Supplementary material for 'Doubly stochastic models for spatio-temporal covariation of replicated point processes'

Daniel Gervini
Department of Mathematical Sciences
University of Wisconsin–Milwaukee

January 22, 2021

1 Estimation

1.1 The model

We have a doubly stochastic process (X, Λ) where $X|\Lambda = \lambda$ is a Poisson process with intensity function λ, and Λ follows the model

$$\Lambda(t, \mathbf{s}) = R\Lambda_t(t)\Lambda_s(\mathbf{s}) \tag{1}$$

with

$$\log R = \tau + Z, \tag{2}$$

$$\log \Lambda_t(t) = \mu(t) + \sum_{k=1}^{p_1} U_k \phi_k(t), \tag{3}$$

and

$$\log \Lambda_s(\mathbf{s}) = \nu(\mathbf{s}) + \sum_{k=1}^{p_2} V_k \psi_k(\mathbf{s}). \tag{4}$$

In (2), Z is a random variable with $E(Z) = 0$ and $\sigma_z^2 = \text{var}(Z)$. In (3) and (4), the ϕ_ks and ψ_ks are orthonormal functions in their respective spaces, $E(U_k) = E(V_k) = 0$ for all k, and $\text{cov}(U_k, U_{k'}) = \text{cov}(V_k, V_{k'}) = 0$ for all $k \neq k'$. Without loss of generality, the components are arranged in decreasing order of variances $\sigma_{uk}^2 = \text{var}(U_k)$ and $\sigma_{vk}^2 = \text{var}(V_k)$. For identifiability we assume that μ, ν, the ϕ_ks and the ψ_ks integrate to zero on their respective domains B_t and B_s. In addition, it is sometimes assumed that the temporal intensity functions are periodic on $B_t = [t_l, t_u]$, in which case we add the constraints $\mu(t_l) = \mu(t_u)$ and $\phi_k(t_l) = \phi_k(t_u)$ for all k.

Let $\mathbf{U} = (U_1, \ldots, U_{p_1})^T$, $\mathbf{V} = (V_1, \ldots, V_{p_2})^T$, $\boldsymbol{\sigma}_u^2 = (\sigma_{u1}^2, \ldots, \sigma_{up_1}^2)^T$, $\boldsymbol{\sigma}_v^2 = (\sigma_{v1}^2, \ldots, \sigma_{vp_2}^2)^T$, $\boldsymbol{\sigma}_{zu} = \text{cov}(Z, \mathbf{U})^T$, $\boldsymbol{\sigma}_{zv} = \text{cov}(Z, \mathbf{V})^T$ and $\boldsymbol{\Sigma}_{uv} = \text{cov}(\mathbf{U}, \mathbf{V})$. We concatenate all random effects into $\mathbf{W} = (Z, \mathbf{U}^T, \mathbf{V}^T)^T$ and assume that $\mathbf{W}$ follows a multivariate normal distribution with mean zero and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_z^2 & \boldsymbol{\sigma}_{zu}^T & \boldsymbol{\sigma}_{zv}^T \\ \boldsymbol{\sigma}_{zu} & \text{diag}(\boldsymbol{\sigma}_u^2) & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\sigma}_{zv} & \boldsymbol{\Sigma}_{uv}^T & \text{diag}(\boldsymbol{\sigma}_v^2) \end{pmatrix}.$$

The functional parameters μ, ν, ϕ_ks and ψ_ks are semiparametrically modeled using splines or similar basis functions: $\mu(t) = \mathbf{c}_0^T \boldsymbol{\beta}_t(t)$, $\phi_k(t) = \mathbf{c}_k^T \boldsymbol{\beta}_t(t)$, $\nu(\mathbf{s}) = \mathbf{d}_0^T \boldsymbol{\beta}_s(\mathbf{s})$, and $\psi_k(\mathbf{s}) = \mathbf{d}_k^T \boldsymbol{\beta}_s(\mathbf{s})$, where $\boldsymbol{\beta}_t(t)$ is the vector of q_1 basis functions of a family $\mathscr{B}_t$ and $\boldsymbol{\beta}_s(\mathbf{s})$ is the vector of q_2 basis functions of a family $\mathscr{B}_s$. The orthonormality constraints on the ϕ_ks can be expressed as $\mathbf{c}_k^T \mathbf{J}_t \mathbf{c}_{k'} = \delta_{kk'}$, where $\delta_{kk'}$ is Kronecker's delta and $\mathbf{J}_t = \int \boldsymbol{\beta}_t(t) \boldsymbol{\beta}_t(t)^T dt$. Similarly for the ψ_ks. The zero-integral constraints for μ and the ϕ_ks can be expressed as $\mathbf{a}_{t0}^T \mathbf{c}_k = 0$ for $k = 0, \ldots, p_1$, where $\mathbf{a}_{t0} = \int_{B_t} \boldsymbol{\beta}_t(t) dt$. Similarly for ν

and the ψ_ks. Finally, the periodicity constraints for μ and the ϕ_ks and their respective derivatives, if present, can be expressed as $\mathbf{A}_P^T \mathbf{c}_k = \mathbf{0}$ for $k = 0, \ldots, p_1$, with $\mathbf{A}_P = [\boldsymbol{\beta}_t(t_u) - \boldsymbol{\beta}_t(t_l), \boldsymbol{\beta}_t'(t_u) - \boldsymbol{\beta}_t'(t_l)]^T$.

We collect all model parameters into a single vector

$$\boldsymbol{\theta} = (\boldsymbol{\sigma}_{zu}, \boldsymbol{\sigma}_{zv}, \operatorname{vec} \boldsymbol{\Sigma}_{uv}, \tau, \sigma_z^2, \mathbf{c}_0, \operatorname{vec} \mathbf{C}, \boldsymbol{\sigma}_u^2, \mathbf{d}_0, \operatorname{vec} \mathbf{D}, \boldsymbol{\sigma}_v^2), \tag{5}$$

where $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_{p_1}]$ and $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_{p_2}]$. For an observation $x = \{(t_1, \mathbf{s}_1), \ldots, (t_m, \mathbf{s}_m)\}$ of the process X on a given bounded region $B = B_t \times B_s$, the joint density of x and the latent $\mathbf{w}$ can be factorized as

$$f_{\boldsymbol{\theta}}(x, \mathbf{w}) = f_{\boldsymbol{\theta}}(x \mid \mathbf{w}) f_{\boldsymbol{\theta}}(\mathbf{w})$$

with

$$f_{\boldsymbol{\theta}}(x \mid \mathbf{w}) = \frac{\exp\{-r I_t(\mathbf{u}) I_s(\mathbf{v})\}}{m!} \; r^m \prod_{j=1}^{m} \lambda_t(t_j; \mathbf{u}) \prod_{j=1}^{m} \lambda_s(\mathbf{s}_j; \mathbf{v}) \tag{6}$$

and

$$f_{\boldsymbol{\theta}}(\mathbf{w}) = \frac{1}{(2\pi)^{p/2} (\det \boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2} \mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w}\right),$$

where $r = \exp(\tau + z)$, $I_t(\mathbf{u}) = \int_{B_t} \lambda_t(t; \mathbf{u}) dt$, $I_s(\mathbf{v}) = \int_{B_s} \lambda_s(\mathbf{s}; \mathbf{v}) d\mathbf{s}$, $\lambda_t(t; \mathbf{u}) = \exp\{\mu(t) + \mathbf{u}^T \boldsymbol{\phi}(t)\}$, $\lambda_s(\mathbf{s}; \mathbf{v}) = \exp\{\nu(\mathbf{s}) + \mathbf{v}^T \boldsymbol{\psi}(\mathbf{s})\}$, and $p = p_1 + p_2 + 1$. The marginal density of X is

$$f_{\boldsymbol{\theta}}(x) = \iint f_{\boldsymbol{\theta}}(x, \mathbf{w}) d\mathbf{w},$$

which we evaluate by Laplace's approximation as explained in Section 1.4.

1.2 EM algorithm

The penalized maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is the maximizer of

$$\ell(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(x_i) - \xi_1 P(\mu) - \xi_2 \sum_{k=1}^{p_1} P(\phi_k) - \xi_3 P(\nu) - \xi_4 \sum_{k=1}^{p_2} P(\psi_k) \tag{7}$$

for $\boldsymbol{\theta}$ in the parameter space (assuming periodicity constraints)

$$\begin{aligned} \Theta \;=\; & \{\boldsymbol{\theta} \in \mathbb{R}^r : h_{kl}^C(\boldsymbol{\theta}) = 0, \;\; k = 1, \ldots, l, \;\; l = 1, \ldots, p_1; \\ & h_{kl}^D(\boldsymbol{\theta}) = 0, \;\; k = 1, \ldots, l, \;\; l = 1, \ldots, p_2; \\ & \mathbf{a}_{t0}^T \mathbf{c}_k = 0, \; k = 0, \ldots, p_1; \; \mathbf{a}_{s0}^T \mathbf{d}_k = 0, \; k = 0, \ldots, p_2; \\ & \mathbf{A}_P \mathbf{c}_k = \mathbf{0}, \; k = 0, \ldots, p_1; \; \boldsymbol{\Sigma} > 0\}, \end{aligned} \tag{8}$$

where r is the dimension of $\boldsymbol{\theta}$, $h_{kl}^C(\boldsymbol{\theta}) = \mathbf{c}_k^T \mathbf{J}_t \mathbf{c}_l - \delta_{kl}$, $h_{kl}^D(\boldsymbol{\theta}) = \mathbf{d}_k^T \mathbf{J}_s \mathbf{d}_l - \delta_{kl}$ and $\boldsymbol{\Sigma} > 0$ denotes that $\boldsymbol{\Sigma}$ is symmetric and positive definite.

The penalty functions are quadratic on the basis coefficients: if $f = \mathbf{c}^T\boldsymbol{\beta}$ then $P(f) = \mathbf{c}^T\boldsymbol{\Omega}\mathbf{c}$ for $\boldsymbol{\Omega}$ that depends only on $\boldsymbol{\beta}$. Specifically, for the temporal functions $P(f) = \int (f'')^2$ and

$$\boldsymbol{\Omega}_t = \int \boldsymbol{\beta}_t''(t)\boldsymbol{\beta}_t''(t)^T dt;$$

for the spatial functions $P(f) = \iint \{(\frac{\partial^2 f}{\partial s_1^2})^2 + 2(\frac{\partial^2 f}{\partial s_1 \partial s_2})^2 + (\frac{\partial^2 f}{\partial s_2^2})^2\}$ and

$$\boldsymbol{\Omega}_s = \mathbf{J}_{11} + 2\mathbf{J}_{12} + \mathbf{J}_{22}$$

with

$$\mathbf{J}_{ij} = \iint \left(\frac{\partial^2 \boldsymbol{\beta}_s(\mathbf{s})}{\partial s_i \partial s_j}\right)\left(\frac{\partial^2 \boldsymbol{\beta}_s(\mathbf{s})}{\partial s_i \partial s_j}\right)^T d\mathbf{s}.$$

Then

$$\ell(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(x_i) - \xi_1 \mathbf{c}_0^T \boldsymbol{\Omega}_t \mathbf{c}_0 - \xi_2 \operatorname{tr}(\mathbf{C}^T\boldsymbol{\Omega}_t\mathbf{C}) - \xi_3 \mathbf{d}_0^T\boldsymbol{\Omega}_s\mathbf{d}_0 - \xi_4 \operatorname{tr}(\mathbf{D}^T\boldsymbol{\Omega}_s\mathbf{D}).$$

The EM algorithm (Dempster et al., 1977) works iteratively as follows: given the current value of the estimator $\hat{\boldsymbol{\theta}}_{(k-1)}$, the updated value $\hat{\boldsymbol{\theta}}_{(k)}$ is defined as the maximizer of

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{(k-1)}) &= \frac{1}{n}\sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}}\{\log f_{\boldsymbol{\theta}}(x_i, \mathbf{w}) \mid x_i\} \\ &\quad -\xi_1 \mathbf{c}_0^T \boldsymbol{\Omega}_t \mathbf{c}_0 - \xi_2 \operatorname{tr}(\mathbf{C}^T\boldsymbol{\Omega}_t\mathbf{C}) - \xi_3 \mathbf{d}_0^T\boldsymbol{\Omega}_s\mathbf{d}_0 - \xi_4 \operatorname{tr}(\mathbf{D}^T\boldsymbol{\Omega}_s\mathbf{D}) \end{aligned}$$

subject to the parameter constraints. Considering the factorization of the joint density and the dependence of each factor on the model parameters, we can write

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_{(k-1)}) = Q_1(\tau, \mathbf{c}_0, \mathbf{C}, \mathbf{d}_0, \mathbf{D} \mid \hat{\boldsymbol{\theta}}_{(k-1)}) + Q_2(\boldsymbol{\Sigma} \mid \hat{\boldsymbol{\theta}}_{(k-1)}),$$

where

$$\begin{aligned} Q_1(\tau, \mathbf{c}_0, \mathbf{C}, \mathbf{d}_0, \mathbf{D} \mid \hat{\boldsymbol{\theta}}_{(k-1)}) &= \frac{1}{n}\sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}}\{\log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) \mid x_i\} \\ &\quad -\xi_1 \mathbf{c}_0^T \boldsymbol{\Omega}_t \mathbf{c}_0 - \xi_2 \operatorname{tr}(\mathbf{C}^T\boldsymbol{\Omega}_t\mathbf{C}) \\ &\quad -\xi_3 \mathbf{d}_0^T\boldsymbol{\Omega}_s\mathbf{d}_0 - \xi_4 \operatorname{tr}(\mathbf{D}^T\boldsymbol{\Omega}_s\mathbf{D}), \end{aligned}$$

and

$$Q_2(\boldsymbol{\Sigma} \mid \hat{\boldsymbol{\theta}}_{(k-1)}) = \frac{1}{n}\sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}}\{\log f_{\boldsymbol{\theta}}(\mathbf{w}) \mid x_i\}.$$

1.3 M-step: updating equations

For τ: Since

$$\begin{aligned}\log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) &= -e^{\tau+z} I_t(\mathbf{u}) I_s(\mathbf{v}) - \log m_i! \\ &\quad + m_i(\tau + z) + \sum_{j=1}^{m_i} \log \lambda_t(t_{ij}; \mathbf{u}) + \sum_{j=1}^{m_i} \log \lambda_s(\mathbf{s}_{ij}; \mathbf{v})\end{aligned}$$

we have

$$\frac{\partial}{\partial \tau} \log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) = -e^{\tau} e^{z} I_t(\mathbf{u}) I_s(\mathbf{v}) + m_i$$

and then

$$\frac{\partial}{\partial \tau} Q_1(\tau, \ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) = -e^{\tau} \frac{1}{n} \sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{e^{z} I_t(\mathbf{u}) I_s(\mathbf{v}) \mid x_i\} + \overline{m}.$$

So the updating equation for $\hat{\tau}_{(k)}$ is

$$\hat{\tau}_{(k)} = \log \left[\overline{m} / \frac{1}{n} \sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{e^{z} I_t(\mathbf{u}) I_s(\mathbf{v}) \mid x_i\} \right].$$

For $\mathbf{c}_0$: Since

$$\begin{aligned}\log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) &= -r I_t(\mathbf{u}) I_s(\mathbf{v}) - \log m_i! \\ &\quad + m_i \log r + \sum_{j=1}^{m_i} \log \lambda_t(t_{ij}; \mathbf{u}) + \sum_{j=1}^{m_i} \log \lambda_s(\mathbf{s}_{ij}; \mathbf{v})\end{aligned}$$

and $\lambda_t(t; \mathbf{u}) = \exp\{\mu(t) + \mathbf{u}^T \boldsymbol{\phi}(t)\}$, we have

$$\mathsf{D}_{\mathbf{c}_0} \log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) = -r I_s(\mathbf{v}) \int_{B_t} \mathsf{D}_{\mathbf{c}_0} \lambda_t(t; \mathbf{u}) dt + \sum_{j=1}^{m_i} \boldsymbol{\beta}_t(t_{ij})^T,$$

where

$$\begin{aligned}\mathsf{D}_{\mathbf{c}_0} \lambda_t(t; \mathbf{u}) &= \lambda_t(t; \mathbf{u}) \mathsf{D}_{\mathbf{c}_0} \log \lambda_t(t; \mathbf{u}) \\ &= \lambda_t(t; \mathbf{u}) \boldsymbol{\beta}_t(t)^T.\end{aligned}$$

