldsymbol{\delta})$ and $\boldsymbol{\mathcal{H}} := \boldsymbol{\mathcal{H}}(\boldsymbol{\delta})$. By using $\boldsymbol{\mathcal{I}} = -\boldsymbol{\mathcal{H}}$, we have that

$$\begin{aligned} -(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{\mathcal{H}} (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) &= (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{\mathcal{I}} (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \\ &= \| \sqrt{\boldsymbol{\mathcal{I}}} \left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} \right) \|^2 \\ &= \| \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} - \boldsymbol{\mathcal{I}} \boldsymbol{\delta} \|^2, \end{aligned}$$

and by applying $\bar{\mathbf{z}} = \sqrt{\boldsymbol{\mathcal{I}}} \boldsymbol{\delta} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g}$ to the above expression we get

$$\begin{aligned} -(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{\mathcal{H}} (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) &= \| \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} - \bar{\mathbf{z}} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \|^2 \\ &= \| - \left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} - \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right) \|^2 \\ &= \| \left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} \right) - \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \|^2 \qquad (28) \\ &= \left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}}, \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} \right\rangle - \\ &\quad 2 \left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right\rangle + \left\langle \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right\rangle \\ &= \| \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} \|^2 - 2 \left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right\rangle + \| \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \|^2, \qquad (29) \end{aligned}$$

where (28) results from the fact that $\| - \tilde{\boldsymbol{\chi}} \|^2 = \| \tilde{\boldsymbol{\chi}} \|^2$, for any vector $\tilde{\boldsymbol{\chi}}$. Similarly, by using the expression for the pseudo-data vector $\bar{\mathbf{z}}$ in the second term in (27) we have

$$\begin{aligned} (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top \boldsymbol{g} &= \left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} \right)^\top \sqrt{\boldsymbol{\mathcal{I}}} \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \\ &= \left(\sqrt{\boldsymbol{\mathcal{I}}} (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \\ &= \left(\sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} - \sqrt{\boldsymbol{\mathcal{I}}} \boldsymbol{\delta} \right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \\ &= \left(\sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} - \bar{\mathbf{z}} + \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \\ &= - \left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} - \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right) \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \\ &= - \left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} \right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} + \left(\sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right)^\top \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \\ &= - \left(\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}} \right)^\top \bullet \left(\sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right) + \boldsymbol{g}^\top \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{g} \\ &= - \left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}} \hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \right\rangle + \| \sqrt{\boldsymbol{\mathcal{I}}}^{-1} \boldsymbol{g} \|^2. \qquad (30) \end{aligned}$$

Substituting both (29) and (30) in (27), we obtain

$$\begin{aligned}
-2\ell(\hat{\boldsymbol{\delta}}) &\approx -2\ell(\boldsymbol{\delta}) - 2\left\{-\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle + \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2\right\} + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2 - \\
&\quad 2\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle + \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 \\
&\approx -2\ell(\boldsymbol{\delta}) + 2\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle - 2\|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2 - \\
&\quad 2\left\langle \bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}, \sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\right\rangle + \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 \\
&\approx -2\ell(\boldsymbol{\delta}) - \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2,
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. It follows that

$$\begin{aligned}
\text{AIC} &\approx 2Q - 2\ell(\boldsymbol{\delta}) - \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2 \\
&\approx 2\text{tr}(\mathbf{C}_{\boldsymbol{\lambda}}) - 2\ell(\boldsymbol{\delta}) - \|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2 + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2, \qquad (31)
\end{aligned}$$

where $\text{tr}(\mathbf{C}_{\boldsymbol{\lambda}})$ denotes the number of estimated parameters in the model and thus $Q = \text{tr}(\mathbf{C}_{\boldsymbol{\lambda}})$. Since we are interested in optimizing a criterion with respect to the smoothing parameter $\boldsymbol{\lambda}$, we drop any terms that are not affected by $\boldsymbol{\lambda}$, i.e., $-2\ell(\boldsymbol{\delta})$ and $-\|\sqrt{\boldsymbol{\mathcal{I}}}^{-1}\boldsymbol{g}\|^2$. Therefore (31) becomes

$$\text{AIC} \approx 2\text{tr}(\mathbf{C}_{\lambda}) + \|\bar{\mathbf{z}} - \sqrt{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{\delta}}\|^2.$$

H.1: Details on the result $\boldsymbol{\delta} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1}\right)$

The rationale for using using the result $\boldsymbol{\delta} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\delta}}, -\hat{\boldsymbol{\mathcal{H}}}_p^{-1}\right)$ for the construction of CIs is provided in Marra & Wood (2012) in a related context, whereas some examples of interval construction are given in Radice et al. (2016). For general smooth models, such as the one considered in this paper, this result can be justified using the distribution of $\bar{\mathbf{z}}$, making the large sample assumption that $\boldsymbol{\mathcal{I}}$ can be treated as fixed, and making the usual Bayesian assumption on the prior of $\boldsymbol{\delta}$ for smooth models (e.g., Silverman, 1985; Wood, 2006). Note that

this result neglects smoothing parameter uncertainty. However, as argued by Marra & Wood (2012) this is not problematic provided that heavy oversmoothing is avoided (so that the bias is not too large a proportion of the sampling variability) and in our experience we found that this result works well in practice (see Section 4.2.2 for some simulation-based evidence). The problem of testing smooth components for equality to zero can be addressed using the results discussed in Wood (2013a) and Wood (2013b).

Supplementary Material I: Data generating processes (DGPs) used in the simulation studies

I.1: Simulation study I (DGP1 & DGP2)

Both DGP1 and DGP2 were based on the following trivariate system of equations

$$\begin{aligned} y_{1i}^* &= 1.6 + 0.9v_{1i} - 1.3z_{1i} + \varepsilon_{1i} \\ y_{2i}^* &= -1.0 - 1.4v_{1i} + 1.0z_{1i} + \varepsilon_{2i} \\ y_{3i}^* &= -1.4 + 2.0v_{1i} - 1.5z_{1i} + \varepsilon_{3i}, \end{aligned}$$

where $\boldsymbol{\varepsilon}_i \sim (\mathbf{0}, \boldsymbol{\Sigma})$ and v_{mi} and z_{mi}, $\forall m$, denote a binary regressor and a continuous covariate, respectively. DGP1 is fully parametric and was used for comparing `mvprobit()` and `SemiParTRIV()`. The correlation parameters were set as ($\vartheta_{12} = -0.8$, $\vartheta_{13} = -0.6$, $\vartheta_{23} = 0.8$). These values were obtained after fitting the trivariate probit model on the North Carolina data set used for the case study in Section 5. DGP2 was based on the following set of correlations ($\vartheta_{12} = -0.1$, $\vartheta_{13} = 0.3$, $\vartheta_{23} = 0.9$), which was selected while trying out different combinations of values; this choice seemed to be problematic from an estimation perspective as convergence was not achieved in most of the replicates used in the simulation study, and the estimates of the correlation parameters were not close to the true values. For each set-up, we generated 250 datasets with sample sizes equal to 1000 and 10000. Note that the responses were unbalanced (similarly as in the case study). Specifically, responses y_{1i}, y_{2i} and y_{3i} had typical observed value 1 proportions of 90.5%, 15.1% and 21.5%, respectively.

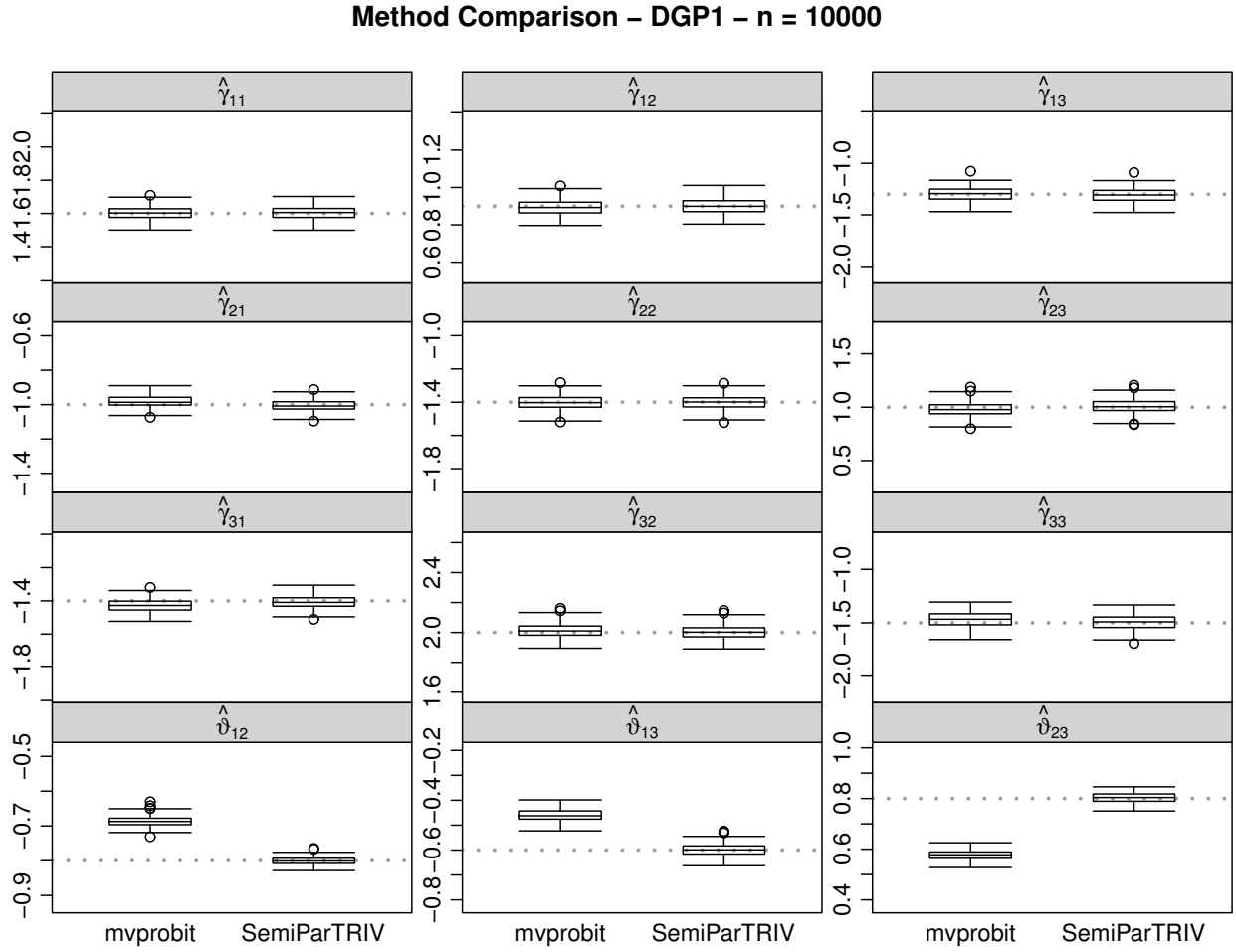

Figure 2: Boxplots of parameter estimates obtained using `mvprobit()` and `SemiParTRIV()` on 250 datasets simulated using the settings described in Supplementary Material I.1. The sample size was equal to 10000 and the true parameter values are represented by horizontal gray dotted lines.

The unsatisfactory performance of `mvprobit()` in estimating the correlation parameters may be attributed to the method used for evaluating normal trivariate integrals, namely the Geweke-Hadjivassiliou-Keane (GHK) smooth recursive simulator (Geweke, 1991; Hajivassiliou & McFadden, 1991; Keane, 1990). Broadly speaking, the GHK approach first applies a Cholesky decomposition on the model's correlation matrix and then expresses the trivariate integrals as a product of three univariate probabilities based on truncated standard normal variables; Trinh & Genz (2015) introduced similar approximations which were found not to yield satisfactory results for highly correlated responses. Furthermore, Cappellari & Jenkins (2003) pointed out that if the correlation matrix obtained at a given

iteration of the optimization is not positive-definite then the GHK method uses the most recent positive-definite estimate of the correlation matrix; this runs the risk of delivering estimates that are far from the optimal values. When we tried different scenarios with higher and lower values for the correlation coefficients, we found that the higher the correlations the worse the estimation results.

I.1.1: STATA and R code for DGP1

```
# Generate some data using STATA:
set obs 1000
matrix Er = (1,  -0.8, -0.6 \ -0.8, 1, 0.8 \ -0.6, 0.8, 1)
forvalues i = 1/250{
set seed ‘i’
drawnorm er1‘i’ er2‘i’ er3‘i’, corr(Er)
matrix C = (1, .5\ .5, 1)
drawnorm x1‘i’ x2‘i’, corr(C)
gen v1‘i’ = round(normal(x1‘i’))
gen z1‘i’ = normal(x2‘i’)
gen y1‘i’ = ( 1.6 + 0.9 * v1‘i’ - 1.3 * z1‘i’ + er1‘i’>0)
gen y2‘i’ = (-1.0 - 1.4 * v1‘i’ + 1.0 * z1‘i’ + er2‘i’>0)
gen y3‘i’ = (-1.4 + 2.0 * v1‘i’ - 1.5 * z1‘i’ + er3‘i’>0)
}

# Fit the model via the mvprobit routine in STATA:
ssc install mvprobit
forvalues i = 1/250{
capture mvprobit( y1‘i’ = v1‘i’ z1‘i’)(y2‘i’ = v1‘i’ z1‘i’)
(y3‘i’ = v1‘i’ z1‘i’)
```

```
matrix estparams‘i’ = e(b)
* creates a matrix of the parameter estimates
}

# Fit the model via the SemiParTRIV routine in R:
library(SemiParBIVProbit)
library(foreign)

SimsSTATADGP2 <- read.dta("SimsDGP2.dta") # extracts the
                                          # simulation
                                          # STATA file
gamma11 <- gamma12 <- gamma13 <- gamma21 <- gamma22 <- NULL
gamma23 <- gamma31 <- gamma32 <- gamma33 <- theta12 <- NULL
theta13 <- theta23 <- NULL

n.rep <- 250 # number of replicates
n <- 1000
p <- 10 # number of variables generated in STATA, i.e.,
        # er1, er2, er3, x1, x2, v1, z1, y1, y2, y3

eqn1 <- y1 ~ v1 + z1
eqn2 <- y2 ~ v1 + z1
eqn3 <- y3 ~ v1 + z1
f.l <- list(eqn1, eqn2, eqn3)

for(i in 1:n.rep){
  j <- ifelse(i>0, i+1, 1)
```

```
  DataSTATADGP2 <- SimsSTATADGP2[1:n, (i * p +1):(j * p)]
  v1 <- DataSTATADGP2[1:n, 6]
  z1 <- DataSTATADGP2[1:n, 7]
  y1 <- DataSTATADGP2[1:n, 8]
  y2 <- DataSTATADGP2[1:n, 9]
  y3 <- DataSTATADGP2[1:n, 10]
  data <- DataSTATADGP2
  out <- SemiParTRIV(f.l, data = data)
  X1.d2 <- out$X1.d2 # number of columns in the design matrix
                     # of first equation
  X2.d2 <- out$X2.d2 # number of columns in the design matrix
                     # of second equation
  X3.d2 <- out$X3.d2 # number of columns in the design matrix
                     # of third equation
  gamma11[i] <- out$fit$argument[1]
  gamma12[i] <- out$fit$argument[2]
  gamma13[i] <- out$fit$argument[X1.d2]
  gamma21[i] <- out$fit$argument[X1.d2 + 1]
  gamma22[i] <- out$fit$argument[X1.d2 + 2]
  gamma23[i] <- out$fit$argument[X1.d2 + X2.d2]
  gamma31[i] <- out$fit$argument[X1.d2 + X2.d2 + 1]
  gamma32[i] <- out$fit$argument[X1.d2 + X2.d2 + 2]
  gamma33[i] <- out$fit$argument[X1.d2 + X2.d2 + X3.d2]
  theta12[i] <- out$theta12
  theta13[i] <- out$theta13
  theta23[i] <- out$theta23
}
```

```
# Note: for sample size 10000 we replace set obs 1000 with
# set obs 10000 in  the STATA code and n <- 1000 with
# n <- 10000 in the R code.
```