Then

$$\begin{aligned}\mathsf{D}_{\mathbf{c}_0} Q_1(\ldots, \mathbf{c}_0, \ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) &= -\frac{1}{n} \sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{r I_s(\mathbf{v}) \int_{B_t} \lambda_t(t; \mathbf{u}) \boldsymbol{\beta}_t(t)^T dt \mid x_i\} \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \boldsymbol{\beta}_t(t_{ij})^T - 2\xi_1 \mathbf{c}_0^T \boldsymbol{\Omega}_t.\end{aligned}$$

Using the Laplace approximation to the conditional expectations, as explained in Section 1.4, we have

$$E_{\hat{\boldsymbol{\theta}}_{(k-1)}}\{rI_s(\mathbf{v})\int_{B_t}\lambda_t(t;\mathbf{u})\boldsymbol{\beta}_t(t)^T dt \mid x_i\} \approx \hat{r}_i I_s(\hat{\mathbf{v}}_i)\int_{B_t}\lambda_t(t;\hat{\mathbf{u}}_i)\boldsymbol{\beta}_t(t)^T dt$$

where $\hat{r}_i$, $\hat{\mathbf{u}}_i$ and $\hat{\mathbf{v}}_i$ are the current predictors of r, $\mathbf{u}$ and $\mathbf{v}$. A Taylor expansion of $\lambda_t(t;\hat{\mathbf{u}}_i)$ on the variable $\mathbf{c}_0$ evaluated at the current $\hat{\mathbf{c}}_{0(k-1)}$ gives

$$\begin{aligned}
\lambda_t(t;\hat{\mathbf{u}}_i) &\approx \hat{\lambda}_{ti}(t) + \mathsf{D}_{\mathbf{c}_0}\lambda_t(t;\hat{\mathbf{u}}_i)|_{\hat{\mathbf{c}}_{0(k-1)}}(\mathbf{c}_0 - \hat{\mathbf{c}}_{0(k-1)}) \\
&= \hat{\lambda}_{ti}(t) + \hat{\lambda}_{ti}(t)\boldsymbol{\beta}_t(t)^T(\mathbf{c}_0 - \hat{\mathbf{c}}_{0(k-1)}) \\
&= \hat{\lambda}_{ti}(t)\{1 - \boldsymbol{\beta}_t(t)^T\hat{\mathbf{c}}_{0(k-1)}\} + \hat{\lambda}_{ti}(t)\boldsymbol{\beta}_t(t)^T\mathbf{c}_0,
\end{aligned}$$

where $\hat{\lambda}_{ti}(t) = \exp\{\hat{\mu}_{(k-1)}(t) + \hat{\mathbf{u}}_i^T\hat{\boldsymbol{\phi}}(t)\}$. Similarly, let $\hat{\lambda}_{si}(\mathbf{s}) = \exp\{\hat{\nu}_{(k-1)}(\mathbf{s}) + \hat{\mathbf{v}}_i^T\hat{\boldsymbol{\psi}}(\mathbf{s})\}$ and $\hat{I}_{si} = \int_{B_s}\hat{\lambda}_{si}(\mathbf{s})d\mathbf{s}$. Then

$$\mathsf{D}_{\mathbf{c}_0}Q_1(\ldots,\mathbf{c}_0,\ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)})$$

$$\begin{aligned}
\approx\ & -\frac{1}{n}\sum_{i=1}^n \hat{r}_i\hat{I}_{si}\int_{B_t}\hat{\lambda}_{ti}(t)\{1-\hat{\mu}_{(k-1)}(t)\}\boldsymbol{\beta}_t(t)^T dt \\
& -\frac{1}{n}\sum_{i=1}^n \hat{r}_i\hat{I}_{si}\int_{B_t}\hat{\lambda}_{ti}(t)\mathbf{c}_0^T\boldsymbol{\beta}_t(t)\boldsymbol{\beta}_t(t)^T dt \\
& +\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{m_i}\boldsymbol{\beta}_t(t_{ij})^T - 2\xi_1\mathbf{c}_0^T\boldsymbol{\Omega}_t.
\end{aligned}$$

The Lagrange condition for the constraints $\mathbf{a}_{t0}^T\mathbf{c}_0 = 0$ and (if present) $\mathbf{A}_P\mathbf{c}_0 = \mathbf{0}$ is

$$\mathsf{D}_{\mathbf{c}_0}Q_1(\ldots,\hat{\mathbf{c}}_{0(k)},\ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) = \boldsymbol{\kappa}^T\mathbf{A},$$

where $\mathbf{A}$ is the matrix with rows $\mathbf{a}_{t0}^T$ and $\mathbf{A}_P$; transposing both sides, we can write it as

$$-\mathbf{Q}\mathbf{c}_{0(k)} + \mathbf{b} = \mathbf{A}^T\boldsymbol{\kappa}$$

with

$$\begin{aligned}
\mathbf{Q} &= \frac{1}{n}\sum_{i=1}^n \hat{r}_i\hat{I}_{si}\int_{B_t}\hat{\lambda}_{ti}(t)\boldsymbol{\beta}_t(t)\boldsymbol{\beta}_t(t)^T dt + 2\xi_1\boldsymbol{\Omega}_t, \\
\mathbf{b} &= -\frac{1}{n}\sum_{i=1}^n \hat{r}_i\hat{I}_{si}\int_{B_t}\hat{\lambda}_{ti}(t)\{1-\hat{\mu}_{(k-1)}(t)\}\boldsymbol{\beta}_t(t)dt + \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{m_i}\boldsymbol{\beta}_t(t_{ij}).
\end{aligned}$$

Then, including the constraints, we solve the system

$$\begin{pmatrix} \mathbf{Q} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{c}}_{0(k)} \\ \boldsymbol{\kappa} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}$$

and obtain $\hat{\mathbf{c}}_{0(k)}$.

For $\mathbf{d}_0$: Given the symmetry of (6) on λ_t and λ_s, from the above derivations for $\mathbf{c}_0$ we have, for $\mathbf{d}_0$, that

$$\begin{aligned} \mathsf{D}_{\mathbf{d}_0} Q_1(\ldots, \mathbf{d}_0, \ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) &= -\frac{1}{n}\sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}}\{rI_t(\mathbf{u}) \int_{B_s} \lambda_s(\mathbf{s};\mathbf{v})\boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} \mid x_i\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{\beta}_s(\mathbf{s}_{ij})^T - 2\xi_3 \mathbf{d}_0^T \boldsymbol{\Omega}_s. \end{aligned}$$

A Taylor expansion of $\lambda_s(\mathbf{s};\hat{\mathbf{v}}_i)$ on the variable $\mathbf{d}_0$ evaluated at the current $\hat{\mathbf{d}}_{0(k-1)}$ gives, as above,

$$\mathsf{D}_{\mathbf{d}_0} Q_1(\ldots, \mathbf{d}_0, \ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)})$$

$$\begin{aligned} \approx\ & -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \hat{I}_{ti} \int_{B_s} \hat{\lambda}_{si}(\mathbf{s})\{1 - \hat{\nu}_{(k-1)}(\mathbf{s})\}\boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} \\ & -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \hat{I}_{ti} \int_{B_s} \hat{\lambda}_{si}(\mathbf{s})\mathbf{d}_0^T\boldsymbol{\beta}_s(\mathbf{s})\boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} \\ & +\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{\beta}_s(\mathbf{s}_{ij})^T - 2\xi_3 \mathbf{d}_0^T \boldsymbol{\Omega}_s, \end{aligned}$$

where $\hat{\lambda}_{ti}(t)$ and $\hat{\lambda}_{si}(\mathbf{s})$ are as above, and $\hat{I}_{ti} = \int_{B_t} \hat{\lambda}_{ti}(t)dt$. The Lagrange condition for the identifiability constraint $\mathbf{a}_{s0}^T \mathbf{d}_0 = 0$ is

$$\mathsf{D}_{\mathbf{d}_0} Q_1(\ldots, \hat{\mathbf{d}}_{0(k)}, \ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) = \kappa \mathbf{a}_{s0}^T,$$

which, transposing both sides, can be written as

$$-\mathbf{Q}\hat{\mathbf{d}}_{0(k)} + \mathbf{b} = \kappa \mathbf{a}_{s0}$$

with

$$\begin{aligned} \mathbf{Q} &= \frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \hat{I}_{ti} \int_{B_s} \hat{\lambda}_{si}(\mathbf{s})\boldsymbol{\beta}_s(\mathbf{s})\boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} + 2\xi_3 \boldsymbol{\Omega}_s, \\ \mathbf{b} &= -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \hat{I}_{ti} \int_{B_s} \hat{\lambda}_{si}(\mathbf{s})\{1 - \hat{\nu}_{(k-1)}(\mathbf{s})\}\boldsymbol{\beta}_s(\mathbf{s}) d\mathbf{s} + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{\beta}_s(\mathbf{s}_{ij}). \end{aligned}$$

Then, including the constraint, we solve the system

$$\begin{pmatrix} \mathbf{Q} & \mathbf{a}_0 \\ \mathbf{a}_0^T & 0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{d}}_{0(k)} \\ \kappa \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix}$$

and obtain $\hat{\mathbf{d}}_{0(k)}$.

For $\mathbf{c}_{p_1}$: The components are estimated sequentially, so we only show the update for $\hat{\mathbf{c}}_{p_1}$. Since

$$\begin{aligned} \log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) &= -rI_t(\mathbf{u})I_s(\mathbf{v}) - \log m_i! \\ &\quad + m_i \log r + \sum_{j=1}^{m_i} \log \lambda_t(t_{ij}; \mathbf{u}) + \sum_{j=1}^{m_i} \log \lambda_s(\mathbf{s}_{ij}; \mathbf{v}) \end{aligned}$$

with $\lambda_t(t; \mathbf{u}) = \exp\{\mu(t) + \boldsymbol{\beta}_t(t)^T \mathbf{C}\mathbf{u}\} = \exp\{\mu(t) + \boldsymbol{\beta}_t(t)^T \sum_{k=1}^{p_1} \mathbf{c}_k u_k\}$, we have

$$\mathsf{D}_{\mathbf{c}_{p_1}} \log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) = -rI_s(\mathbf{v}) \int_{B_t} \mathsf{D}_{\mathbf{c}_{p_1}} \lambda_t(t; \mathbf{u}) dt + \sum_{j=1}^{m_i} u_{p_1} \boldsymbol{\beta}_t(t_{ij})^T,$$

where

$$\begin{aligned} \mathsf{D}_{\mathbf{c}_{p_1}} \lambda_t(t; \mathbf{u}) &= \lambda_t(t; \mathbf{u}) \mathsf{D}_{\mathbf{c}_{p_1}} \log \lambda_t(t; \mathbf{u}) \\ &= \lambda_t(t; \mathbf{u}) u_{p_1} \boldsymbol{\beta}_t(t)^T. \end{aligned}$$

Then

$$\begin{aligned} \mathsf{D}_{\mathbf{c}_{p_1}} Q_1(\ldots, \mathbf{c}_{p_1}, \ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) &= -\frac{1}{n} \sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{ rI_s(\mathbf{v}) \int_{B_t} \lambda_t(t; \mathbf{u}) u_{p_1} \boldsymbol{\beta}_t(t)^T dt \mid x_i \} \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{ u_{p_1} \boldsymbol{\beta}_t(t_{ij})^T \mid x_i \} \\ &\quad - 2\xi_2 \mathbf{c}_{p_1}^T \boldsymbol{\Omega}_t. \end{aligned}$$

Using Laplace approximations to conditional expectations, as above, we have

$$\begin{aligned} E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{ rI_s(\mathbf{v}) \int_{B_t} \lambda_t(t; \mathbf{u}) u_{p_1} \boldsymbol{\beta}_t(t)^T dt \mid x_i \} &\approx \hat{r}_i \quad I_s(\hat{\mathbf{v}}_i) \int_{B_t} \lambda_t(t; \hat{\mathbf{u}}_i) \hat{u}_{ip_1} \boldsymbol{\beta}_t(t)^T dt, \\ E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{ u_{p_1} \boldsymbol{\beta}_t(t_{ij})^T \mid x_i \} &\approx \hat{u}_{ip_1} \boldsymbol{\beta}_t(t_{ij})^T. \end{aligned}$$

A Taylor expansion of $\lambda_t(t;\hat{\mathbf{u}}_i)$ on the variable $\mathbf{c}_{p_1}$ evaluated at the current $\hat{\mathbf{c}}_{p_1(k-1)}$ gives

$$\begin{aligned}
\lambda_t(t;\hat{\mathbf{u}}_i) &\approx \hat{\lambda}_{ti}(t) + \mathsf{D}_{\mathbf{c}_{p_1}}\lambda_t(t;\hat{\mathbf{u}}_i)|_{\hat{\mathbf{c}}_{p_1(k-1)}}(\mathbf{c}_{p_1} - \hat{\mathbf{c}}_{p_1(k-1)}) \\
&= \hat{\lambda}_{ti}(t) + \hat{\lambda}_{ti}(t)\hat{u}_{ip_1}\boldsymbol{\beta}_t(t)^T(\mathbf{c}_{p_1} - \hat{\mathbf{c}}_{p_1(k-1)}) \\
&= \hat{\lambda}_{ti}(t) + \hat{\lambda}_{ti}(t)\hat{u}_{ip_1}\boldsymbol{\beta}_t(t)^T\mathbf{c}_{p_1} - \hat{\lambda}_{ti}(t)\hat{u}_{ip_1}\hat{\phi}_{p_1(k-1)}(t) \\
&= \hat{\lambda}_{ti}(t)\{1 - \hat{u}_{ip_1}\hat{\phi}_{p_1(k-1)}(t)\} + \hat{\lambda}_{ti}(t)\hat{u}_{ip_1}\boldsymbol{\beta}_t(t)^T\mathbf{c}_{p_1}.
\end{aligned}$$

Then

$$\mathsf{D}_{\mathbf{c}_{p_1}}Q_1(\ldots,\mathbf{c}_{p_1},\ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)})$$

$$\begin{aligned}
\approx\ & -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\hat{I}_{si}\int_{B_t}\hat{\lambda}_{ti}(t)\{1 - \hat{u}_{ip_1}\hat{\phi}_{p_1(k-1)}(t)\}\hat{u}_{ip_1}\boldsymbol{\beta}_t(t)^T dt \\
& -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\hat{I}_{si}\int_{B_t}\hat{\lambda}_{ti}(t)\hat{u}_{ip_1}^2\mathbf{c}_{p_1}^T\boldsymbol{\beta}_t(t)\boldsymbol{\beta}_t(t)^T dt \\
& +\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{u}_{ip_1}\boldsymbol{\beta}_t(t_{ij})^T - 2\xi_2\mathbf{c}_{p_1}^T\boldsymbol{\Omega}_t.
\end{aligned}$$

The orthonormality constraints $\mathbf{c}_k^T\mathbf{J}_t\mathbf{c}_{p_1} - \delta_{kp_1}$ can be linearized using the current values of the $\hat{\mathbf{c}}_k$s. So let $\mathbf{A}$ be the matrix with rows $\hat{\mathbf{c}}_k^T\mathbf{J}_t$, $k = 1,\ldots,p_1$, in addition to $\mathbf{a}_{t0}^T$ and $\mathbf{A}_P$ (if periodicity is imposed). Then the Lagrange condition for $\hat{\mathbf{c}}_{p_1(k)}$ is $\mathsf{D}_{\mathrm{vec}\,\mathbf{C}}Q_1(\ldots,\hat{\mathbf{c}}_{p_1(k)},\ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) = \boldsymbol{\kappa}^T\mathbf{A}$ which, transposing both sides, can be written as

$$-\mathbf{Q}\hat{\mathbf{c}}_{p_1(k)} + \mathbf{b} = \mathbf{A}^T\boldsymbol{\kappa}$$

with

$$\begin{aligned}
\mathbf{Q} &= \frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\hat{I}_{si}\int_{B_t}\hat{\lambda}_{ti}(t)\hat{u}_{ip_1}^2\boldsymbol{\beta}_t(t)\boldsymbol{\beta}_t(t)^T dt + 2\xi_2\boldsymbol{\Omega}_t, \\
\mathbf{b} &= -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\hat{I}_{si}\int_{B_t}\hat{\lambda}_{ti}(t)\{1 - \hat{u}_{ip_1}\hat{\phi}_{p_1(k-1)}(t)\}\hat{u}_{ip_1}\boldsymbol{\beta}_t(t)dt + \\
&\quad \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{u}_{ip_1}\boldsymbol{\beta}_t(t_{ij}).
\end{aligned}$$

Then, including the constraints, we solve the system

$$\begin{pmatrix} \mathbf{Q} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{pmatrix}\begin{pmatrix} \hat{\mathbf{c}}_{p_1(k)} \\ \boldsymbol{\kappa} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{f} \end{pmatrix}$$

where $\mathbf{f}$ is a vector with elements $f_{p_1} = 1$ and $f_k = 0$ for $k \neq p_1$, and obtain $\hat{\mathbf{c}}_{p_1(k)}$.