I.1.2: R code for DGP2

```
library(SemiParBIVProbit)


theta12.sim <- -0.1
theta13.sim <-  0.3
theta23.sim <-  0.9
n.rep <- 250
n <- 1000 # then n <- 10000
Sigma.er <- matrix( c( 1, theta12.sim, theta13.sim,
                       theta12.sim, 1, theta23.sim,
                       theta13.sim, theta23.sim, 1), 3 , 3)
theta.cov <- 0.5
SigmaCov       <- matrix(theta.cov, 2, 2)
diag(SigmaCov) <- 1
f.l <- list(y1 ~ v1 + z1, y2 ~ v1 + z1, y3 ~ v1 + z1 )
gamma11 <- gamma12 <- gamma13 <- gamma21 <- gamma22 <- NULL
gamma23 <- gamma31 <- gamma32 <- gamma33 <- theta12 <- NULL
theta13 <- theta23 <- NULL


for(i in 1:n.rep){
  set.seed(i)
  er  <- rMVN(n, rep(0,3), Sigma.er)
```

```
cov <- rMVN(n, rep(0,2), SigmaCov)
cov <- pnorm(cov)
v1 <- round(cov[,1])
z1 <- cov[,2]
y1 <- ifelse(1.6  + 0.9 * v1 - 1.3 * z1 + er[,1] > 0, 1, 0)
y2 <- ifelse(-1.0 - 1.4 * v1 + 1.0 * z1 + er[,2] > 0, 1, 0)
y3 <- ifelse(-1.4 + 2.0 * v1 - 1.5 * z1 + er[,3] > 0, 1, 0)
dataSim <- data.frame(y1, y2, y3, v1, z1)
out <- SemiParTRIV(f.l, data = dataSim) # penCor = "lasso"
                                        # or penCor = "ridge"
                                        # or penCor = "alasso"
                                        # for penalised
                                        # correlation study

X1.d2 <- out$X1.d2
X2.d2 <- out$X2.d2
X3.d2 <- out$X3.d2
gamma11[i] <- coef(out)[1]
gamma12[i] <- coef(out)[2]
gamma13[i] <- coef(out)[X1.d2]
gamma21[i] <- coef(out)[X1.d2 + 1]
gamma22[i] <- coef(out)[X1.d2 + 2]
gamma23[i] <- coef(out)[X1.d2 + X2.d2]
gamma31[i] <- coef(out)[X1.d2 + X2.d2 + 1]
gamma32[i] <- coef(out)[X1.d2 + X2.d2 + 2]
gamma33[i] <- coef(out)[X1.d2 + X2.d2 + X3.d2]
theta12[i] <- out$theta12
```

```
  theta13[i] <- out$theta13
  theta23[i] <- out$theta23
}
```

I.2: Simulation study II (DGP3)

The trivariate system of equations was based on

$$\begin{aligned} y_{1i}^* &= 1.05 + 0.90v_{1i} + s_1(z_{1i}) + \varepsilon_{1i} \\ y_{2i}^* &= -1.45 - 1.40v_{1i} + s_2(z_{1i}) + \varepsilon_{2i} \\ y_{3i}^* &= -1.60 + 2.00v_{1i} + s_3(z_{1i}) + \varepsilon_{3i} \end{aligned}$$

where $\boldsymbol{\varepsilon}_i \sim (\mathbf{0}, \boldsymbol{\Sigma})$ and s_m, for all m, corresponds to the smooth component which was represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives. The correlation parameters were set to the same values as those used for DGP2, while the smooth functions are given by $s_1(z_{1i}) = 0.5\cos(2\pi z_{1i})$, $s_2(z_{1i}) = z_{1i} + \exp\{-30(z_{1i} - 0.5)^2\}$ and $s_3(z_{1i}) = -0.5(z_{1i} + 3z_{1i}^3)$. The other settings are similar to those described in Supplementary Material I.1. For each replicate and fitted model the estimated smooth functions were evaluated at 200 fixed values in the ranges of the respective covariates. Parameter estimation was carried out using a Lasso-type penalty for the correlations, i.e. $\boldsymbol{\Gamma}_{\boldsymbol{\lambda}}^{\mathrm{L}} = \tilde{\boldsymbol{S}}_{\boldsymbol{\lambda}} + \boldsymbol{\Lambda}_{\lambda_{\vartheta^*}}^{\mathrm{L}}$; using Ridge and Adaptive Lasso did led to virtually identical results.

I.2.1: `R` code for DGP3

```
library(SemiParBIVProbit)


# Simulate some data:
```

```
n     <- 1000 # then n <- 10000
n.rep <- 250
theta12.sim <- -0.1
theta13.sim <-  0.3
theta23.sim <-  0.9
Sigma.er <- matrix( c( 1, theta12.sim, theta13.sim,
                       theta12.sim, 1, theta23.sim,
                       theta13.sim, theta23.sim, 1 ), 3 , 3)
SigmaCov <- matrix(0.5, 2, 2)
diag(SigmaCov) <- 1
f.l <- list(y1 ~ v1 + s(z1), y2 ~ v1 + s(z1), y3 ~ v1 + s(z1) )
F1 <- F2 <- F3 <- matrix(NA, 200, n.rep)
theta12 <- theta13 <- theta23 <- NULL
# smooth functions
f1 <- function(x) 0.5*cos(pi*2*x)
f2 <- function(x) x+exp(-30*(x-0.5)^2)
f3 <- function(x) -0.5*(x+3*x^3)
xt <- seq(0.0000001, 0.9999999, length.out = 200) # grid to evaluate
smooth functions
dt <- data.frame(z = xt)
f1t <- f1(xt) - mean(f1(xt))
f2t <- f2(xt) - mean(f2(xt))
f3t <- f3(xt) - mean(f3(xt))

for(i in 1:n.rep){
set.seed(i)
u   <- rMVN(n, rep(0,3), Sigma.er)
```

```
cov <- rMVN(n, rep(0,2), SigmaCov)
cov <- pnorm(cov)
v1  <- round(cov[, 1])
z1  <- cov[, 2]
y1 <- ifelse( 1.05 + 0.9*v1  + f1(z1) + u[,1]  > 0, 1, 0)
y2 <- ifelse(-1.45 - 1.4*v1  + f2(z1) + u[,2]  > 0, 1, 0)
y3 <- ifelse(-1.6  + 2.0*v1  + f3(z1) + u[,3]  > 0, 1, 0)
dataSim <- data.frame(y1, y2, y3, v1, z1)
out <- SemiParTRIV(f.l, data = dataSim, penCor = "lasso")
X1 <- PredictMat( out$gam1$smooth[[1]], dt )
X2 <- PredictMat( out$gam2$smooth[[1]], dt )
X3 <- PredictMat( out$gam3$smooth[[1]], dt )
lg1 <- length(coef(out$gam1))
lg2 <- length(coef(out$gam2))
F1[,i] <- X1%*%
coef(out)[(out$gam1$smooth[[1]]$first.para:out$gam1$
smooth[[1]]$last.para)]
F2[,i] <- X2%*%
coef(out)[lg1 + (out$gam2$smooth[[1]]$first.para:out$
gam2$smooth[[1]]$
last.para)]
F3[,i] <- X3%*%
coef(out)[lg1 + lg2 + (out$gam3$smooth[[1]]$first.para:out$
gam3$smooth[[1]]
$last.para)]
F1[,i] <- F1[,i] - mean(F1[,i])
F2[,i] <- F2[,i] - mean(F2[,i])
```

```
F3[,i] <- F3[,i] - mean(F3[,i])

theta12[i] <- out$theta12

theta13[i] <- out$theta13

theta23[i] <- out$theta23

}
```