For $\mathbf{d}_{p_2}$: Again, due to the symmetry of (6) on λ_t and λ_s the derivations for $\mathbf{d}_{p_2}$ are

analogous to those for $\mathbf{c}_{p_1}$. We have

$$\begin{aligned}\log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) &= -rI_t(\mathbf{u})I_s(\mathbf{v}) - \log m_i! \\ &\quad + m_i \log r + \sum_{j=1}^{m_i} \log \lambda_t(t_{ij}; \mathbf{u}) + \sum_{j=1}^{m_i} \log \lambda_s(\mathbf{s}_{ij}; \mathbf{v})\end{aligned}$$

with $\lambda_s(\mathbf{s}; \mathbf{v}) = \exp\{\nu(\mathbf{s}) + \boldsymbol{\beta}_s(\mathbf{s})^T \sum_{k=1}^{p_2} v_k \mathbf{d}_k\}$, so

$$\mathsf{D}_{\mathbf{d}_{p_2}} \log f_{\boldsymbol{\theta}}(x_i \mid \mathbf{w}) = -rI_t(\mathbf{u}) \int_{B_s} \mathsf{D}_{\mathbf{d}_{p_2}} \lambda_s(\mathbf{s}; \mathbf{v}) d\mathbf{s} + \sum_{j=1}^{m_i} v_{p_2} \boldsymbol{\beta}_s(\mathbf{s}_{ij})^T,$$

where

$$\begin{aligned}\mathsf{D}_{\mathbf{d}_{p_2}} \lambda_s(\mathbf{s}; \mathbf{v}) &= \lambda_s(\mathbf{s}; \mathbf{v}) \mathsf{D}_{\mathbf{d}_{p_2}} \log \lambda_s(\mathbf{s}; \mathbf{v}) \\ &= \lambda_s(\mathbf{s}; \mathbf{v}) v_{p_2} \boldsymbol{\beta}_s(\mathbf{s})^T.\end{aligned}$$

Then

$$\begin{aligned}\mathsf{D}_{\mathbf{d}_{p_2}} Q_1(\ldots, \mathbf{d}_{p_2}, \ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) &= -\frac{1}{n} \sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{rI_t(\mathbf{u}) \int_{B_s} \lambda_s(\mathbf{s}; \mathbf{v}) v_{p_2} \boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} \mid x_i\} \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{v_{p_2} \boldsymbol{\beta}_s(\mathbf{s}_{ij})^T \mid x_i\} \\ &\quad - 2\xi_4 \mathbf{d}_{p_2}^T \boldsymbol{\Omega}_s.\end{aligned}$$

Using Laplace approximations for the conditional expectations we have

$$\begin{aligned}E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{rI_t(\mathbf{u}) \int_{B_s} \lambda_s(\mathbf{s}; \mathbf{v}) v_{p_2} \boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} \mid x_i\} &\approx \hat{r}_i I_t(\hat{\mathbf{u}}_i) \int_{B_s} \lambda_s(\mathbf{s}; \hat{\mathbf{v}}_i) \hat{v}_{ip_2} \boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s}, \\ E_{\hat{\boldsymbol{\theta}}_{(k-1)}} \{v_{p_2} \boldsymbol{\beta}_s(\mathbf{s}_{ij})^T \mid x_i\} &\approx \hat{v}_{ip_2} \boldsymbol{\beta}_s(\mathbf{s}_{ij})^T.\end{aligned}$$

A Taylor expansion of $\lambda_s(\mathbf{s}; \hat{\mathbf{v}}_i)$ on the variable $\mathbf{d}_{p_2}$ evaluated at the current $\hat{\mathbf{d}}_{p_2(k-1)}$ gives

$$\begin{aligned}\lambda_s(\mathbf{s}; \hat{\mathbf{v}}_i) &\approx \hat{\lambda}_{si}(\mathbf{s}) + \mathsf{D}_{\mathbf{d}_{p_2}} \lambda_s(\mathbf{s}; \hat{\mathbf{v}}_i)|_{\hat{\mathbf{d}}_{p_2(k-1)}} (\mathbf{d}_{p_2} - \hat{\mathbf{d}}_{p_2(k-1)}) \\ &= \hat{\lambda}_{si}(\mathbf{s}) + \hat{\lambda}_{si}(\mathbf{s}) \hat{v}_{ip_2} \boldsymbol{\beta}_s(\mathbf{s})^T (\mathbf{d}_{p_2} - \hat{\mathbf{d}}_{p_2(k-1)}) \\ &= \hat{\lambda}_{si}(\mathbf{s}) + \hat{\lambda}_{si}(\mathbf{s}) \hat{v}_{ip_2} \boldsymbol{\beta}_s(\mathbf{s})^T \mathbf{d}_{p_2} - \hat{\lambda}_{si}(\mathbf{s}) \hat{v}_{ip_2} \hat{\psi}_{p_2(k-1)}(\mathbf{s}) \\ &= \hat{\lambda}_{si}(\mathbf{s}) \{1 - \hat{v}_{ip_2} \hat{\psi}_{p_2(k-1)}(\mathbf{s})\} + \hat{\lambda}_{si}(\mathbf{s}) \hat{v}_{ip_2} \boldsymbol{\beta}_s(\mathbf{s})^T \mathbf{d}_{p_2}.\end{aligned}$$

Then

$$\mathsf{D}_{\mathbf{d}_{p_2}} Q_1(\ldots, \mathbf{d}_{p_2}, \ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)})$$

$$
\begin{aligned}
\approx\ & -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\hat{I}_{ti}\int_{B_s}\hat{\lambda}_{si}(\mathbf{s})\{1-\hat{v}_{ip_2}\hat{\psi}_{p_2(k-1)}(\mathbf{s})\}\hat{v}_{ip_2}\boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} \\
& -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\hat{I}_{ti}\int_{B_s}\hat{\lambda}_{si}(\mathbf{s})\hat{v}_{ip_2}^2\mathbf{d}_{p_2}^T\boldsymbol{\beta}_s(\mathbf{s})\boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} \\
& +\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{v}_{ip_2}\boldsymbol{\beta}_s(\mathbf{s}_{ij})^T - 2\xi_4\mathbf{d}_{p_2}^T\boldsymbol{\Omega}_s.
\end{aligned}
$$

We can write

$$
\mathsf{D}_{\mathbf{d}_{p_2}}Q_1(\ldots,\mathbf{d}_{p_2},\ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)})^T \approx -\mathbf{Q}\mathbf{d}_{p_2}+\mathbf{b}
$$

with

$$
\begin{aligned}
\mathbf{Q} &= \frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\hat{I}_{ti}\int_{B_s}\hat{\lambda}_{si}(\mathbf{s})\hat{v}_{ip_2}^2\boldsymbol{\beta}_s(\mathbf{s})\boldsymbol{\beta}_s(\mathbf{s})^T d\mathbf{s} + 2\xi_4\boldsymbol{\Omega}_s, \\
\mathbf{b} &= -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\hat{I}_{ti}\int_{B_s}\hat{\lambda}_{si}(\mathbf{s})\{1-\hat{v}_{ip_2}\hat{\psi}_{p_2(k-1)}(\mathbf{s})\}\hat{v}_{ip_2}\boldsymbol{\beta}_s(\mathbf{s})d\mathbf{s} + \\
& \quad \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{v}_{ip_2}\boldsymbol{\beta}_s(\mathbf{s}_{ij}).
\end{aligned}
$$

The constraints are handled as before. Let $\mathbf{A}$ be the matrix with rows $\hat{\mathbf{d}}_k^T\mathbf{J}_s$, $k=1,\ldots,p_2$, in addition to $\mathbf{a}_{s0}^T$. Then the Lagrange condition for $\hat{\mathbf{d}}_{p_2(k)}$ is $\mathsf{D}_{\mathbf{d}_{p_2}}Q_1(\ldots,\hat{\mathbf{d}}_{p_2(k)},\ldots \mid \hat{\boldsymbol{\theta}}_{(k-1)}) = \boldsymbol{\kappa}^T\mathbf{A}$. Transposing both sides and using then above approximations we get the equation

$$
-\mathbf{Q}\hat{\mathbf{d}}_{p_2(k)}+\mathbf{b} = \mathbf{A}^T\boldsymbol{\kappa},
$$

which together with the constraints can be written as a single system

$$
\begin{pmatrix} \mathbf{Q} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{pmatrix}\begin{pmatrix} \hat{\mathbf{d}}_{p_2(k)} \\ \boldsymbol{\kappa} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{f} \end{pmatrix},
$$

where $\mathbf{f}$ is a vector with elements $f_{p_2}=1$ and $f_k=0$ for $k\neq p_2$. Solving this linear system gives the updated $\hat{\mathbf{d}}_{p_2(k)}$.

For $\boldsymbol{\Sigma}$: We have

$$
\begin{aligned}
\log f_{\boldsymbol{\theta}}(\mathbf{w}) &\propto -\frac{1}{2}\log\det\boldsymbol{\Sigma} - \frac{1}{2}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w} \\
&= -\frac{1}{2}\log\det\boldsymbol{\Sigma} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{w}\mathbf{w}^T),
\end{aligned}
$$

so

$$
Q_2(\boldsymbol{\Sigma} \mid \hat{\boldsymbol{\theta}}_{(k-1)}) \propto -\frac{1}{2}\log\det\boldsymbol{\Sigma} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})
$$

where

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}_{(k-1)}}\left(\mathbf{w}\mathbf{w}^T \mid x_i\right).$$

This $Q_2(\boldsymbol{\Sigma} \mid \hat{\boldsymbol{\theta}}_{(k-1)})$ is the classical log-likelihood function of a multivariate normal density, and it is well-known that the (unconstrained) maximizer is $\mathbf{S}$. However, $\mathbf{S}$ must be rotated to satisfy the constraints that $\boldsymbol{\Sigma}_{uu}$ and $\boldsymbol{\Sigma}_{vv}$ are diagonal, while maintaining the positive-definitness of the whole $\hat{\boldsymbol{\Sigma}}$. To this end we compute the spectral decompositions of the blocks $\mathbf{S}_{uu}$ and $\mathbf{S}_{vv}$,

$$\begin{aligned}
\boldsymbol{\Gamma}_1\mathbf{L}_1\boldsymbol{\Gamma}_1^T &= \mathbf{S}_{uu}, \\
\boldsymbol{\Gamma}_2\mathbf{L}_2\boldsymbol{\Gamma}_2^T &= \mathbf{S}_{vv},
\end{aligned}$$

with the $\boldsymbol{\Gamma}$s orthogonal and the $\mathbf{L}$s diagonal, and let

$$\hat{\boldsymbol{\Sigma}}_{(k)} = \begin{pmatrix} 1 & \mathbf{0}^T & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{\Gamma}_1^T & \mathbf{O} \\ \mathbf{0} & \mathbf{O} & \boldsymbol{\Gamma}_2^T \end{pmatrix} \mathbf{S} \begin{pmatrix} 1 & \mathbf{0}^T & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{\Gamma}_1 & \mathbf{O} \\ \mathbf{0} & \mathbf{O} & \boldsymbol{\Gamma}_2 \end{pmatrix}.$$

Then the blocks $\hat{\boldsymbol{\Sigma}}_{(k),uu}$ and $\hat{\boldsymbol{\Sigma}}_{(k),vv}$ are diagonal and equal to $\mathbf{L}_1$ and $\mathbf{L}_2$, respectively. Then $\hat{\boldsymbol{\sigma}}^2_{u(k)} = \operatorname{diag}\hat{\boldsymbol{\Sigma}}_{(k),uu}$, $\hat{\boldsymbol{\sigma}}^2_{v(k)} = \operatorname{diag}\hat{\boldsymbol{\Sigma}}_{(k),vv}$ and $\hat{\boldsymbol{\Sigma}}_{uv(k)} = \hat{\boldsymbol{\Sigma}}_{(k),uv}$. Similarly for $\hat{\sigma}^2_{z(k)}$, $\hat{\boldsymbol{\sigma}}_{zu(k)}$ and $\hat{\boldsymbol{\sigma}}_{zv(k)}$.

The respective component scores and component basis coefficients must be rotated as well, to preserve the values of $\hat{\boldsymbol{\phi}}(t)^T\hat{\mathbf{u}}_i$ and $\hat{\boldsymbol{\psi}}(\mathbf{s})^T\hat{\mathbf{v}}_i$ and preserve the fact that $\hat{\boldsymbol{\Sigma}}_{(k)}$ is the covariance matrix of the $\hat{\mathbf{w}}_i$s:

$$\begin{aligned}
\hat{\mathbf{u}}_i &\longleftarrow \boldsymbol{\Gamma}_1^T\hat{\mathbf{u}}_i, \\
\hat{\mathbf{v}}_i &\longleftarrow \boldsymbol{\Gamma}_2^T\hat{\mathbf{v}}_i, \\
\hat{\mathbf{C}} &\longleftarrow \hat{\mathbf{C}}\boldsymbol{\Gamma}_1, \\
\hat{\mathbf{D}} &\longleftarrow \hat{\mathbf{D}}\boldsymbol{\Gamma}_2.
\end{aligned}$$

1.4 Laplace's approximation of integrals

The marginal densities $f(x)$ are computed by Laplace approximation. We have

$$\begin{aligned}
f(x) &= \iint f(x \mid \mathbf{w}) f(\mathbf{w}) d\mathbf{w} \\
&= \iint \exp g(\mathbf{w}) d\mathbf{w}
\end{aligned}$$

with

$$g(\mathbf{w}) = \log f(x \mid \mathbf{w}) + \log f(\mathbf{w}).$$

If $\hat{\mathbf{w}} = \arg\max g(\mathbf{w})$ then $g(\mathbf{w}) \approx g(\hat{\mathbf{w}}) + .5(\mathbf{w} - \hat{\mathbf{w}})^T \mathsf{H} g(\hat{\mathbf{w}})(\mathbf{w} - \hat{\mathbf{w}})$ and

$$f(x) \approx \exp\{g(\hat{\mathbf{w}})\}(2\pi)^{p/2} \det(\mathbf{S})^{1/2}$$

with $p = p_1 + p_2 + 1$ and

$$\mathbf{S} = \{-\mathsf{H} g(\hat{\mathbf{w}})\}^{-1}.$$

In effect, we are approximating

$$f(x \mid \mathbf{w}) f(\mathbf{w}) \approx \exp\{g(\hat{\mathbf{w}})\}(2\pi)^{p/2} \det(\mathbf{S})^{1/2} \varphi_{(\hat{\mathbf{w}}, \mathbf{S})}(\mathbf{w})$$

where $\varphi_{(\hat{\mathbf{w}}, \mathbf{S})}(\mathbf{w})$ denotes the pdf of a $N_p(\hat{\mathbf{w}}, \mathbf{S})$, so $\mathbf{W} \mid x \approx N_p(\hat{\mathbf{w}}, \mathbf{S})$. Then we can also approximate the moments:

$$\begin{aligned} E(\mathbf{W} \mid x) &\approx \hat{\mathbf{w}}, \\ E(\mathbf{W}\mathbf{W}^T \mid x) &\approx \mathbf{S} + \hat{\mathbf{w}}\hat{\mathbf{w}}^T. \end{aligned}$$