Supplementary Material J: Correlation–based penalty

J.1: The penalty functions

$$
\begin{aligned}
\textbf{Lasso: } \mathcal{P}^{\mathrm{L}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &= \mathcal{P}^{\mathrm{L}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_1) \\
&= \lambda_{\vartheta^*}\|\mathbf{R}_q\boldsymbol{\delta}\|_1 \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}|\mathbf{R}_q\boldsymbol{\delta}| \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\left\{(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta}\right\}^{1/2} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\left\{\left(\boldsymbol{e}_q^\top\boldsymbol{\delta}\right)^2\right\}^{1/2} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}|\boldsymbol{e}_q^\top\boldsymbol{\delta}| \\
&= \lambda_{\vartheta^*}\left\{|\boldsymbol{e}_{Q-2}^\top\boldsymbol{\delta}| + |\boldsymbol{e}_{Q-1}^\top\boldsymbol{\delta}| + |\boldsymbol{e}_{Q}^\top\boldsymbol{\delta}|\right\} \\
&= \lambda_{\vartheta^*}\left(|\vartheta^*_{12}| + |\vartheta^*_{13}| + |\vartheta^*_{23}|\right),
\end{aligned}
$$

where $\boldsymbol{e}_q = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ with a one at the q^{th} position, $\forall q$.

$$
\begin{aligned}
\textbf{Ridge: } \mathcal{P}^{\text{R}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &= \frac{1}{2}\mathcal{P}^{\text{R}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_2^2) \\
&= \frac{1}{2}\lambda_{\vartheta^*}\|\mathbf{R}_q\boldsymbol{\delta}\|_2^2 \\
&= \frac{1}{2}\lambda_{\vartheta^*}\left\{\left[\sum_{q=Q-2}^{Q}\left((\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta}\right)\right]^{1/2}\right\}^2 \\
&= \frac{1}{2}\lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\left((\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta}\right) \\
&= \frac{1}{2}\lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\left(\boldsymbol{e}_q^\top\boldsymbol{\delta}\right)^2 \\
&= \frac{1}{2}\lambda_{\vartheta^*}\left\{(\boldsymbol{e}_{Q-2}^\top\boldsymbol{\delta})^2 + (\boldsymbol{e}_{Q-1}^\top\boldsymbol{\delta})^2 + (\boldsymbol{e}_{Q}^\top\boldsymbol{\delta})^2\right\} \\
&= \frac{1}{2}\lambda_{\vartheta^*}\left(\vartheta_{12}^{*2} + \vartheta_{13}^{*2} + \vartheta_{23}^{*2}\right).
\end{aligned}
$$

$$
\begin{aligned}
\textbf{Ad. Lasso: } \mathcal{P}^{\text{AL}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &= \mathcal{P}^{\text{AL}}_{\lambda_{\vartheta^*}}(\|\mathbf{R}_q\boldsymbol{\delta}\|_1) \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\frac{|\mathbf{R}_q\boldsymbol{\delta}|}{|\mathbf{R}_q\hat{\boldsymbol{\delta}}^{\text{MLE}}|^{\bar{\gamma}}} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\frac{\left\{\left(\mathbf{R}_q\boldsymbol{\delta}\right)^\top\mathbf{R}_q\boldsymbol{\delta}\right\}^{1/2}}{\left\{\left(\mathbf{R}_q\hat{\boldsymbol{\delta}}^{\text{MLE}}\right)^\top\mathbf{R}_q\hat{\boldsymbol{\delta}}^{\text{MLE}}\right\}^{\bar{\gamma}/2}} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\frac{\left\{\left(\boldsymbol{e}_q^\top\boldsymbol{\delta}\right)^2\right\}^{1/2}}{\left\{\left(\boldsymbol{e}_q^\top\hat{\boldsymbol{\delta}}^{\text{MLE}}\right)^2\right\}^{\bar{\gamma}/2}} \\
&= \lambda_{\vartheta^*}\sum_{q=Q-2}^{Q}\frac{|\boldsymbol{e}_q^\top\boldsymbol{\delta}|}{|\boldsymbol{e}_q^\top\hat{\boldsymbol{\delta}}^{\text{MLE}}|^{\bar{\gamma}}} \\
&= \lambda_{\vartheta^*}\left\{\frac{|\boldsymbol{e}_{Q-2}^\top\boldsymbol{\delta}|}{|\boldsymbol{a}_{Q-2}^\top\hat{\boldsymbol{\delta}}^{\text{MLE}}|^{\bar{\gamma}}} + \frac{|\boldsymbol{e}_{Q-1}^\top\boldsymbol{\delta}|}{|\boldsymbol{e}_{Q-1}^\top\hat{\boldsymbol{\delta}}^{\text{MLE}}|^{\bar{\gamma}}} + \frac{|\boldsymbol{e}_{Q}^\top\boldsymbol{\delta}|}{|\boldsymbol{a}_{Q}^\top\hat{\boldsymbol{\delta}}^{\text{MLE}}|^{\bar{\gamma}}}\right\} \\
&= \lambda_{\vartheta^*}\left(\frac{|\vartheta_{12}^*|}{|\hat{\vartheta}_{12}^{*\text{MLE}}|^{\bar{\gamma}}} + \frac{|\vartheta_{13}^*|}{|\hat{\vartheta}_{13}^{*\text{MLE}}|^{\bar{\gamma}}} + \frac{|\vartheta_{23}^*|}{|\hat{\vartheta}_{23}^{*\text{MLE}}|^{\bar{\gamma}}}\right).
\end{aligned}
$$

J.2: LQA of the penalty function $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$

The approximated penalty functions for both Lasso and Adaptive Lasso belong to the L_1-type family. Based on (18) and by applying the chain rule, it follows that $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ can be written as

$$\begin{aligned}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \nabla_{\tilde{\boldsymbol{\delta}}}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})^\top(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}) \\ &\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \frac{\partial \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})}{\partial \|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\cdot\frac{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}{\partial \mathbf{R}_q\tilde{\boldsymbol{\delta}}}\cdot\frac{\partial \mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\partial\tilde{\boldsymbol{\delta}}^\top}\cdot\left(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right). \end{aligned} \tag{32}$$

By using the local approximation $(\mathbf{R}_q\boldsymbol{\delta})^\top / \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top \approx 1$ for $\tilde{\boldsymbol{\delta}} \approx \boldsymbol{\delta}$ (Fan & Li, 2001) as well as the following approximation (Ulbricht, 2010)

$$\begin{aligned}(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\left(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}}\right) &= (\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta} - (\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}} \\ &= \frac{1}{2}\left\{(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta} - 2\,(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}} + \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right\} + \\ &\quad \frac{1}{2}\left\{(\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta} - \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right\} \\ &= \frac{1}{2}\left(\mathbf{R}_q^\top(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}})^\top(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}})\mathbf{R}_q\right) + \\ &\quad \frac{1}{2}\left((\mathbf{R}_q\boldsymbol{\delta})^\top\mathbf{R}_q\boldsymbol{\delta} - \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right) \\ &\approx \frac{1}{2}(\boldsymbol{\delta}^\top\mathbf{R}_q^\top\mathbf{R}_q\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}^\top\mathbf{R}_q^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}),\end{aligned}$$

equation (32) becomes

$$\begin{aligned}
\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) &\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top} \cdot \left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top \cdot \mathbf{R}_q \cdot \left(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right) \\
&\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top} \cdot \frac{1}{2}(\boldsymbol{\delta}^\top\mathbf{R}_q^\top\mathbf{R}_q\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}^\top\mathbf{R}_q^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}) \\
&\approx \mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) + \frac{1}{2}\boldsymbol{\delta}^\top\left\{\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top}\mathbf{R}_q\mathbf{R}_q^\top\right\}\boldsymbol{\delta} - \\
&\qquad \frac{1}{2}\tilde{\boldsymbol{\delta}}^\top\left\{\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top}\mathbf{R}_q\mathbf{R}_q^\top\right\}\tilde{\boldsymbol{\delta}},
\end{aligned}$$

where $\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) = \partial\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})/\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1$, $\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}}) = \partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1/\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}$ and $\mathbf{R}_q = \partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}/\partial\tilde{\boldsymbol{\delta}}^\top$. The constant terms do not affect the PMLE problem (10) in Section 4 and hence they can be eliminated. Therefore $\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta})$ can be locally approximated (except for a constant term) by

$$\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\boldsymbol{\delta}) \approx \frac{1}{2}\boldsymbol{\delta}^\top\left\{\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top}\mathbf{R}_q\mathbf{R}_q^\top\right\}\boldsymbol{\delta}.$$

J.3: Derivation of $\boldsymbol{\Lambda}^{\mathrm{L}}_{\lambda_{\vartheta^*}}$ and $\boldsymbol{\Lambda}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}$

Based on the approximation derived in Supplementary Material J.2, we have that the penalty matrix $\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}}$ is equal to

$$\boldsymbol{\Lambda}^{\mathcal{G}}_{\lambda_{\vartheta^*}} = \nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) \cdot \frac{\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})}{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top}\mathbf{R}_q\mathbf{R}_q^\top.$$

Quantity $\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathcal{G}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})$ for Lasso and Adaptive Lasso, respectively, is equal to

$$\begin{aligned}
\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathrm{L}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) &= \frac{\partial\mathcal{P}^{\mathrm{L}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} = \frac{\partial\left(\lambda_{\vartheta^*}\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1\right)}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} = \lambda_{\vartheta^*}, \\
\nabla_{\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}\mathcal{P}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}}) &= \frac{\partial\mathcal{P}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}(\tilde{\boldsymbol{\delta}})}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} = \frac{\partial\left(\lambda_{\vartheta^*}\sum_q w_q|\mathbf{R}_q\tilde{\boldsymbol{\delta}}|\right)}{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1} = \lambda_{\vartheta^*}w_q.
\end{aligned}$$

Derivative $\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}})$ is equal to

$$\begin{aligned}
\mathcal{D}_1(\mathbf{R}_q\tilde{\boldsymbol{\delta}}) &= \frac{\partial\|\mathbf{R}_q\tilde{\boldsymbol{\delta}}\|_1}{\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}} \\
&= \frac{\partial}{\partial\mathbf{R}_q\tilde{\boldsymbol{\delta}}}\sum_{q=Q-2}^{Q}\left\{\left(\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^{1/2}\right\} \\
&= \sum_{q=Q-2}^{Q}\left\{\frac{1}{2}\left(\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^{-1/2}\cdot 2\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right\} \\
&= \sum_{q=Q-2}^{Q}\frac{\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\sqrt{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}}} \\
&\approx \sum_{q=Q-2}^{Q}\frac{\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\sqrt{\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top\mathbf{R}_q\tilde{\boldsymbol{\delta}}+\bar{c}}},
\end{aligned}$$

where the denominator was approximated by $\left(\left(\mathbf{R}_q\tilde{\boldsymbol{\delta}}\right)^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}\right)^{-1/2}$ which allows for $\tilde{\boldsymbol{\delta}} = \mathbf{0}$. It follows that $\boldsymbol{\Lambda}^{\mathrm{L}}_{\lambda_{\vartheta^*}}$ can be expressed as

$$
\begin{aligned}
\boldsymbol{\Lambda}^{\mathrm{L}}_{\lambda_{\vartheta^*}} &= \sum_{q=Q-2}^{Q} \left\{ \frac{\lambda_{\vartheta^*}}{\mathbf{R}_q\tilde{\boldsymbol{\delta}}} \cdot \frac{\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\sqrt{(\mathbf{R}_q\tilde{\boldsymbol{\delta}})^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}}} \cdot \mathbf{R}_q\mathbf{R}_q^\top \right\} \\
&= \sum_{q=Q-2}^{Q} \left\{ \frac{\lambda_{\vartheta^*}}{\sqrt{(\mathbf{R}_q\tilde{\boldsymbol{\delta}})^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}}} \cdot \mathbf{R}_q\mathbf{R}_q^\top \right\} \\
&= \lambda_{\vartheta^*} \left\{ \frac{1}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 1, 0, 0) + \right. \\
&\quad \frac{1}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 0, 1, 0) + \\
&\quad \left. \frac{1}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 0, 0, 1) \right\} \\
&= \lambda_{\vartheta^*} \operatorname{diag}\left(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, \frac{1}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}}, \frac{1}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}}, \frac{1}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \right),
\end{aligned}
$$

while $\boldsymbol{\Lambda}^{\mathrm{AL}}_{\lambda_{\vartheta^*}}$ is equal to

$$
\begin{aligned}
\boldsymbol{\Lambda}^{\mathrm{AL}}_{\lambda_{\vartheta^*}} &= \sum_{q=Q-2}^{Q} \left\{ \frac{\lambda_{\vartheta^*}\mathrm{w}_q}{\mathbf{R}_q\tilde{\boldsymbol{\delta}}} \cdot \frac{\mathbf{R}_q\tilde{\boldsymbol{\delta}}}{\sqrt{(\mathbf{R}_q\tilde{\boldsymbol{\delta}})^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}}} \cdot \mathbf{R}_q\mathbf{R}_q^\top \right\} \\
&= \sum_{q=Q-2}^{Q} \left\{ \frac{\lambda_{\vartheta^*}\mathrm{w}_q}{\sqrt{(\mathbf{R}_q\tilde{\boldsymbol{\delta}})^\top \mathbf{R}_q\tilde{\boldsymbol{\delta}} + \bar{c}}} \cdot \mathbf{R}_q\mathbf{R}_q^\top \right\} \\
&= \lambda_{\vartheta^*} \left\{ \frac{\mathrm{w}_{12}}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 1, 0, 0) + \right. \\
&\quad \frac{\mathrm{w}_{13}}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 0, 1, 0) + \\
&\quad \left. \frac{\mathrm{w}_{23}}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \operatorname{diag}(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, 0, 0, 1) \right\} \\
&= \lambda_{\vartheta^*} \operatorname{diag}\left(\mathbf{0}_{P_1\times P_1}, \mathbf{0}_{P_2\times P_2}, \mathbf{0}_{P_3\times P_3}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{12}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{12} + \bar{c}}}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{13}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{13} + \bar{c}}}, \frac{1/|\hat{\vartheta}^{*\mathrm{MLE}}_{23}|^{\bar{\gamma}}}{\sqrt{\vartheta^{*2}_{23} + \bar{c}}} \right).
\end{aligned}
$$

Supplementary Material K: Some theoretical aspects

In what follows we assume that $s_{m\nu_m}(z_{m\nu_m i})$ is approximated by a spline basis with fixed high dimension, $\forall m, \nu_m, i$. Although this may be regarded as a strong assumption, in practice estimation is achieved with finite bases which, if rich enough, will allow one to assume that, compared to estimation variability, the modelling bias resulting from this approximation may be ignored (Kauermann, 2005). We also assume that both $\tilde{\mathbf{S}}_{\boldsymbol{\lambda}}$ and $\boldsymbol{\Lambda}_{\lambda_{\vartheta^*}}$ (superscripts $\mathcal{G}$ and R have been suppressed to avoid clutter) are employed and denote the MLE as $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ and the PMLE as $\hat{\boldsymbol{\delta}}$.