We find $\hat{\mathbf{w}}$ by (a few steps of) Newton–Raphson for each x_i. Since

$$\begin{aligned} g(\mathbf{w}) &= -e^{\tau + z} I_t(\mathbf{u}) I_s(\mathbf{v}) - \log m! + m(\tau + z) \\ &\quad + \sum_{j=1}^{m} \log \lambda_t(t_j; \mathbf{u}) + \sum_{j=1}^{m} \log \lambda_s(\mathbf{s}_j; \mathbf{v}) \\ &\quad - \frac{p}{2} \log 2\pi - \frac{1}{2} \log \det \boldsymbol{\Sigma} - \frac{1}{2} \mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w}, \end{aligned}$$

then, defining

$$\begin{aligned} I_t'(\mathbf{u}) &= \int \lambda_t(t; \mathbf{u}) \boldsymbol{\phi}(t) dt, \\ I_s'(\mathbf{v}) &= \int \lambda_s(\mathbf{s}; \mathbf{v}) \boldsymbol{\psi}(\mathbf{s}) d\mathbf{s}, \\ I_t''(\mathbf{u}) &= \int \lambda_t(t; \mathbf{u}) \boldsymbol{\phi}(t) \boldsymbol{\phi}(t)^T dt, \\ I_s''(\mathbf{v}) &= \int \lambda_s(\mathbf{s}; \mathbf{v}) \boldsymbol{\psi}(\mathbf{s}) \boldsymbol{\psi}(\mathbf{s})^T d\mathbf{s}, \end{aligned}$$

the derivatives with respect to $\mathbf{w} = (z, \mathbf{u}, \mathbf{v})$ are

$$\nabla g(\mathbf{w}) = \begin{bmatrix} -r I_t(\mathbf{u}) I_s(\mathbf{v}) + m \\ -r I_s(\mathbf{v}) I_t'(\mathbf{u}) + \sum_{j=1}^{m} \boldsymbol{\phi}(t_j) \\ -r I_t(\mathbf{u}) I_s'(\mathbf{v}) + \sum_{j=1}^{m} \boldsymbol{\psi}(\mathbf{s}_j) \end{bmatrix} - \boldsymbol{\Sigma}^{-1} \mathbf{w}$$

and

$$\mathsf{H}g(\mathbf{w}) = \begin{bmatrix} -rI_t(\mathbf{u})I_s(\mathbf{v}) & -rI_s(\mathbf{v})I_t'(\mathbf{u})^T & -rI_t(\mathbf{u})I_s'(\mathbf{v})^T \\ -rI_t'(\mathbf{u})I_s(\mathbf{v}) & -rI_s(\mathbf{v})I_t''(\mathbf{u}) & -rI_t'(\mathbf{u})I_s'(\mathbf{v})^T \\ -rI_s'(\mathbf{v})I_t(\mathbf{u}) & -rI_s'(\mathbf{v})I_t'(\mathbf{u})^T & -rI_t(\mathbf{u})I_s''(\mathbf{v}) \end{bmatrix} - \boldsymbol{\Sigma}^{-1}.$$

2 Asymptotics

2.1 Explicit Fisher's information matrix

Fisher's information matrix $\mathbf{F}_0 = E_{\boldsymbol{\theta}_0}\{\nabla \log f_{\boldsymbol{\theta}_0}(X) \nabla \log f_{\boldsymbol{\theta}_0}(X)^T\}$, used in the asymptotic results below, is estimated by

$$\hat{\mathbf{F}} = \frac{1}{n}\sum_{i=1}^{n} \nabla \log f_{\hat{\boldsymbol{\theta}}}(x_i) \nabla \log f_{\hat{\boldsymbol{\theta}}}(x_i)^T.$$

Here we derive $\nabla \log f_{\boldsymbol{\theta}}(x)$ by blocks for $\boldsymbol{\theta} = (\boldsymbol{\sigma}_{zu}, \boldsymbol{\sigma}_{zv}, \operatorname{vec} \boldsymbol{\Sigma}_{uv}, \tau, \sigma_z^2, \mathbf{c}_0, \operatorname{vec} \mathbf{C}, \boldsymbol{\sigma}_u^2, \mathbf{d}_0, \operatorname{vec} \mathbf{D}, \boldsymbol{\sigma}_v^2)$.

► For $\boldsymbol{\sigma}_{zu}$, since only $f(\mathbf{w})$ depends on $\boldsymbol{\sigma}_{zu}$, we have

$$\begin{aligned} & \nabla_{\boldsymbol{\sigma}_{zu}} \log f_{\boldsymbol{\theta}}(x) \\ = \;& \frac{1}{f_{\boldsymbol{\theta}}(x)} \iint f(x \mid \mathbf{w}) \nabla_{\boldsymbol{\sigma}_{zu}} f(\mathbf{w}) \; d\mathbf{w} \\ = \;& \iint \frac{\nabla_{\boldsymbol{\sigma}_{zu}} f(\mathbf{w})}{f(\mathbf{w})} \; \frac{f(x \mid \mathbf{w}) f(\mathbf{w})}{f_{\boldsymbol{\theta}}(x)} \; d\mathbf{w} \\ = \;& \iint \nabla_{\boldsymbol{\sigma}_{zu}} \log f(\mathbf{w}) \; f(\mathbf{w} \mid x) \; d\mathbf{w}. \end{aligned}$$

Since

$$\log f(\mathbf{w}) \propto -\frac{1}{2} \log \det \boldsymbol{\Sigma} - \frac{1}{2} \mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w},$$

the differential with respect to $\boldsymbol{\Sigma}$ is

$$\mathsf{d} \log f(\mathbf{w}) = -\frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \mathsf{d}\boldsymbol{\Sigma}) + \frac{1}{2} \mathbf{w}^T \boldsymbol{\Sigma}^{-1} (\mathsf{d}\boldsymbol{\Sigma}) \boldsymbol{\Sigma}^{-1} \mathbf{w}.$$

Differentiating specifically with respect to $\boldsymbol{\sigma}_{zu}$,

$$\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} 0 & \mathsf{d}\boldsymbol{\sigma}_{zu}^T & \mathbf{0}^T \\ \mathsf{d}\boldsymbol{\sigma}_{zu} & \mathbf{O} & \mathbf{O} \\ \mathbf{0} & \mathbf{O} & \mathbf{O} \end{pmatrix}.$$

If we split $\boldsymbol{\Sigma}^{-1}$ into nine blocks commensurate with those of $\boldsymbol{\Sigma}$, denoted by $\boldsymbol{\Sigma}_{11}^{-1}$, $\boldsymbol{\Sigma}_{12}^{-1}$, …, $\boldsymbol{\Sigma}_{33}^{-1}$, and the vector $\boldsymbol{\Sigma}^{-1}\mathbf{w}$ into three sub-vectors of respectively one, p_1 and p_2 elements,

we have

$$\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{12}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu} & \boldsymbol{\Sigma}_{11}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu}^T & \mathbf{0}^T \\ \boldsymbol{\Sigma}_{22}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu} & \boldsymbol{\Sigma}_{21}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu}^T & \mathbf{O} \\ \boldsymbol{\Sigma}_{32}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu} & \boldsymbol{\Sigma}_{31}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu}^T & \mathbf{O} \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{w}^T\boldsymbol{\Sigma}^{-1}(\mathsf{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{w} &= 2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T(\mathsf{d}\boldsymbol{\sigma}_{zu}^T)(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2 \\ &= 2\,\mathrm{tr}\{\mathsf{d}\boldsymbol{\sigma}_{zu}^T(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T\} \\ &= 2\mathsf{d}\boldsymbol{\sigma}_{zu}^T\,\mathrm{vec}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T\}. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma}) &= \mathrm{tr}(\boldsymbol{\Sigma}_{12}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu}) + \mathrm{tr}(\boldsymbol{\Sigma}_{21}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu}^T) \\ &= 2\,\mathrm{tr}(\mathsf{d}\boldsymbol{\sigma}_{zu}^T\boldsymbol{\Sigma}_{21}^{-1}) \\ &= 2\mathsf{d}\boldsymbol{\sigma}_{zu}^T\,\mathrm{vec}(\boldsymbol{\Sigma}_{21}^{-1}), \end{aligned}$$

so

$$\mathsf{d}\log f(\mathbf{w}) = -\mathsf{d}\boldsymbol{\sigma}_{zu}^T\,\mathrm{vec}(\boldsymbol{\Sigma}_{21}^{-1}) + \mathsf{d}\boldsymbol{\sigma}_{zu}^T\,\mathrm{vec}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T\},$$

which implies

$$\nabla_{\boldsymbol{\sigma}_{zu}} f(\mathbf{w}) = -\,\mathrm{vec}(\boldsymbol{\Sigma}_{21}^{-1}) + \mathrm{vec}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T\}$$

and then

$$\boxed{\nabla_{\boldsymbol{\sigma}_{zu}} \log f_{\boldsymbol{\theta}}(x) = -\,\mathrm{vec}(\boldsymbol{\Sigma}_{21}^{-1}) + \mathrm{vec}\,\mathbb{E}_{\boldsymbol{\theta}}\left\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T \mid x\right\}.}$$

The second term can be written out more explicitly in terms of $\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{w}\mathbf{w}^T \mid x)$: since $(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2 = [\mathbf{0}, \mathbf{I}_{p_1}, \mathbf{O}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}$ and $(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1 = \left[1, \mathbf{0}^T, \mathbf{0}^T\right]\boldsymbol{\Sigma}^{-1}\mathbf{w}$, we have

$$\begin{aligned} (\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T &= [\mathbf{0}, \mathbf{I}_{p_1}, \mathbf{O}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\begin{bmatrix} 1 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \\ &= \left[\boldsymbol{\Sigma}_{21}^{-1}, \boldsymbol{\Sigma}_{22}^{-1}, \boldsymbol{\Sigma}_{23}^{-1}\right]\mathbf{w}\mathbf{w}^T\begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} \\ \boldsymbol{\Sigma}_{21}^{-1} \\ \boldsymbol{\Sigma}_{31}^{-1} \end{bmatrix} \end{aligned}$$

and then we take $\mathbb{E}_{\boldsymbol{\theta}}$.

► For $\boldsymbol{\sigma}_{zv}$ we proceed as above, since only $f(\mathbf{w})$ depends on $\boldsymbol{\sigma}_{zv}$. We have

$$\nabla_{\boldsymbol{\sigma}_{zv}} \log f_{\boldsymbol{\theta}}(x) = \iint \nabla_{\boldsymbol{\sigma}_{zv}} \log f(\mathbf{w})\; f(\mathbf{w} \mid x)\; d\mathbf{w}$$

and $\nabla_{\boldsymbol{\sigma}_{zv}} \log f(\mathbf{w})$ is derived via differentials as before. Differentiating $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\sigma}_{zv}$ we get

$$\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} 0 & \mathbf{0}^T & \mathsf{d}\boldsymbol{\sigma}_{zv}^T \\ \mathbf{0} & \mathbf{O} & \mathbf{O} \\ \mathsf{d}\boldsymbol{\sigma}_{zv} & \mathbf{O} & \mathbf{O} \end{pmatrix}$$

and then

$$\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{13}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu} & \mathbf{0}^T & \boldsymbol{\Sigma}_{11}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu}^T \\ \boldsymbol{\Sigma}_{23}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu} & \mathbf{O} & \boldsymbol{\Sigma}_{21}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu}^T \\ \boldsymbol{\Sigma}_{33}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu} & \mathbf{O} & \boldsymbol{\Sigma}_{31}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zu}^T \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{w}^T\boldsymbol{\Sigma}^{-1}(\mathsf{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{w} &= 2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T(\mathsf{d}\boldsymbol{\sigma}_{zv}^T)(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3 \\ &= 2\,\mathrm{tr}\{\mathsf{d}\boldsymbol{\sigma}_{zv}^T(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T\} \\ &= 2\mathsf{d}\boldsymbol{\sigma}_{zv}^T\,\mathrm{vec}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T\}. \end{aligned}$$

Then

$$\begin{aligned} \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma}) &= \mathrm{tr}(\boldsymbol{\Sigma}_{13}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zv}) + \mathrm{tr}(\boldsymbol{\Sigma}_{31}^{-1}\mathsf{d}\boldsymbol{\sigma}_{zv}^T) \\ &= 2\,\mathrm{tr}(\mathsf{d}\boldsymbol{\sigma}_{zv}^T\boldsymbol{\Sigma}_{31}^{-1}) \\ &= 2\mathsf{d}\boldsymbol{\sigma}_{zv}^T\,\mathrm{vec}(\boldsymbol{\Sigma}_{31}^{-1}), \end{aligned}$$

and following the same steps as above, we get

$$\boxed{\nabla_{\boldsymbol{\sigma}_{zv}} \log f_{\boldsymbol{\theta}}(x) = -\,\mathrm{vec}(\boldsymbol{\Sigma}_{31}^{-1}) + \mathrm{vec}\,\mathbb{E}_{\boldsymbol{\theta}}\left\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T \mid x\right\}.}$$

The second term can again be written out more explicitly in terms of $\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{w}\mathbf{w}^T \mid x)$: since $(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3 = [\mathbf{0}, \mathbf{O}, \mathbf{I}_{p_2}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}$ and $(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1 = \left[1, \mathbf{0}^T, \mathbf{0}^T\right]\boldsymbol{\Sigma}^{-1}\mathbf{w}$, we have

$$\begin{aligned} (\boldsymbol{\Sigma}^{-1}\mathbf{w})_3(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^T &= [\mathbf{0}, \mathbf{O}, \mathbf{I}_{p_2}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\begin{bmatrix} 1 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \\ &= \left[\boldsymbol{\Sigma}_{31}^{-1}, \boldsymbol{\Sigma}_{32}^{-1}, \boldsymbol{\Sigma}_{33}^{-1}\right]\mathbf{w}\mathbf{w}^T\begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} \\ \boldsymbol{\Sigma}_{21}^{-1} \\ \boldsymbol{\Sigma}_{31}^{-1} \end{bmatrix} \end{aligned}$$

and then we take $\mathbb{E}_{\boldsymbol{\theta}}$.