Theorem 1. *Under certain regularity conditions, it can be proved that*

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1}\right),$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) = -\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ *and* $\boldsymbol{\delta}_0$ *denotes the true value vector of* $\boldsymbol{\delta}$.

Proof. By definition, the gradient of the log-likelihood function at $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ is equal to zero, that is $\boldsymbol{g}(\hat{\boldsymbol{\delta}}^{\text{MLE}}) = \mathbf{0}$. If $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ is close to $\boldsymbol{\delta}_0$, then $\boldsymbol{g}(\hat{\boldsymbol{\delta}}^{\text{MLE}})$ can be approximated by a Taylor series around the true parameter $\boldsymbol{\delta}_0$. We apply the mean value theorem in order to truncate the Taylor series at the second term, that is

$$\boldsymbol{g}(\hat{\boldsymbol{\delta}}^{\text{MLE}}) \approx \boldsymbol{g}(\boldsymbol{\delta}_0) + \boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \quad = \quad \mathbf{0}.$$

Multiplying both sides by $\sqrt{n}$ and rearranging, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \quad \approx \quad \{-\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\}^{-1}\left\{\sqrt{n}\boldsymbol{g}(\boldsymbol{\delta}_0)\right\},$$

and by dividing $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ and $\boldsymbol{g}(\boldsymbol{\delta}_0)$ by n we obtain

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \approx \left\{-\frac{1}{n}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\left\{\sqrt{n}\frac{\boldsymbol{g}(\boldsymbol{\delta}_0)}{n}\right\}.$$

Since $\boldsymbol{g}(\boldsymbol{\delta}_0)/n$ is the mean of a random sample, we may apply the Central Limit Theorem (CLT) to $\sqrt{n}\boldsymbol{g}(\boldsymbol{\delta}_0)/n$. According to the theorem and given that $\mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\delta}_0)) = \mathbf{0}$ (as $\boldsymbol{\delta}_0$ is the maximizer of $\ell(\boldsymbol{\delta}_0)$, $\forall i$) we have that

$$\sqrt{n}\left\{\frac{1}{n}\boldsymbol{g}(\boldsymbol{\delta}_0) - \cancelto{0}{\mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\delta}_0))}\right\} \to \mathcal{N}\left(\mathbf{0}, \mathbf{Cov}\left(\boldsymbol{g}_i(\boldsymbol{\delta}_0)\right)\right),$$

where

$$\mathbf{Cov}\left(\boldsymbol{g}_i(\boldsymbol{\delta}_0)\right) = \mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\delta}_0)\boldsymbol{g}_i(\boldsymbol{\delta}_0)^\top) = \left\{-\mathbb{E}\left(\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0)\right)\right\} = -\mathbb{E}\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0) = -\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0).$$

It follows that

$$\sqrt{n}\left\{\frac{1}{n}\boldsymbol{g}(\boldsymbol{\delta}_0)\right\} \to \mathcal{N}\left(\mathbf{0}, -\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right).$$

By using the limiting distribution $\mathbb{P}\left(\lim\left(-1/n\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right)\right) = -1/n\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ we have that

$$\left\{-\frac{1}{n}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\left\{\sqrt{n}\frac{\boldsymbol{g}(\boldsymbol{\delta}_0)}{n}\right\} \;\to\; \mathcal{N}\left(\mathbf{0}, \bar{\boldsymbol{\Gamma}}\right),$$

for

$$\bar{\boldsymbol{\Gamma}} = \left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1} = \left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}.$$

Equivalently

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \;\to\; \mathcal{N}\left(\mathbf{0}, \left\{-\frac{1}{n}\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\right),$$

or

$$\sqrt{n}(\hat{\boldsymbol{\delta}}^{\text{MLE}} - \boldsymbol{\delta}_0) \;\to\; \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1}\right),$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) = -\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ and $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)$ denotes the Fisher information matrix, as required. $\square$

Note that although $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ is unbiased, when $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)$ is near singular then $\hat{\boldsymbol{\delta}}^{\text{MLE}}$ has a large covariance matrix.

In what follows we consider the following assumptions (Cox & Barndorff-Nielsen, 1994, Ch. 3, pp. 82-83): (i) $\boldsymbol{g}(\boldsymbol{\delta}_0) \equiv \sqrt{n}\bar{\boldsymbol{g}}(\boldsymbol{\delta}_0) = \mathcal{O}_P(n^{1/2})$; (ii) $\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) = -\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) \equiv -n\boldsymbol{\mathcal{I}}_i(\boldsymbol{\delta}_0) = \mathcal{O}(n)$; (iii) $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) - \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) = \mathcal{O}_P(n^{1/2})$; (iv) $\bar{\boldsymbol{\lambda}} = o(n^{1/2})$, where $\bar{\boldsymbol{g}}(\boldsymbol{\delta}_0) = \mathcal{O}_P(1)$, $\boldsymbol{\mathcal{I}}_i(\boldsymbol{\delta}_0) = \mathcal{O}(1)$, $\bar{\boldsymbol{g}}(\boldsymbol{\delta}_0)$ is a normalized score function defined as $\bar{\boldsymbol{g}}(\boldsymbol{\delta}_0) = 1/n\boldsymbol{g}(\boldsymbol{\delta}_0) - \mathbb{E}\boldsymbol{g}(\boldsymbol{\delta}_0) = 1/n\boldsymbol{g}(\boldsymbol{\delta}_0)$ for $\mathbb{E}\boldsymbol{g}(\boldsymbol{\delta}_0) \approx \mathbf{0}$ and $\boldsymbol{\mathcal{I}}_i(\boldsymbol{\delta}_0)$ and $\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0)$ denote the expected and the observed Fisher information for a single observation, respectively, for $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) \equiv n\boldsymbol{\mathcal{I}}_i(\boldsymbol{\delta}_0)$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) \equiv n\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0)$. Assumption (iii) results by decomposing $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$ in its mean and stochastic part, that is $\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) = \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\epsilon}$ where we assume that $\boldsymbol{\epsilon} = \mathcal{O}_P(n^{1/2})$ (Kauermann, 2005). Assumptions (i) - (iii) are the classical conditions for the consistency of the MLE, while assumption (iv) ensures that the smoothing parameter increases with the sample size; this is equivalent to $\boldsymbol{\Gamma}_{\bar{\lambda}} = o(n^{1/2})$.

Theorem 2. *Under certain regularity conditions, the PMLE has the following asymptotic distribution*

$$\sqrt{n}\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}\left[\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 + \left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\lambda}}\boldsymbol{\delta}_0\right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, n\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right),$$

and thus the asymptotic covariance of $\hat{\boldsymbol{\delta}}$ *is equal to* $\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}^{-1}$ *while its asymptotic bias is* $-\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\lambda}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\lambda}}\boldsymbol{\delta}_0$.

Proof. The first-order Taylor expansion of $\boldsymbol{g}_p(\cdot)$ around $\boldsymbol{\delta}_0$ is as follows

$$\boldsymbol{g}_p(\hat{\boldsymbol{\delta}}) \approx \boldsymbol{g}_p(\boldsymbol{\delta}_0) + \boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0). \tag{33}$$

By using the fact that $\boldsymbol{g}_p(\hat{\boldsymbol{\delta}}) = \mathbf{0}$ and multiplying all terms by $\sqrt{n}$ leads to

$$\sqrt{n}\boldsymbol{g}_p(\boldsymbol{\delta}_0) + \sqrt{n}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;\; = \;\; 0.$$

Inverting the above series results to

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;\; = \;\; -\left\{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{\sqrt{n}\boldsymbol{g}_p(\boldsymbol{\delta}_0)\right\}.$$

We then divide both $\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)$ and $\boldsymbol{g}_p(\boldsymbol{\delta}_0)$ by n, that is

$$\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \;\; = \;\; -\left\{\frac{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\sqrt{n}\frac{\boldsymbol{g}_p(\boldsymbol{\delta}_0)}{n}\right\}.$$

By using the CLT on $\sqrt{n}\boldsymbol{g}_p(\boldsymbol{\delta}_0)/n$ we obtain the following

$$\sqrt{n}\left\{\frac{\boldsymbol{g}_p(\boldsymbol{\delta})}{n} - \mathbb{E}(\boldsymbol{g}_{pi}(\boldsymbol{\delta}_0))\right\} \;\; \rightarrow \;\; \mathcal{N}\left(\mathbf{0}, \mathbf{Cov}(\boldsymbol{g}_{pi}(\boldsymbol{\delta}_0)\right), \tag{34}$$

where $\mathbb{E}(\boldsymbol{g}_{pi}(\boldsymbol{\delta}_0)) = 1/n\mathbb{E}\left(\boldsymbol{g}_p(\boldsymbol{\delta}_0)\right) = 1/n\mathbb{E}\left(\boldsymbol{g}(\boldsymbol{\delta}_0)\right) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0) = 1/n\left[\mathbb{E}\left(\boldsymbol{g}(\boldsymbol{\delta}_0)\right) - \mathbb{E}\left(\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right)\right] = 1/n\left[\mathbf{0} - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] = -1/n\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0$ and $\mathbf{Cov}(\boldsymbol{g}_{pi}(\boldsymbol{\delta}_0)) = \mathbf{Cov}(\boldsymbol{g}_i(\boldsymbol{\delta}_0) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0) = \mathbf{Cov}(\boldsymbol{g}_i(\boldsymbol{\delta}_0)) = -\mathbb{E}\boldsymbol{\mathcal{H}}_i(\boldsymbol{\delta}_0) = -1/n\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)$. Therefore, (34) can be re-expressed as

$$\sqrt{n}\left\{\frac{\boldsymbol{g}_p(\boldsymbol{\delta}) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{n}\right\} \;\; \rightarrow \;\; \mathcal{N}\left(\mathbf{0}, -1/n\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right),$$

and thus

$$\sqrt{n}\frac{\boldsymbol{g}_p(\boldsymbol{\delta})}{n} \;\; \rightarrow \;\; \mathcal{N}\left(-\frac{\sqrt{n}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{n}, -1/n\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right).$$

Next we use the law of large numbers that says that the observed information converges to the expected Fisher information as the sample size increases. That is, $-\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0) \rightarrow -\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)$.

Therefore

$$-\left\{\frac{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\sqrt{n}\frac{\boldsymbol{g}_p(\boldsymbol{\delta}_0)}{n}\right\} \;\rightarrow\; \mathcal{N}\left(\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{\sqrt{n}}\right\},\right.$$
$$\left.\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{n}\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\right),$$

which implies that

$$\sqrt{n}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0) \;\rightarrow\; \mathcal{N}\left(\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{\sqrt{n}}\right\},\right.$$
$$\left.\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{n}\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\right), \tag{35}$$

From the above result we can calculate the bias of the estimator $\hat{\boldsymbol{\delta}}$, that is

$$\begin{aligned}
\mathbf{Bias}(\hat{\boldsymbol{\delta}}) &= \mathbb{E}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)\\
&\approx \frac{1}{\sqrt{n}}\mathbb{E}\left[\sqrt{n}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)\right]\\
&\approx \frac{1}{\sqrt{n}}\left[\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{\sqrt{n}}\right\}\right]\\
&\approx -\frac{1}{\sqrt{n}}n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0}{\sqrt{n}}\right\}\\
&\approx -\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\\
&\approx -\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\\
&\approx -\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0.
\end{aligned}$$

as well as its asymptotic covariance matrix

$$\begin{aligned}
\mathbf{Cov}(\hat{\boldsymbol{\delta}}) &\approx \frac{1}{n}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{n}\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)}{n}\right\}^{-1} \\
&\approx \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1} \\
&\approx \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \\
&\approx \left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}.
\end{aligned}$$

Rearranging (35) leads to

$$\sqrt{n}\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}\left[(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)+\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] \;\rightarrow\; \mathcal{N}\left(\mathbf{0}, n\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right),$$

which results from the following: expression (35) can be re-written as

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0) \;\rightarrow\; \mathcal{N}\Big(&-\sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\}, \\
&n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\Big),
\end{aligned}$$

or

$$\sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0) \;\rightarrow\; \mathcal{N}\left(-\sqrt{n}\left\{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\}, n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\right),$$

and therefore,

$$\begin{aligned}
\sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)+\sqrt{n}\left\{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\} &\rightarrow \mathcal{N}\left(\mathbf{0}, n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\right) \Longrightarrow \\
\Longrightarrow \sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}\left[\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0+\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] &\rightarrow \mathcal{N}\left(\mathbf{0}, n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\right) \Longrightarrow \\
\Longrightarrow \sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}\left[\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0+\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] &\rightarrow \mathcal{N}\left(\mathbf{0}, n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\right) \Longrightarrow \\
\Longrightarrow \sqrt{n}\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}\left[(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)+\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right] &\rightarrow \mathcal{N}\left(\mathbf{0}, n\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right).
\end{aligned}$$

□

Note that when $\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)$ is near singular then $\mathbf{Cov}(\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}) \to \infty$ and $\mathbf{Cov}(\hat{\boldsymbol{\delta}}) \to \mathbf{0}$. This verifies that asymptotically the PMLE has smaller variance than the MLE and thus may perform better.

Under assumptions (i)-(iv) we have that

$$\begin{aligned}
\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 &= -\left\{\boldsymbol{\mathcal{H}}_p(\boldsymbol{\delta}_0)\right\}^{-1} \boldsymbol{g}_p(\boldsymbol{\delta}_0) + \dots \\
&= -\left\{\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \left\{\boldsymbol{g}(\boldsymbol{\delta}_0) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\} + \dots \\
&= -\left\{\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) - \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \left\{\boldsymbol{g}(\boldsymbol{\delta}_0) - \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0\right\} + \dots \\
&= -\left\{\mathcal{O}_P(n^{1/2}) + \mathcal{O}(n) - o(n^{1/2})\right\}^{-1} \left\{\mathcal{O}_P(n^{1/2}) - o(n^{1/2})\right\} \\
&= \left\{\mathcal{O}_P(n)\right\}^{-1} \left\{\mathcal{O}_P(n^{1/2})\right\} \\
&= \mathcal{O}_P(n^{-1})\mathcal{O}_P(n^{1/2}) \\
&= \mathcal{O}_P(n^{-1/2}),
\end{aligned}$$

where the first line results by rearranging equation (33). Assumptions (ii) and (iv) imply that

$$\begin{aligned}
\mathbf{Cov}(\hat{\boldsymbol{\delta}}) &\approx \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\} \left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \\
&= \left\{\mathcal{O}(n) + o(n^{1/2})\right\}^{-1} \left\{\mathcal{O}(n)\right\} \left\{\mathcal{O}(n) + o(n^{1/2})\right\}^{-1} \\
&= \left\{\mathcal{O}(n)\right\}^{-1} \left\{\mathcal{O}(n)\right\} \left\{\mathcal{O}(n)\right\}^{-1} \\
&= \mathcal{O}(n^{-1}),
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{Bias}(\hat{\boldsymbol{\delta}}) &\approx -\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0 \\
&= -\left\{-\mathcal{O}(n) + o(n^{1/2})\right\}^{-1} o(n^{1/2}) \\
&= \left\{\mathcal{O}(n)\right\}^{-1} o(n^{1/2}) \\
&= \mathcal{O}(n^{-1})\, o(n^{1/2}) \\
&= o(n^{-1/2}).
\end{aligned}$$