▶ For $\mathrm{vec}\,\boldsymbol{\Sigma}_{uv}$ we proceed as above again, since only $f(\mathbf{w})$ depends on $\boldsymbol{\Sigma}_{uv}$. We have

$$\nabla_{\mathrm{vec}\,\boldsymbol{\Sigma}_{uv}} \log f_{\boldsymbol{\theta}}(x) = \iint \nabla_{\mathrm{vec}\,\boldsymbol{\Sigma}_{uv}} \log f(\mathbf{w})\; f(\mathbf{w} \mid x)\; d\mathbf{w},$$

and $\nabla_{\mathrm{vec}\,\boldsymbol{\Sigma}_{uv}} \log f(\mathbf{w})$ is obtained via differentials. Differentiating $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Sigma}_{uv}$ we get

$$\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} 0 & \mathbf{0}^T & \mathbf{0}^T \\ \mathbf{0} & \mathbf{O} & \mathsf{d}\boldsymbol{\Sigma}_{uv} \\ \mathbf{0} & \mathsf{d}\boldsymbol{\Sigma}_{uv}^T & \mathbf{O} \end{pmatrix}.$$

Then

$$\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} 0 & \boldsymbol{\Sigma}_{13}^{-1}\mathsf{d}\boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{12}^{-1}\mathsf{d}\boldsymbol{\Sigma}_{uv} \\ \mathbf{0} & \boldsymbol{\Sigma}_{23}^{-1}\mathsf{d}\boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{22}^{-1}\mathsf{d}\boldsymbol{\Sigma}_{uv} \\ \mathbf{0} & \boldsymbol{\Sigma}_{33}^{-1}\mathsf{d}\boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_{32}^{-1}\mathsf{d}\boldsymbol{\Sigma}_{uv} \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{w}^T\boldsymbol{\Sigma}^{-1}(\mathsf{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{w} &= 2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^T(\mathsf{d}\boldsymbol{\Sigma}_{uv}^T)(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2 \\ &= 2\,\mathrm{tr}\{\mathsf{d}\boldsymbol{\Sigma}_{uv}^T(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^T\} \\ &= 2\,\mathrm{vec}(\mathsf{d}\boldsymbol{\Sigma}_{uv})^T\,\mathrm{vec}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^T\}, \end{aligned}$$

so

$$\begin{aligned} \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma}) &= \mathrm{tr}(\boldsymbol{\Sigma}_{23}^{-1}\mathsf{d}\boldsymbol{\Sigma}_{uv}^T) + \mathrm{tr}(\boldsymbol{\Sigma}_{32}^{-1}\mathsf{d}\boldsymbol{\Sigma}_{uv}) \\ &= 2\,\mathrm{tr}(\mathsf{d}\boldsymbol{\Sigma}_{uv}^T\boldsymbol{\Sigma}_{23}^{-1}) \\ &= 2\,\mathrm{vec}(\mathsf{d}\boldsymbol{\Sigma}_{uv})^T\,\mathrm{vec}(\boldsymbol{\Sigma}_{23}^{-1}) \end{aligned}$$

and then

$$\mathsf{d}\log f(\mathbf{w}) = -\,\mathrm{vec}(\mathsf{d}\boldsymbol{\Sigma}_{uv})^T\,\mathrm{vec}(\boldsymbol{\Sigma}_{23}^{-1}) + \mathrm{vec}(\mathsf{d}\boldsymbol{\Sigma}_{uv})^T\,\mathrm{vec}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^T\},$$

which implies

$$\nabla_{\mathrm{vec}\,\boldsymbol{\Sigma}_{uv}} f(\mathbf{w}) = -\,\mathrm{vec}(\boldsymbol{\Sigma}_{23}^{-1}) + \mathrm{vec}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^T\}$$

and therefore

$$\boxed{\nabla_{\mathrm{vec}\,\boldsymbol{\Sigma}_{uv}} \log f_{\boldsymbol{\theta}}(x) = -\,\mathrm{vec}(\boldsymbol{\Sigma}_{23}^{-1}) + \mathrm{vec}\,\mathbb{E}_{\boldsymbol{\theta}}\left\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^T \mid x\right\}.}$$

As before, the second term can be written more explicitly in terms of $\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{w}\mathbf{w}^T \mid x)$: since

$(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2 = [\mathbf{0}, \mathbf{I}_{p_1}, \mathbf{O}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}$ and $(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3 = [\mathbf{0}, \mathbf{O}, \mathbf{I}_{p_2}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}$, we have

$$\begin{aligned}(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^T &= [\mathbf{0}, \mathbf{I}_{p_1}, \mathbf{O}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\begin{bmatrix}\mathbf{0}^T\\ \mathbf{O}\\ \mathbf{I}_{p_2}\end{bmatrix}\\ &= \left[\boldsymbol{\Sigma}_{21}^{-1}, \boldsymbol{\Sigma}_{22}^{-1}, \boldsymbol{\Sigma}_{23}^{-1}\right]\mathbf{w}\mathbf{w}^T\begin{bmatrix}\boldsymbol{\Sigma}_{13}^{-1}\\ \boldsymbol{\Sigma}_{23}^{-1}\\ \boldsymbol{\Sigma}_{33}^{-1}\end{bmatrix}\end{aligned}$$

and then we take $\mathbb{E}_{\boldsymbol{\theta}}$.

▶ For τ, since only $f(x \mid \mathbf{w})$ depends on this parameter, we have

$$\frac{\partial}{\partial\tau}\log f_{\boldsymbol{\theta}}(x) = \iint \frac{\partial}{\partial\tau}\log f(x \mid \mathbf{w})\; f(\mathbf{w} \mid x)\; d\mathbf{w}.$$

Since

$$\log f(x \mid \mathbf{w}) = -rI_t(\mathbf{u})I_s(\mathbf{v}) - \log m! + m\log r + \sum_{j=1}^{m}\log\lambda_t(t_j;\mathbf{u}) + \sum_{j=1}^{m}\log\lambda_s(\mathbf{s}_j;\mathbf{v})$$

with $r = \exp(\tau + z)$, we have

$$\frac{\partial}{\partial\tau}\log f(x \mid \mathbf{w}) = -rI_t(\mathbf{u})I_s(\mathbf{v}) + m$$

and then

$$\boxed{\frac{\partial}{\partial\tau}\log f_{\boldsymbol{\theta}}(x) = -\mathbb{E}_{\boldsymbol{\theta}}\{rI_t(\mathbf{u})I_s(\mathbf{v}) \mid x\} + m.}$$

▶ For σ_z^2, only $f(\mathbf{w})$ depends on this parameter, so

$$\frac{\partial}{\partial\sigma_z^2}\log f_{\boldsymbol{\theta}}(x) = \iint \frac{\partial}{\partial\sigma_z^2}\log f(\mathbf{w})\; f(\mathbf{w} \mid x)\; d\mathbf{w}.$$

As before,

$$\mathsf{d}\log f(\mathbf{w}) = -\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma}) + \frac{1}{2}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}(\mathsf{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{w},$$

and differentiating with respect to σ_z^2 we have

$$\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix}\mathsf{d}\sigma_z^2 & \mathbf{0}^T & \mathbf{0}^T\\ \mathbf{0} & \mathbf{O} & \mathbf{O}\\ \mathbf{0} & \mathbf{O} & \mathbf{O}\end{pmatrix}.$$

Then

$$\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1}\mathsf{d}\sigma_z^2 & \mathbf{0}^T & \mathbf{0}^T \\ \boldsymbol{\Sigma}_{21}^{-1}\mathsf{d}\sigma_z^2 & \mathbf{O} & \mathbf{O} \\ \boldsymbol{\Sigma}_{31}^{-1}\mathsf{d}\sigma_z^2 & \mathbf{O} & \mathbf{O} \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{w}^T\boldsymbol{\Sigma}^{-1}(\mathsf{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{w} &= (\boldsymbol{\Sigma}^{-1}\mathbf{w})_1(\mathsf{d}\sigma_z^2)(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1 \\ &= (\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^2\mathsf{d}\sigma_z^2, \end{aligned}$$

so

$$\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}_{11}^{-1}\mathsf{d}\sigma_z^2.$$

Then

$$\mathsf{d}\log f(\mathbf{w}) = -\frac{1}{2}\boldsymbol{\Sigma}_{11}^{-1}\mathsf{d}\sigma_z^2 + \frac{1}{2}(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^2\mathsf{d}\sigma_z^2,$$

so

$$\frac{\partial}{\partial\sigma_z^2}\log f(\mathbf{w}) = -\frac{1}{2}\boldsymbol{\Sigma}_{11}^{-1} + \frac{1}{2}(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^2$$

and then

$$\boxed{\frac{\partial}{\partial\sigma_z^2}\log f_{\boldsymbol{\theta}}(x) = -\frac{1}{2}\boldsymbol{\Sigma}_{11}^{-1} + \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^2 \mid x\}.}$$

The second term can be written out more explicitly in terms of $\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{w}\mathbf{w}^T \mid x)$: since $(\boldsymbol{\Sigma}^{-1}\mathbf{w})_1 = \left[1, \mathbf{0}^T, \mathbf{0}^T\right]\boldsymbol{\Sigma}^{-1}\mathbf{w}$, we have

$$\begin{aligned} (\boldsymbol{\Sigma}^{-1}\mathbf{w})_1^2 &= \left[1, \mathbf{0}^T, \mathbf{0}^T\right]\boldsymbol{\Sigma}^{-1}\mathbf{w}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\begin{bmatrix} 1 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \\ &= \left[\boldsymbol{\Sigma}_{11}^{-1}, \boldsymbol{\Sigma}_{12}^{-1}, \boldsymbol{\Sigma}_{13}^{-1}\right]\mathbf{w}\mathbf{w}^T\begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} \\ \boldsymbol{\Sigma}_{21}^{-1} \\ \boldsymbol{\Sigma}_{31}^{-1} \end{bmatrix}. \end{aligned}$$

and then we take $\mathbb{E}_{\boldsymbol{\theta}}$.

► For $\mathbf{c}_0$, since only $f(x \mid \mathbf{w})$ depends on $\mathbf{c}_0$, we have

$$\nabla_{\mathbf{c}_0}\log f_{\boldsymbol{\theta}}(x) = \iint \nabla_{\mathbf{c}_0}\log f(x \mid \mathbf{w})\ f(\mathbf{w} \mid x)\ d\mathbf{w}.$$

Here

$$\log f(x \mid \mathbf{w}) = -rI_t(\mathbf{u})I_s(\mathbf{v}) - \log m! + m\log r + \sum_{j=1}^{m}\log\lambda_t(t_j;\mathbf{u}) + \sum_{j=1}^{m}\log\lambda_s(\mathbf{s}_j;\mathbf{v})$$

with $I_t(\mathbf{u}) = \int \lambda_t(t;\mathbf{u})dt$ and $\lambda_t(t;\mathbf{u}) = \exp\{\mathbf{c}_0^T\boldsymbol{\beta}_t(t) + \mathbf{u}^T\boldsymbol{\phi}(t)\}$, so

$$\nabla_{\mathbf{c}_0}\log f(x \mid \mathbf{w}) = -rI_s(\mathbf{v})\int \lambda_t(t;\mathbf{u})\boldsymbol{\beta}_t(t)dt + \sum_{j=1}^{m}\boldsymbol{\beta}_t(t_j)$$

and then

$$\boxed{\nabla_{\mathbf{c}_0}\log f_{\boldsymbol{\theta}}(x) = -\mathbb{E}_{\boldsymbol{\theta}}\{rI_s(\mathbf{v})\textstyle\int \lambda_t(t;\mathbf{u})\boldsymbol{\beta}_t(t)dt \mid x\} + \sum_{j=1}^{m}\boldsymbol{\beta}_t(t_j).}$$

▶ For $\operatorname{vec}\mathbf{C}$, again only $f(x \mid \mathbf{w})$ depends on $\operatorname{vec}\mathbf{C}$, so

$$\nabla_{\operatorname{vec}\mathbf{C}}\log f_{\boldsymbol{\theta}}(x) = \iint \nabla_{\operatorname{vec}\mathbf{C}}\log f(x \mid \mathbf{w})\ f(\mathbf{w} \mid x)\ d\mathbf{w}$$

as above. Since $\lambda_t(t;\mathbf{u}) = \exp\{\mu(t) + \boldsymbol{\beta}_t(t)^T\mathbf{C}\mathbf{u}\} = \exp[\mu(t) + \{\mathbf{u}^T \otimes \boldsymbol{\beta}_t(t)^T\}\operatorname{vec}\mathbf{C}]$, we have

$$\nabla_{\operatorname{vec}\mathbf{C}}\log f(x \mid \mathbf{w}) = -rI_s(\mathbf{v})\int \lambda_t(t;\mathbf{u})\{\mathbf{u}\otimes\boldsymbol{\beta}_t(t)\}dt + \sum_{j=1}^{m}\{\mathbf{u}\otimes\boldsymbol{\beta}_t(t_j)\}$$

and then

$$\boxed{\nabla_{\operatorname{vec}\mathbf{C}}\log f_{\boldsymbol{\theta}}(x) = -\mathbb{E}_{\boldsymbol{\theta}}[rI_s(\mathbf{v})\textstyle\int \lambda_t(t;\mathbf{u})\{\mathbf{u}\otimes\boldsymbol{\beta}_t(t)\}dt \mid x] + \mathbb{E}_{\boldsymbol{\theta}}(\mathbf{u} \mid x)\otimes\sum_{j=1}^{m}\boldsymbol{\beta}_t(t_j).}$$

▶ For $\boldsymbol{\sigma}_u^2$, only $f(\mathbf{w})$ depends on this parameter, so

$$\nabla_{\boldsymbol{\sigma}_u^2}\log f_{\boldsymbol{\theta}}(x) = \iint \nabla_{\boldsymbol{\sigma}_u^2}\log f(\mathbf{w})\ f(\mathbf{w} \mid x)\ d\mathbf{w}.$$

As before,

$$\mathsf{d}\log f(\mathbf{w}) = -\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma}) + \frac{1}{2}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}(\mathsf{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{w},$$

and differentiating with respect to $\boldsymbol{\sigma}_u^2$ we get

$$\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} 0 & \mathbf{0}^T & \mathbf{0}^T \\ \mathbf{0} & \operatorname{diag}(\mathsf{d}\boldsymbol{\sigma}_u^2) & \mathbf{O} \\ \mathbf{0} & \mathbf{O} & \mathbf{O} \end{pmatrix}.$$

Then

$$\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma} = \begin{pmatrix} 0 & \boldsymbol{\Sigma}_{12}^{-1}\operatorname{diag}(\mathsf{d}\boldsymbol{\sigma}_u^2) & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1}\operatorname{diag}(\mathsf{d}\boldsymbol{\sigma}_u^2) & \mathbf{O} \\ \mathbf{0} & \boldsymbol{\Sigma}_{32}^{-1}\operatorname{diag}(\mathsf{d}\boldsymbol{\sigma}_u^2) & \mathbf{O} \end{pmatrix}$$

and

$$\begin{aligned}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}(\mathsf{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{w} &= (\boldsymbol{\Sigma}^{-1}\mathbf{w})_2^T\,\mathrm{diag}(\mathsf{d}\boldsymbol{\sigma}_u^2)(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2\\ &= \{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2^{\odot 2}\}^T\mathsf{d}\boldsymbol{\sigma}_u^2,\end{aligned}$$

where $^{\odot 2}$ denotes element-wise squaring. Therefore

$$\begin{aligned}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathsf{d}\boldsymbol{\Sigma}) &= \mathrm{tr}\{\boldsymbol{\Sigma}_{22}^{-1}\,\mathrm{diag}(\mathsf{d}\boldsymbol{\sigma}_u^2)\}\\ &= \mathrm{diag}(\boldsymbol{\Sigma}_{22}^{-1})^T\mathsf{d}\boldsymbol{\sigma}_u^2,\end{aligned}$$

so

$$\mathsf{d}\log f(\mathbf{w}) = -\frac{1}{2}\,\mathrm{diag}(\boldsymbol{\Sigma}_{22}^{-1})^T\mathsf{d}\boldsymbol{\sigma}_u^2 + \frac{1}{2}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2^{\odot 2}\}^T\mathsf{d}\boldsymbol{\sigma}_u^2$$

and then

$$\nabla_{\boldsymbol{\sigma}_u^2}\log f(\mathbf{w}) = -\frac{1}{2}\,\mathrm{diag}(\boldsymbol{\Sigma}_{22}^{-1}) + \frac{1}{2}(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2^{\odot 2},$$

which implies

$$\boxed{\nabla_{\boldsymbol{\sigma}_u^2}\log f_{\boldsymbol{\theta}}(x) = -\frac{1}{2}\,\mathrm{diag}(\boldsymbol{\Sigma}_{22}^{-1}) + \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2^{\odot 2}\mid x\}.}$$

The second term can be written more explicitly in terms of $\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{w}\mathbf{w}^T\mid x)$: since $(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2 = [\mathbf{0},\mathbf{I}_{p_1},\mathbf{O}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}$ and $(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2^{\odot 2} = \mathrm{diag}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2^T\}$, we have

$$\begin{aligned}(\boldsymbol{\Sigma}^{-1}\mathbf{w})_2^{\odot 2} &= \mathrm{diag}\{[\mathbf{0},\mathbf{I}_{p_1},\mathbf{O}]\,\boldsymbol{\Sigma}^{-1}\mathbf{w}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\begin{bmatrix}\mathbf{0}^T\\ \mathbf{I}_{p_1}\\ \mathbf{O}\end{bmatrix}\}\\ &= \mathrm{diag}\{\left[\boldsymbol{\Sigma}_{21}^{-1},\boldsymbol{\Sigma}_{22}^{-1},\boldsymbol{\Sigma}_{23}^{-1}\right]\mathbf{w}\mathbf{w}^T\begin{bmatrix}\boldsymbol{\Sigma}_{12}^{-1}\\ \boldsymbol{\Sigma}_{22}^{-1}\\ \boldsymbol{\Sigma}_{32}^{-1}\end{bmatrix}\}.\end{aligned}$$

Then we take $\mathbb{E}_{\boldsymbol{\theta}}$, which commutes with the diag operator.