Theorem 3. *If* $\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0| = o(n^{1/2})$ *and* $\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}| = o(n^{1/2})$*, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\right) \sim \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1}\right).$$

Proof. If $\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0| = o(n^{1/2})$ and $\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}| = o(n^{1/2})$, then as $n \to \infty$ we have that $1/\sqrt{n}\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\boldsymbol{\delta}_0| \to \mathbf{0}$ and $1/\sqrt{n}\max|\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}| \to \mathbf{0}$. Given these two conditions, it follows that

$$\begin{aligned}
\mathbb{E}\left(\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)\right) &= \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{n}\right\}^{-1} \left\{-\frac{\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}\boldsymbol{\delta}_0}}{\sqrt{n}}\right\} \\
&\to \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0) + \boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{n}\right\}^{-1} \cdot \mathbf{0} \\
&\to \mathbf{0},
\end{aligned}$$

and

$$
\begin{aligned}
\mathbf{Cov}\left(\sqrt{n}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)\right) &= \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{n}\right\}^{-1}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{n}\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{n}\right\}^{-1} \\
&= n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}\right\}^{-1} \\
&= \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{\sqrt{n}}\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)+\boldsymbol{\Gamma}_{\bar{\boldsymbol{\lambda}}}}{\sqrt{n}}\right\}^{-1} \\
&\to \left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{\sqrt{n}}+\mathbf{0}\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{\frac{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)}{\sqrt{n}}+\mathbf{0}\right\}^{-1} \\
&\to \sqrt{n}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1}\sqrt{n} \\
&\to n\left\{-\mathbb{E}\boldsymbol{\mathcal{H}}(\boldsymbol{\delta}_0)\right\}^{-1} \\
&\to n\left\{\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1} \\
&\to \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta})\right\}^{-1},
\end{aligned}
$$

and thus

$$
\sqrt{n}\left(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0\right) \sim \mathcal{N}\left(\mathbf{0}, \left\{\frac{1}{n}\boldsymbol{\mathcal{I}}(\boldsymbol{\delta}_0)\right\}^{-1}\right).
$$

□

Theorem 4. *Suppose that $\bar{\boldsymbol{\lambda}} \in [0,\infty)$ is fixed. Then the PMLE $\hat{\boldsymbol{\delta}}$ that minimizes $-\ell_p(\boldsymbol{\delta})$ is consistent, that is $\lim_{n\to\infty}\mathbb{P}(\|\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0\|^2 > \bar{\varepsilon}) = 0$, $\forall \bar{\varepsilon} > 0$.*

Proof. If $\hat{\boldsymbol{\delta}}$ minimizes $-\ell_p(\boldsymbol{\delta})$, then it also minimizes $-\ell_p(\boldsymbol{\delta})/n$. Similarly, $\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}$ minimizes $-\ell(\boldsymbol{\delta})$ as well as $-\ell(\boldsymbol{\delta})/n$. Because $\bar{\boldsymbol{\lambda}}$ is fixed, we have that $-\ell_p(\hat{\boldsymbol{\delta}})/n \to -\ell(\hat{\boldsymbol{\delta}}^{\mathrm{MLE}})/n$ and $-\ell_p(\hat{\boldsymbol{\delta}})/n \to -\ell(\hat{\boldsymbol{\delta}})/n$; thus $-\ell(\hat{\boldsymbol{\delta}})/n \to -\ell(\hat{\boldsymbol{\delta}}^{\mathrm{MLE}})/n$ hold as well. Since $\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}$ is a unique minimizer of $-\ell(\boldsymbol{\delta})/n$ and $-\ell(\boldsymbol{\delta})/n$ is convex, it follows that $\hat{\boldsymbol{\delta}} \to \hat{\boldsymbol{\delta}}^{\mathrm{MLE}}$. The consistency of $\hat{\boldsymbol{\delta}}$ follows form the consistency of $\hat{\boldsymbol{\delta}}^{\mathrm{MLE}}$. □

The above theorems have mainly been adapted from Fan & Li (2001), Li & Sudjianto (2005) and Oelker et al. (2014).

Theorem 3 shows that as the sample size grows large, under certain conditions, the asymptotic distribution of the PMLE coincides with that of MLE. This is a desirable property as it is well-known that the MLE is the most efficient estimator. The above theorem also suggests that PMLE is essentially needed when the sample size is small. This is in line with the results obtained in the simulation studies in Sections 3.3 and 4.2.

References

Anderson, T. W., Anderson, T. W., Anderson, T. W., & Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. Wiley New York.

Azzalini, M. A. (2014). *The Multivariate Normal and t Distributions*. R package version 1.5-3.

Cappellari, L. & Jenkins, S. P. (2003). Multivariate probit regression using simulated maximum likelihood. *The Stata Journal*, 3(3), 278–294.

Cox, D. & Barndorff-Nielsen, O. (1994). *Inference and Asymptotics*. CRC Press.

Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2), 89–102.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

Genz, A. (1991). An adaptive numerical integration algorithm for simplices. In *Computing in the 90's* (pp. 279–285). Springer.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2), 141–149.

Genz, A. & Bretz, F. (2002). Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, 11(4), 950–971.

Genz, A. & Kass, R. E. (1997). Subregion-adaptive integration of functions having a dominant peak. *Journal of Computational and Graphical Statistics*, 6(1), 92–111.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities.

Glass, G. V. & Collins, J. R. (1970). Geometric proof of the restriction on the possible values of rxy when ryz are fixed. *Educational and Psychological Measurement*, 30, 37–39.

Hajivassiliou, V. A. & McFadden, D. (1991). *The method of simulated scores for the estimation of LDV models with an application to external debt crises*. Yale University, Cowles Foundation for Research in Economics.

Hubert, L. J. (1972). A note on the restriction of range for pearson product-moment correlation coefficients. *Educational and Psychological Measurement*, 32, 767–770.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, 49(1), 169–186.

Keane, M. P. (1990). *Four essays in empirical macro and labor economics*. Ph.D. dissertation, Brown University.

Leung, C.-K. & Lam, K. (1975). A note on the geometric representation of the correlation coefficients. *The American Statistician*, 29(3), 128–130.

Li, R. & Sudjianto, A. (2005). Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, 47(2), 111–120.

Marius Hofert, Ivan Kojadinovic, M. M. & Yan, J. (2015). *Multivariate Dependence with Copulas*. R package version 0.999-14.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., McGovern, M. E., et al. (2016). A simultaneous equation approach to estimating hiv prevalence with non-ignorable missing responses. *Journal of the American Statistical Association*.

Marra, G. & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1), 53–74.

Nocedal, J. & Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Oelker, M.-R., Gertheiss, J., & Tutz, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling*, 14(2), 157–177.

Plackett, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika*, 41(3/4), 351–360.

Radice, R., Marra, G., & Wojtyś, M. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, 26(5), 981–995.

Rossi, P. (2015). *Bayesian Inference for Marketing/Micro-Econometrics*. R package version Version 3.0-2.

Rousseeuw, P. J. & Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics–Theory and Methods*, 22(4), 965–984.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields*. New Haven: Chapman & Hall/CRC, Boca Raton, FL.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge university press.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of The Royal Statistical Society Series B*, 47, 1–52.

Stanley, J. C. & Wang, M. D. (1969). Restrictions on the possible values of r12 given r13 and r23. *Educational and Psychological Measurement*, 29, 579–581.

Team, R. C. & contributors worldwide (2015). *The R Stats Package*. R package version 3.1.3.

Trinh, G. & Genz, A. (2015). Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing*, 25(5), 989–996.

Ulbricht, J. (2010). *Variable selection in generalized linear models*. Verlag Dr. Hut.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. Chapman & Hall/CRC, London.

Wood, S. N. (2013a). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S. N. (2013b). A simple test for random effects in regression models. *Biometrika*, 100(4), 1005–1010.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models*. R package version 0.9-7.