▶ For $\mathbf{d}_0$, given the symmetry of $f(x,\mathbf{w})$ on the parameters of λ_t and λ_s, we have

$$\boxed{\nabla_{\mathbf{d}_0}\log f_{\boldsymbol{\theta}}(x) = -\mathbb{E}_{\boldsymbol{\theta}}\{rI_t(\mathbf{u})\textstyle\int\lambda_s(\mathbf{s};\mathbf{v})\boldsymbol{\beta}_s(\mathbf{s})d\mathbf{s}\mid x\} + \sum_{j=1}^m\boldsymbol{\beta}_s(\mathbf{s}_j).}$$

▶ For $\mathrm{vec}\,\mathbf{D}$, again by the symmetry of $f(x,\mathbf{w})$ on the parameters, we have

$$\boxed{\nabla_{\mathrm{vec}\,\mathbf{D}}\log f_{\boldsymbol{\theta}}(x) = -\mathbb{E}_{\boldsymbol{\theta}}[rI_t(\mathbf{u})\textstyle\int\lambda_s(\mathbf{s};\mathbf{v})\{\mathbf{v}\otimes\boldsymbol{\beta}_s(\mathbf{s})\}d\mathbf{s}\mid x] + \mathbb{E}_{\boldsymbol{\theta}}(\mathbf{v}\mid x)\otimes\sum_{j=1}^m\boldsymbol{\beta}_s(\mathbf{s}_j).}$$

▶ For $\boldsymbol{\sigma}_v^2$, given the symmetry of the dependence of $\boldsymbol{\Sigma}$ on $\boldsymbol{\sigma}_u^2$ and $\boldsymbol{\sigma}_v^2$, we have

$$\boxed{\nabla_{\boldsymbol{\sigma}_v^2} \log f_{\boldsymbol{\theta}}(x) = -\frac{1}{2}\operatorname{diag}(\boldsymbol{\Sigma}_{33}^{-1}) + \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}}\{(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^{\odot 2} \mid x\}.}$$

Again, the second term can be written out in terms of $\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{w}\mathbf{w}^T \mid x)$ as

$$(\boldsymbol{\Sigma}^{-1}\mathbf{w})_3^{\odot 2} = \operatorname{diag}\{\left[\boldsymbol{\Sigma}_{31}^{-1}, \boldsymbol{\Sigma}_{32}^{-1}, \boldsymbol{\Sigma}_{33}^{-1}\right] \mathbf{w}\mathbf{w}^T \begin{bmatrix} \boldsymbol{\Sigma}_{13}^{-1} \\ \boldsymbol{\Sigma}_{23}^{-1} \\ \boldsymbol{\Sigma}_{33}^{-1} \end{bmatrix}\}.$$

Then we take $\mathbb{E}_{\boldsymbol{\theta}}$, which commutes with the diag operator.

2.2 Consistency

The consistency proof follows the usual steps for maximum likelihood estimators and M-estimators; see e.g. Pollard (1984) and Van der Vaart (2000). First we show that the asymptotic objective function has a unique maximum at the true parameter $\boldsymbol{\theta}_0$, then that $\{\hat{\boldsymbol{\theta}}_n\}$ is bounded in probability, and finally, via the Argmax Theorem, that $\hat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$ in probability. We define $\boldsymbol{\xi}_n = (\xi_{1n}, \xi_{2n}, \xi_{3n}, \xi_{4n})^T$ and $\mathbf{P}(\boldsymbol{\theta}) = (P(\mu), \sum_{k=1}^{p_1} P(\phi_k), P(\nu), \sum_{k=1}^{p_2} P(\psi_k))^T$.

For identifiability of the components we need to constrain the parameter space Θ further, so as to rule out sign ambiguity. Then we take

$$\begin{aligned}
\Theta \;=\; & \{\boldsymbol{\theta} \in \mathbb{R}^r : h_{kl}^C(\boldsymbol{\theta}) = 0, \quad k = 1, \ldots, l, \quad l = 1, \ldots, p_1; \\
& h_{kl}^D(\boldsymbol{\theta}) = 0, \quad k = 1, \ldots, l, \quad l = 1, \ldots, p_2; \\
& \mathbf{a}_{t0}^T \mathbf{c}_k = 0, \; k = 0, \ldots, p_1; \\
& \mathbf{a}_{s0}^T \mathbf{d}_k = 0, \; k = 0, \ldots, p_2; \\
& \mathbf{A}_P \mathbf{c}_k = \mathbf{0}, \; k = 0, \ldots, p_1; \\
& \boldsymbol{\Sigma} > 0; \; \sigma_{u1} > \cdots > \sigma_{up_1} > 0; \; \sigma_{v1} > \cdots > \sigma_{vp_2} > 0; \\
& c_{k1} \geq 0, \; k = 1, \ldots, p_1; \; d_{k1} \geq 0, \; k = 1, \ldots, p_2\}.
\end{aligned} \tag{9}$$

We make the following assumptions:

A1 The signs of the functional components $\hat{\phi}_{k,n}$ and $\hat{\psi}_{k,n}$ are specified so that the first non-zero basis coefficient of each $\hat{\phi}_{k,n}$ and $\hat{\psi}_{k,n}$ is positive (then $\hat{\boldsymbol{\theta}}_n \in \Theta$ for Θ defined in (9).)

A2 The true functional parameters μ_0, ν_0, ϕ_{k0}s and ψ_{k0}s of models (3) and (4) belong to the functional spaces $\mathscr{B}_t$ and $\mathscr{B}_s$ used for estimation, and the basis coefficients $c_{k1,0}$ and $d_{k1,0}$ are not zero. The signs of ϕ_{k0} and ψ_{k0} are then specified so that $c_{k1,0} > 0$

and $d_{k1,0} > 0$; therefore there is a unique $\boldsymbol{\theta}_0$ in Θ such that $f_{\boldsymbol{\theta}_0}(x)$ is the true density of the data.

A3 $\boldsymbol{\xi}_n \to \mathbf{0}$ as $n \to \infty$, where $\boldsymbol{\xi}_n = (\xi_{1n}, \xi_{2n}, \xi_{3n}, \xi_{4n})^T$ is the vector of smoothing parameters in (7).

A4 $\sqrt{n}\boldsymbol{\xi}_n \to \boldsymbol{\kappa}$ as $n \to \infty$, for a finite $\boldsymbol{\kappa}$.

Lemma 1 *Under assumption A2, the function* $M(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}\{\log f_{\boldsymbol{\theta}}(X)\}$ *has a unique maximum at* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Proof. This is a consequence of Jensen's Inequality and model identifiability:

$$E_{\boldsymbol{\theta}_0}\left\{\log \frac{f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}_0}(X)}\right\} \leq \log E_{\boldsymbol{\theta}_0}\left\{\frac{f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}_0}(X)}\right\} = 0 \tag{10}$$

because

$$E_{\boldsymbol{\theta}_0}\left\{\frac{f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}_0}(X)}\right\} = 1$$

for all $\boldsymbol{\theta}$. Moreover, inequality (10) is strict unless $P_{\boldsymbol{\theta}_0}\{f_{\boldsymbol{\theta}}(X)/f_{\boldsymbol{\theta}_0}(X) = 1\} = 1$, which happens only for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ by identifiability. Then $E_{\boldsymbol{\theta}_0}\{\log f_{\theta}(X)\} < E_{\boldsymbol{\theta}_0}\{\log f_{\boldsymbol{\theta}_0}(X)\}$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. ■

Lemma 2 *Under assumptions A1 and A3,* $\|\hat{\boldsymbol{\theta}}_n\| = O_P(1)$.

Proof. Let

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(X_i).$$

By definition, $\hat{\boldsymbol{\theta}}_n$ maximizes

$$\ell_n(\boldsymbol{\theta}) = M_n(\boldsymbol{\theta}) - \boldsymbol{\xi}_n^T \mathbf{P}(\boldsymbol{\theta}),$$

so we have

$$M_n(\hat{\boldsymbol{\theta}}_n) - M_n(\boldsymbol{\theta}_0) \geq \boldsymbol{\xi}_n^T\{\mathbf{P}(\hat{\boldsymbol{\theta}}_n) - \mathbf{P}(\boldsymbol{\theta}_0)\}.$$

Since $\mathbf{P}(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta}$, this implies

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{\hat{\boldsymbol{\theta}}_n}(X_i)}{f_{\boldsymbol{\theta}_0}(X_i)} \geq -\boldsymbol{\xi}_n^T \mathbf{P}(\boldsymbol{\theta}_0), \tag{11}$$

with the right-hand side going to zero as $n \to \infty$. As in Van der Vaart (2000, p. 63), consider the surrogate functions

$$g(x; \boldsymbol{\theta}) = \log\left\{\frac{f_{\boldsymbol{\theta}}(x) + f_{\boldsymbol{\theta}_0}(x)}{2 f_{\boldsymbol{\theta}_0}(x)}\right\}$$

which satisfy

$$\log(\frac{1}{2}) \leq g(x;\boldsymbol{\theta}) \leq \log\left\{\frac{c(x)+f_{\boldsymbol{\theta}_0}(x)}{2f_{\boldsymbol{\theta}_0}(x)}\right\}$$

where $c(x) \geq f_{\boldsymbol{\theta}}(x)$ for all $\boldsymbol{\theta}$. By concavity of the logarithm,

$$g(x;\boldsymbol{\theta}) \geq \frac{1}{2}\log\frac{f_{\boldsymbol{\theta}}(x)}{f_{\boldsymbol{\theta}_0}(x)} + \frac{1}{2}\log(1) = \frac{1}{2}\log\frac{f_{\boldsymbol{\theta}}(x)}{f_{\boldsymbol{\theta}_0}(x)},$$

so (11) implies

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i;\hat{\boldsymbol{\theta}}_n) \geq \frac{1}{2}\{-\boldsymbol{\xi}_n^T\mathbf{P}(\boldsymbol{\theta}_0)\}. \tag{12}$$

For any $K > 0$, if $\|\hat{\boldsymbol{\theta}}_n\| \geq K$ we have

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i;\hat{\boldsymbol{\theta}}_n) \leq \frac{1}{n}\sum_{i=1}^{n}\psi(X_i) \tag{13}$$

with

$$\psi(x) = \sup_{\|\boldsymbol{\theta}\|\geq K} g(x;\boldsymbol{\theta}).$$

By Law of Large Numbers $n^{-1}\sum_{i=1}^{n}\psi(X_i) \xrightarrow{P} E_{\boldsymbol{\theta}_0}\{\psi(X)\}$, and by Bounded Convergence Theorem we can switch supremum and expectation:

$$E_{\boldsymbol{\theta}_0}\{\psi(X)\} = \sup_{\|\boldsymbol{\theta}\|\geq K} E_{\boldsymbol{\theta}_0}\{g(X;\boldsymbol{\theta})\}.$$

Now, as in the proof of Lemma 1, by Jensen's Inequality we have

$$E_{\boldsymbol{\theta}_0}\{g(X;\boldsymbol{\theta})\} \leq \log E_{\boldsymbol{\theta}_0}\left\{\frac{f_{\boldsymbol{\theta}}(X)+f_{\boldsymbol{\theta}_0}(X)}{2f_{\boldsymbol{\theta}_0}(X)}\right\} = 0 = E_{\boldsymbol{\theta}_0}\{g(X;\boldsymbol{\theta}_0)\}$$

with strict inequality for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. So $\max E_{\boldsymbol{\theta}_0}\{g(X;\boldsymbol{\theta})\} = 0$ and the maximum is attained only at $\boldsymbol{\theta}_0$. We can rule out the possibility of $E_{\boldsymbol{\theta}_0}\{g(X;\boldsymbol{\theta})\}$ approaching zero at infinity because $\lim_{\|\boldsymbol{\theta}\|\to\infty} f_{\boldsymbol{\theta}}(x) = 0$ and then

$$\lim_{\|\boldsymbol{\theta}\|\to\infty} E_{\boldsymbol{\theta}_0}\{g(X;\boldsymbol{\theta})\} = E_{\boldsymbol{\theta}_0}\{\lim_{\|\boldsymbol{\theta}\|\to\infty} g(X;\boldsymbol{\theta})\} = \log(\frac{1}{2}) < 0.$$

Therefore, there exists an $\varepsilon > 0$ and a $K > 0$ such that $E_{\boldsymbol{\theta}_0}\{\psi(X)\} < -\varepsilon$. This fact together with (12) and (13) imply that $P(\|\hat{\boldsymbol{\theta}}_n\| \geq K)$ goes to zero as $n \to \infty$. ■

Lemma 3 *Under assumptions A1–A3,* $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$ *as* $n \to \infty$.

Proof. By Lemma 2, for any $\varepsilon > 0$ we can choose $K > 0$ such that $P\{\|\hat{\boldsymbol{\theta}}_n\| > K\} < \varepsilon/2$ for all n, and we can choose it so that $K \geq \|\boldsymbol{\theta}_0\|$. On the other hand, for $\|\hat{\boldsymbol{\theta}}_n\| \leq K$ we

have

$$\hat{\boldsymbol{\theta}}_n = \underset{\Theta \cap \{\|\boldsymbol{\theta}\| \leq K\}}{\operatorname{argmax}} \ell_n(\boldsymbol{\theta}).$$

The penalty function $\mathbf{P}(\boldsymbol{\theta})$ is continuous and therefore uniformly continuous on compact sets, and the process $M_n(\boldsymbol{\theta})$ is stochastically equicontinuous (Pollard, 1984, ch. 7), so $\ell_n(\boldsymbol{\theta})$ converges in probability to $M(\boldsymbol{\theta})$ uniformly over bounded sets. Then by the Argmax Theorem (Van der Vaart, 2000, ch. 5.9),

$$\underset{\Theta \cap \{\|\boldsymbol{\theta}\| \leq K\}}{\operatorname{argmax}} \ell_n(\boldsymbol{\theta}) \xrightarrow{P} \underset{\Theta \cap \{\|\boldsymbol{\theta}\| \leq K\}}{\operatorname{argmax}} M(\boldsymbol{\theta}) = \boldsymbol{\theta}_0,$$

so for any $\delta > 0$ we can choose N such that $P\{\|\hat{\boldsymbol{\theta}}_n\| \leq K, \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > \delta\} < \varepsilon/2$ for every $n \geq N$. This completes the proof. ■

2.3 Asymptotic normality

To prove the asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ we will follow the approach of Geyer (1994), which makes use of the tangent cone of the parameter space. The definition and properties of tangent cones can be found in Rockafellar and Wets (1998, ch. 6). Using Theorem 6.31 of Rockafellar and Wets (1998), the tangent cone of Θ at $\boldsymbol{\theta}_0$ is

$$\begin{aligned}
\mathscr{T}_0 \;=\; & \{\boldsymbol{\delta} \in \mathbb{R}^r : \nabla h_{kl}^C(\boldsymbol{\theta}_0)^T \boldsymbol{\delta} = 0,\ k = 1, \ldots, l,\ l = 1, \ldots, p_1; \\
& \nabla h_{kl}^D(\boldsymbol{\theta}_0)^T \boldsymbol{\delta} = 0,\ k = 1, \ldots, l,\ l = 1, \ldots, p_2; \\
& \mathbf{a}_{t0}^T \mathbf{K}_{\mathbf{c}_k} \boldsymbol{\delta} = 0,\ k = 0, \ldots, p_1; \\
& \mathbf{a}_{s0}^T \mathbf{K}_{\mathbf{d}_k} \boldsymbol{\delta} = 0,\ k = 0, \ldots, p_2; \\
& \mathbf{A}_P \mathbf{K}_{\mathbf{c}_k} \boldsymbol{\delta} = \mathbf{0},\ k = 0, \ldots, p_1\},
\end{aligned}$$

where $\mathbf{K}_{\mathbf{d}_k}$ and $\mathbf{K}_{\mathbf{c}_k}$ are the 'extraction' matrices such that $\mathbf{d}_k = \mathbf{K}_{\mathbf{d}_k}\boldsymbol{\theta}$ and $\mathbf{c}_k = \mathbf{K}_{\mathbf{c}_k}\boldsymbol{\theta}$. Note that $c_{k1,0}$ and $d_{k1,0}$ are strictly positive, so they do not contribute restrictions to the tangent cone. The explicit forms of $\nabla h_{kl}^C(\boldsymbol{\theta})$ and $\nabla h_{kl}^D(\boldsymbol{\theta})$ are derived in Section 2.4. Fisher's information matrix $\mathbf{F}_0$, which appears in the results below, was derived in Section 2.1.

Lemma 4 $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_P(n^{-1/2})$ *if* $\sqrt{n}\|\boldsymbol{\xi}_n\| = O_P(1)$.

Proof. The estimator $\hat{\boldsymbol{\theta}}_n$ maximizes $\ell_n(\boldsymbol{\theta})$, or equivalently

$$\tilde{\ell}_n(\boldsymbol{\theta}) = n\{\ell_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta}_0)\},$$

over $\boldsymbol{\theta} \in \Theta$. Let $r(x, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ be such that

$$\begin{aligned} \log f_{\boldsymbol{\theta}}(x) &= \log f_{\boldsymbol{\theta}_0}(x) + \nabla \log f_{\boldsymbol{\theta}_0}(x)^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \, r(x, \boldsymbol{\theta}, \boldsymbol{\theta}_0), \end{aligned}$$

and $M(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}\{\log f_{\boldsymbol{\theta}}(X)\}$ as above. Then

$$\begin{aligned} \tilde{M}_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \nabla \log f_{\boldsymbol{\theta}_0}(X_i)^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \sum_{i=1}^{n} [r(X_i, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - E_{\boldsymbol{\theta}_0}\{r(X, \boldsymbol{\theta}, \boldsymbol{\theta}_0)\}] \\ &\quad + n\{M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}_0)\} - n\boldsymbol{\xi}_n^T\{\mathbf{P}(\boldsymbol{\theta}) - \mathbf{P}(\boldsymbol{\theta}_0)\}. \end{aligned}$$

Note that $E_{\theta_0}\{\nabla \log f_{\boldsymbol{\theta}_0}(X)\} = \nabla M(\boldsymbol{\theta}_0) = \mathbf{0}$ because $f_{\boldsymbol{\theta}}(x)$ is a density function; the fact that $\boldsymbol{\theta}_0$ maximizes $M(\boldsymbol{\theta})$ does not necessarily imply $\nabla M(\boldsymbol{\theta}_0) = \mathbf{0}$ because $\boldsymbol{\theta}_0$ may be on the border of Θ. Let

$$R_n(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [r(X_i, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - E_{\boldsymbol{\theta}_0}\{r(X, \boldsymbol{\theta}, \boldsymbol{\theta}_0)\}]$$

and

$$\mathbf{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \log f(Z_i, \boldsymbol{\theta}_0).$$

Since $\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) \geq \tilde{\ell}_n(\boldsymbol{\theta}_0) = 0$,

$$\begin{aligned} & \sqrt{n}\mathbf{Z}_n^T(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| R_n(\hat{\boldsymbol{\theta}}_n) - n\boldsymbol{\xi}_n^T\{\mathbf{P}(\hat{\boldsymbol{\theta}}_n) - \mathbf{P}(\boldsymbol{\theta}_0)\} \\ \geq \; & -n\{M(\hat{\boldsymbol{\theta}}_n) - M(\boldsymbol{\theta}_0)\}. \end{aligned} \tag{14}$$

Clearly $\|\mathbf{Z}_n\| = O_P(1)$ because $\mathbf{Z}_n \xrightarrow{D} \mathrm{N}(0, \mathbf{F}_0)$. The mean value theorem applied to $\mathbf{P}(\boldsymbol{\theta})$ implies

$$\begin{aligned} n\boldsymbol{\xi}_n^T\{\mathbf{P}(\hat{\boldsymbol{\theta}}_n) - \mathbf{P}(\boldsymbol{\theta}_0)\} &= n\|\boldsymbol{\xi}_n\| O_P(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|) \\ &= \sqrt{n}\|\boldsymbol{\xi}_n\| O_P(1) \sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|. \end{aligned}$$

The process $R_n(\boldsymbol{\theta})$ is equicontinuous in $\boldsymbol{\theta}$ (Pollard, 1984, ch. 7) and $R_n(\boldsymbol{\theta}) \xrightarrow{D} \mathrm{N}(0, v(\boldsymbol{\theta}, \boldsymbol{\theta}_0))$ with $v(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = 0$, so $R_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{P} 0$. Then it follows from (14) that

$$\begin{aligned} & \{O_P(1) + o_P(1) - \sqrt{n}\|\boldsymbol{\xi}_n\| O_P(1)\} \sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \\ \geq \; & -n\{M(\hat{\boldsymbol{\theta}}_n) - M(\boldsymbol{\theta}_0)\}. \end{aligned}$$

Now,

$$M(\hat{\boldsymbol{\theta}}_n) - M(\boldsymbol{\theta}_0) = \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T \nabla^2 M(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_P(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2)$$

and $\nabla^2 M(\boldsymbol{\theta}_0) = -\mathbf{F}_0$, so if $\lambda_1 > 0$ is the smallest eigenvalue of $\mathbf{F}_0$,

$$-n\{M(\hat{\boldsymbol{\theta}}_n) - M(\boldsymbol{\theta}_0)\} \geq n\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2\lambda_1 - no_P(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2).$$

Then from the last two inequalities we have

$$\{O_P(1) + o_P(1) - \sqrt{n}\|\boldsymbol{\xi}_n\|O_P(1)\}\sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \geq n\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2\{\lambda_1 - o_P(1)\},$$

which implies $\sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_P(1)$. ■

Theorem 5 *Under assumption A4,* $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} \boldsymbol{\delta}(\mathbf{Z})$*, with* $\boldsymbol{\delta}(\mathbf{Z})$ *the maximizer of*

$$W(\boldsymbol{\delta}) = \{\mathbf{Z}^T - \boldsymbol{\kappa}^T \mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0)\}\boldsymbol{\delta} - \frac{1}{2}\boldsymbol{\delta}^T\mathbf{F}_0\boldsymbol{\delta}$$

over $\boldsymbol{\delta} \in \mathscr{T}_0$*, where* $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{F}_0)$.

Proof. Let $W_n(\boldsymbol{\delta}) = \tilde{\ell}_n(\boldsymbol{\theta}_0 + \boldsymbol{\delta}/\sqrt{n})$ with $\tilde{\ell}_n(\boldsymbol{\theta})$ as in the previous lemma. Then

$$\begin{aligned} W_n(\boldsymbol{\delta}) &= \mathbf{Z}_n^T\boldsymbol{\delta} \\ &\quad + \|\boldsymbol{\delta}\|\, R_n(\boldsymbol{\theta}_0 + \boldsymbol{\delta}/\sqrt{n}) \\ &\quad + n\{M(\boldsymbol{\theta}_0 + \boldsymbol{\delta}/\sqrt{n}) - M(\boldsymbol{\theta}_0)\} \\ &\quad - n\boldsymbol{\xi}_n^T\{\mathbf{P}(\boldsymbol{\theta}_0 + \boldsymbol{\delta}/\sqrt{n}) - \mathbf{P}(\boldsymbol{\theta}_0)\}, \end{aligned}$$

and $\hat{\boldsymbol{\delta}}_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ maximizes $W_n(\boldsymbol{\delta})$ over $\mathscr{T}_n = \sqrt{n}(\Theta - \{\boldsymbol{\theta}_0\})$. Having already proved that $\|\hat{\boldsymbol{\delta}}_n\| = O_P(1)$, given $\varepsilon > 0$ we can take K such that $P(\|\hat{\boldsymbol{\delta}}_n\| \leq K) \geq 1 - \varepsilon$ for every n, and focus on the set $\mathscr{T}_n \cap \{\|\boldsymbol{\delta}\| \leq K\}$. In the limit, as $n \to \infty$, we have:

$$\mathscr{T}_n \to \mathscr{T}_0, \text{ the tangent cone of } \Theta \text{ at } \boldsymbol{\theta}_0$$

(Geyer, 1994);

$$\mathbf{Z}_n \xrightarrow{D} \mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{F}_0);$$

$$R_n(\boldsymbol{\theta}_0 + \boldsymbol{\delta}_n/\sqrt{n}) \xrightarrow{P} 0 \text{ for any bounded sequence } \{\boldsymbol{\delta}_n\}$$

by stochastic equicontinuity of $R_n(\boldsymbol{\theta})$;

$$n\{M(\boldsymbol{\theta}_0 + \boldsymbol{\delta}/\sqrt{n}) - M(\boldsymbol{\theta}_0)\} = \frac{1}{2}\boldsymbol{\delta}^T\{-\mathbf{F}_0 + o_P(1)\}\boldsymbol{\delta};$$

and

$$n\boldsymbol{\xi}_n^T\{\mathbf{P}(\boldsymbol{\theta}_0 + \boldsymbol{\delta}/\sqrt{n}) - \mathbf{P}(\boldsymbol{\theta}_0)\} = \sqrt{n}\boldsymbol{\xi}_n^T\{\mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0) + o_P(1)\}\boldsymbol{\delta}.$$

All this implies that $W_n(\boldsymbol{\delta}) \xrightarrow{D} W(\boldsymbol{\delta})$ with

$$W(\boldsymbol{\delta}) = \{\mathbf{Z}^T - \boldsymbol{\kappa}^T \mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0)\}\boldsymbol{\delta} - \frac{1}{2}\boldsymbol{\delta}^T \mathbf{F}_0 \boldsymbol{\delta},$$

and the convergence is uniform in $\boldsymbol{\delta}$, i.e. $\sup_{\mathscr{T}_n \cap \{\|\boldsymbol{\delta}\| \le K\}} |W_n(\boldsymbol{\delta}) - W(\boldsymbol{\delta})| \xrightarrow{P} 0$. Then

$$\operatorname*{argmax}_{\mathscr{T}_n \cap \{\|\boldsymbol{\delta}\| \le K\}} W_n(\boldsymbol{\delta}) \xrightarrow{D} \operatorname*{argmax}_{\mathscr{T}_0} W(\boldsymbol{\delta}),$$

which implies that $\hat{\boldsymbol{\delta}}_n \xrightarrow{D} \boldsymbol{\delta}(\mathbf{Z})$ as stated. ■

We have $\mathscr{T}_0 = \{\boldsymbol{\delta} \in \mathbb{R}^r : \mathbf{A}\boldsymbol{\delta} = \mathbf{0}\}$, with $\mathbf{A}$ the $r_1 \times r$ matrix with rows $\nabla h^C_{kl}(\boldsymbol{\theta}_0)^T$, $\nabla h^D_{kl}(\boldsymbol{\theta}_0)^T$, $\mathbf{a}^T_{t0}\mathbf{K}_{\mathbf{c}_k}$, $\mathbf{a}^T_{s0}\mathbf{K}_{\mathbf{d}_k}$ and $\mathbf{A}_P\mathbf{K}_{\mathbf{c}_k}$. Then a $\boldsymbol{\delta} \in \mathscr{T}_0$ is of the form $\boldsymbol{\delta} = \mathbf{B}^T\tilde{\boldsymbol{\delta}}$ with $\mathbf{B}$ an orthogonal $(r - r_1) \times r$ matrix with rows orthogonal to those of $\mathbf{A}$ and $\tilde{\boldsymbol{\delta}} \in \mathbb{R}^{r-r_1}$ free. So we can reparameterize the process $W(\boldsymbol{\delta})$ above in terms of $\tilde{\boldsymbol{\delta}}$:

$$W(\boldsymbol{\delta}) = W(\mathbf{B}^T\tilde{\boldsymbol{\delta}}) = \{\mathbf{Z}^T - \boldsymbol{\kappa}^T \mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0)\}\mathbf{B}^T\tilde{\boldsymbol{\delta}} - \frac{1}{2}\tilde{\boldsymbol{\delta}}^T\mathbf{B}\mathbf{F}_0\mathbf{B}^T\tilde{\boldsymbol{\delta}},$$

which is maximized by $\tilde{\boldsymbol{\delta}}(\mathbf{Z}) = (\mathbf{B}\mathbf{F}_0\mathbf{B}^T)^{-1}\mathbf{B}\{\mathbf{Z} - \mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0)^T\boldsymbol{\kappa}\}$, and then $\boldsymbol{\delta}(\mathbf{Z}) = \mathbf{B}^T\tilde{\boldsymbol{\delta}}(\mathbf{Z})$. Since $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \mathbf{F}_0)$, we have $\tilde{\boldsymbol{\delta}}(\mathbf{Z}) \sim \mathrm{N}(-(\mathbf{B}\mathbf{F}_0\mathbf{B}^T)^{-1}\mathbf{B}\mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0)^T\boldsymbol{\kappa}, (\mathbf{B}\mathbf{F}_0\mathbf{B}^T)^{-1})$ and then

$$\boldsymbol{\delta}(\mathbf{Z}) \sim \mathrm{N}(-\mathbf{V}\mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0)^T\boldsymbol{\kappa}, \mathbf{V})$$

with $\mathbf{V} = \mathbf{B}^T(\mathbf{B}\mathbf{F}_0\mathbf{B}^T)^{-1}\mathbf{B}$. The explicit form of $\mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0)$ is derived in Section 2.4.

2.4 Derivatives of constraints and smoothness penalties

The explicit forms of $\nabla h^C_{kl}(\boldsymbol{\theta})$ and $\nabla h^D_{kl}(\boldsymbol{\theta})$ can be derived as follows. Let $\mathbf{K}_{\mathbf{c}_k}$ be the $q_1 \times r$ matrix that 'extracts' $\mathbf{c}_k$ from $\boldsymbol{\theta}$, that is, $\mathbf{c}_k = \mathbf{K}_{\mathbf{c}_k}\boldsymbol{\theta}$. Then we can write $h^C_{kl}(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\mathbf{K}^T_{\mathbf{c}_k}\mathbf{J}_t\mathbf{K}_{\mathbf{c}_l}\boldsymbol{\theta} - \delta_{kl}$ and it follows that

$$\nabla h^C_{kl}(\boldsymbol{\theta}) = (\mathbf{K}^T_{\mathbf{c}_k}\mathbf{J}_t\mathbf{K}_{\mathbf{c}_l} + \mathbf{K}^T_{\mathbf{c}_l}\mathbf{J}_t\mathbf{K}_{\mathbf{c}_k})\boldsymbol{\theta}.$$

Similarly, if $\mathbf{K}_{\mathbf{d}_k}$ is the $q_2 \times r$ matrix such that $\mathbf{d}_k = \mathbf{K}_{\mathbf{d}_k}\boldsymbol{\theta}$, we have $h^D_{kl}(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\mathbf{K}^T_{\mathbf{d}_k}\mathbf{J}_s\mathbf{K}_{\mathbf{d}_k}\boldsymbol{\theta} - \delta_{kl}$ and then

$$\nabla h^D_{kl}(\boldsymbol{\theta}) = (\mathbf{K}^T_{\mathbf{d}_k}\mathbf{J}_s\mathbf{K}_{\mathbf{d}_l} + \mathbf{K}^T_{\mathbf{d}_l}\mathbf{J}_s\mathbf{K}_{\mathbf{d}_k})\boldsymbol{\theta}.$$

The explicit form of $\mathsf{D}\mathbf{P}(\boldsymbol{\theta})$ is derived in a similar way. Using extraction matrices $\mathbf{K}$ as above and the smoothing matrices $\boldsymbol{\Omega}_t$ and $\boldsymbol{\Omega}_s$ specified in Section 1.2, we have $\mathbf{P}(\boldsymbol{\theta}) = (P(\mu), \sum_{k=1}^{p_1} P(\phi_k), P(\nu), \sum_{k=1}^{p_2} P(\psi_k))^T$ with

$$\begin{aligned} P(\mu) &= \mathbf{c}_0^T\boldsymbol{\Omega}_t\mathbf{c}_0 \\ &= \boldsymbol{\theta}^T\mathbf{K}^T_{\mathbf{c}_0}\boldsymbol{\Omega}_t\mathbf{K}_{\mathbf{c}_0}\boldsymbol{\theta}, \end{aligned}$$

$$\begin{aligned}\sum_{k=1}^{p_1} P(\phi_k) &= \operatorname{tr}(\mathbf{C}^T\boldsymbol{\Omega}_t\mathbf{C}) \\ &= \operatorname{vec}\mathbf{C}^T(\mathbf{I}_{p_1}\otimes\boldsymbol{\Omega}_t)\operatorname{vec}\mathbf{C} \\ &= \boldsymbol{\theta}^T\mathbf{K}_{\operatorname{vec}\mathbf{C}}^T(\mathbf{I}_{p_1}\otimes\boldsymbol{\Omega}_t)\mathbf{K}_{\operatorname{vec}\mathbf{C}}\boldsymbol{\theta},\end{aligned}$$

$$\begin{aligned}P(\nu) &= \mathbf{d}_0^T\boldsymbol{\Omega}_s\mathbf{d}_0 \\ &= \boldsymbol{\theta}^T\mathbf{K}_{\mathbf{d}_0}^T\boldsymbol{\Omega}_s\mathbf{K}_{\mathbf{d}_0}\boldsymbol{\theta},\end{aligned}$$

and

$$\begin{aligned}\sum_{k=1}^{p_2} P(\psi_k) &= \operatorname{tr}(\mathbf{D}^T\boldsymbol{\Omega}_s\mathbf{D}) \\ &= \boldsymbol{\theta}^T\mathbf{K}_{\operatorname{vec}\mathbf{D}}^T(\mathbf{I}_{p_2}\otimes\boldsymbol{\Omega}_s)\mathbf{K}_{\operatorname{vec}\mathbf{D}}\boldsymbol{\theta}.\end{aligned}$$

Then

$$\mathsf{D}\mathbf{P}(\boldsymbol{\theta}) = \begin{bmatrix} 2\boldsymbol{\theta}^T\mathbf{K}_{\mathbf{c}_0}^T\boldsymbol{\Omega}_t\mathbf{K}_{\mathbf{c}_0} \\ 2\boldsymbol{\theta}^T\mathbf{K}_{\operatorname{vec}\mathbf{C}}^T(\mathbf{I}_{p_1}\otimes\boldsymbol{\Omega}_t)\mathbf{K}_{\operatorname{vec}\mathbf{C}} \\ 2\boldsymbol{\theta}^T\mathbf{K}_{\mathbf{d}_0}^T\boldsymbol{\Omega}_s\mathbf{K}_{\mathbf{d}_0} \\ 2\boldsymbol{\theta}^T\mathbf{K}_{\operatorname{vec}\mathbf{D}}^T(\mathbf{I}_{p_2}\otimes\boldsymbol{\Omega}_s)\mathbf{K}_{\operatorname{vec}\mathbf{D}} \end{bmatrix}.$$

3 Simulations

We provide here some additional plots with simulation outputs, specifically for the temporal mean and components μ, ϕ_1 and ϕ_2; the spatial mean and components, unfortunately, cannot be visualized this way. Figures 1 and 2 show plots of the true μ and the simulated $\hat{\mu}$s for different sample sizes, for $\tau = \log 10$ and $\tau = \log 30$ respectively. Figures 3 and 4 do the same for $\hat{\phi}_1$, and Figures 5 and 6 for $\hat{\phi}_2$.

4 Application: Chicago's Divvy bike sharing system

In this section we include additional plots for the Divvy data analysis presented in the paper.

5 References

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39** 1–38.

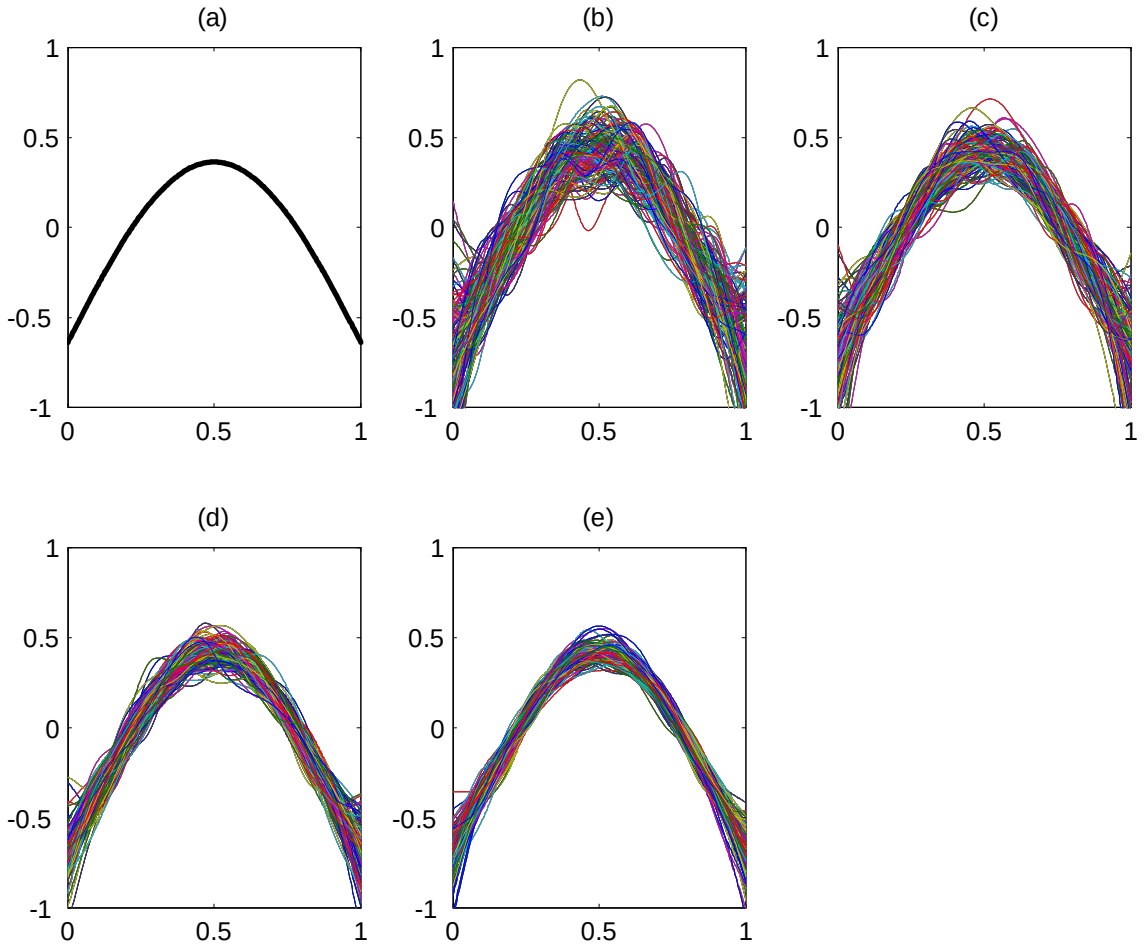

Figure 1: Simulation results. (a) True temporal mean μ and (b)–(e) simulated estimators $\hat{\mu}$ for sample sizes (b) $n = 50$, (c) $n = 100$, (d) $n = 200$ and (e) $n = 400$. Rate parameter $\tau = \log 10$.

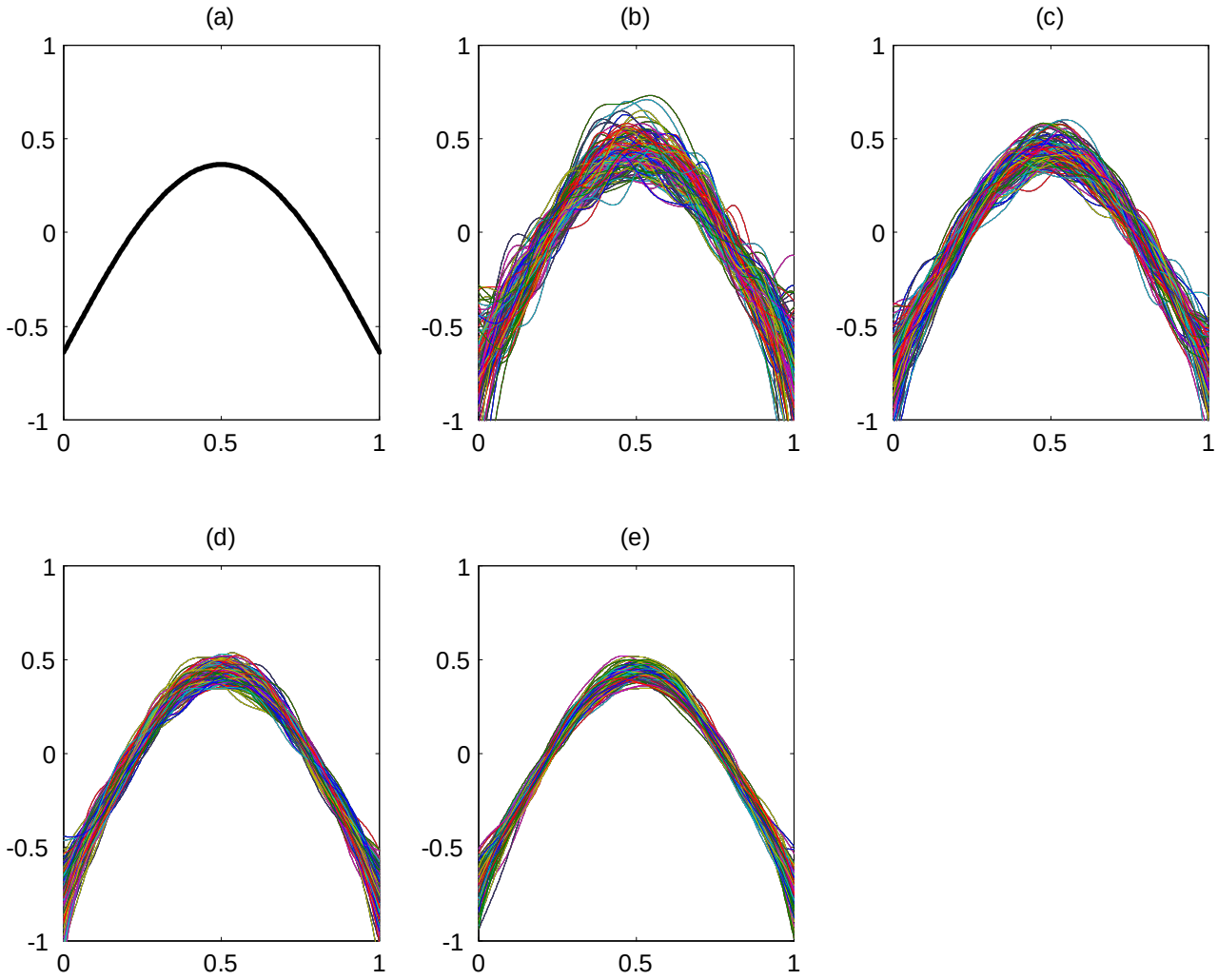

Figure 2: Simulation results. (a) True temporal mean μ and (b)–(e) simulated estimators $\hat{\mu}$ for sample sizes (b) $n = 50$, (c) $n = 100$, (d) $n = 200$ and (e) $n = 400$. Rate parameter $\tau = \log 30$.

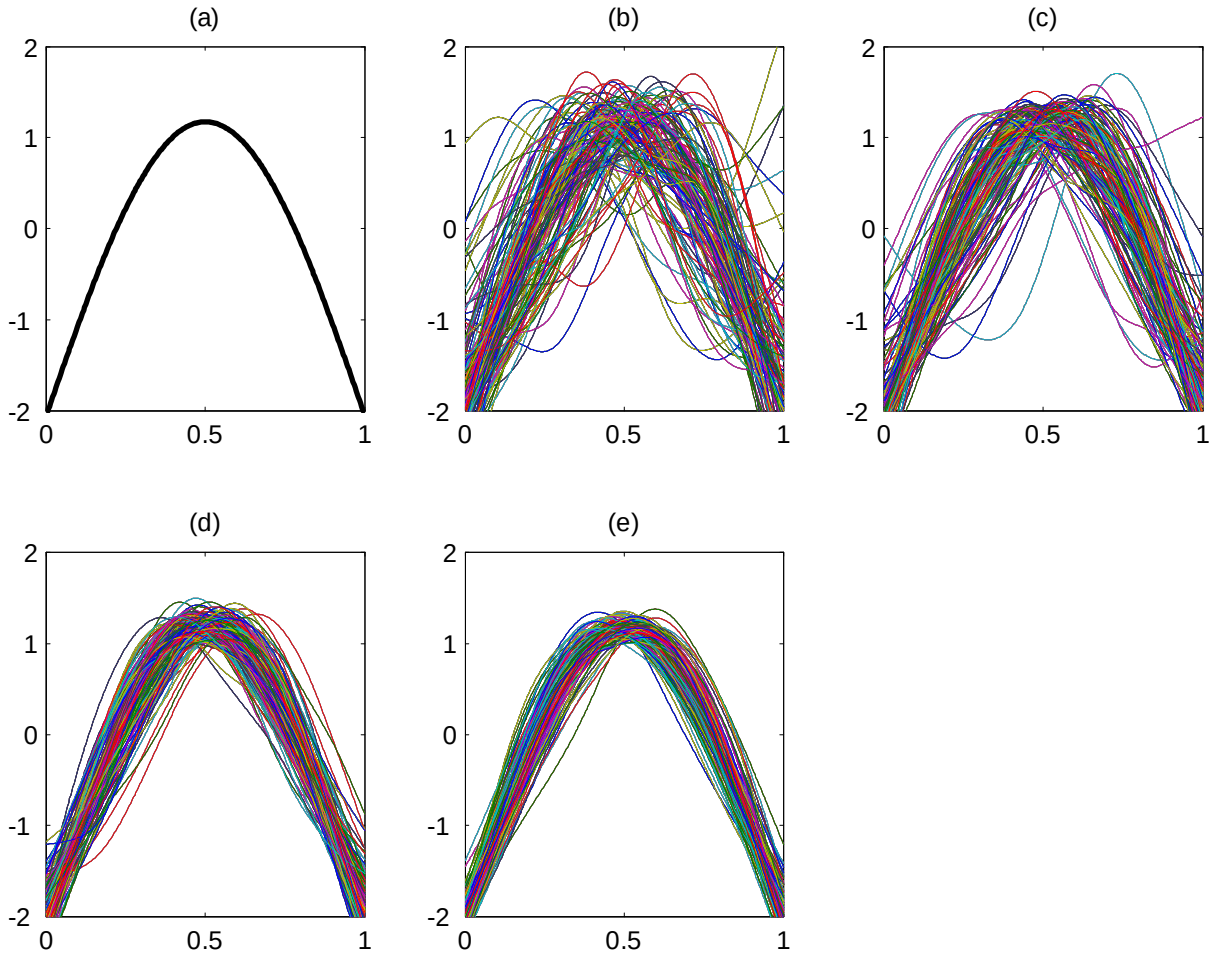

Figure 3: Simulation results. (a) True temporal component ϕ_1 and (b)–(e) simulated estimators $\hat{\phi}_1$ for sample sizes (b) $n = 50$, (c) $n = 100$, (d) $n = 200$ and (e) $n = 400$. Rate parameter $\tau = \log 10$.

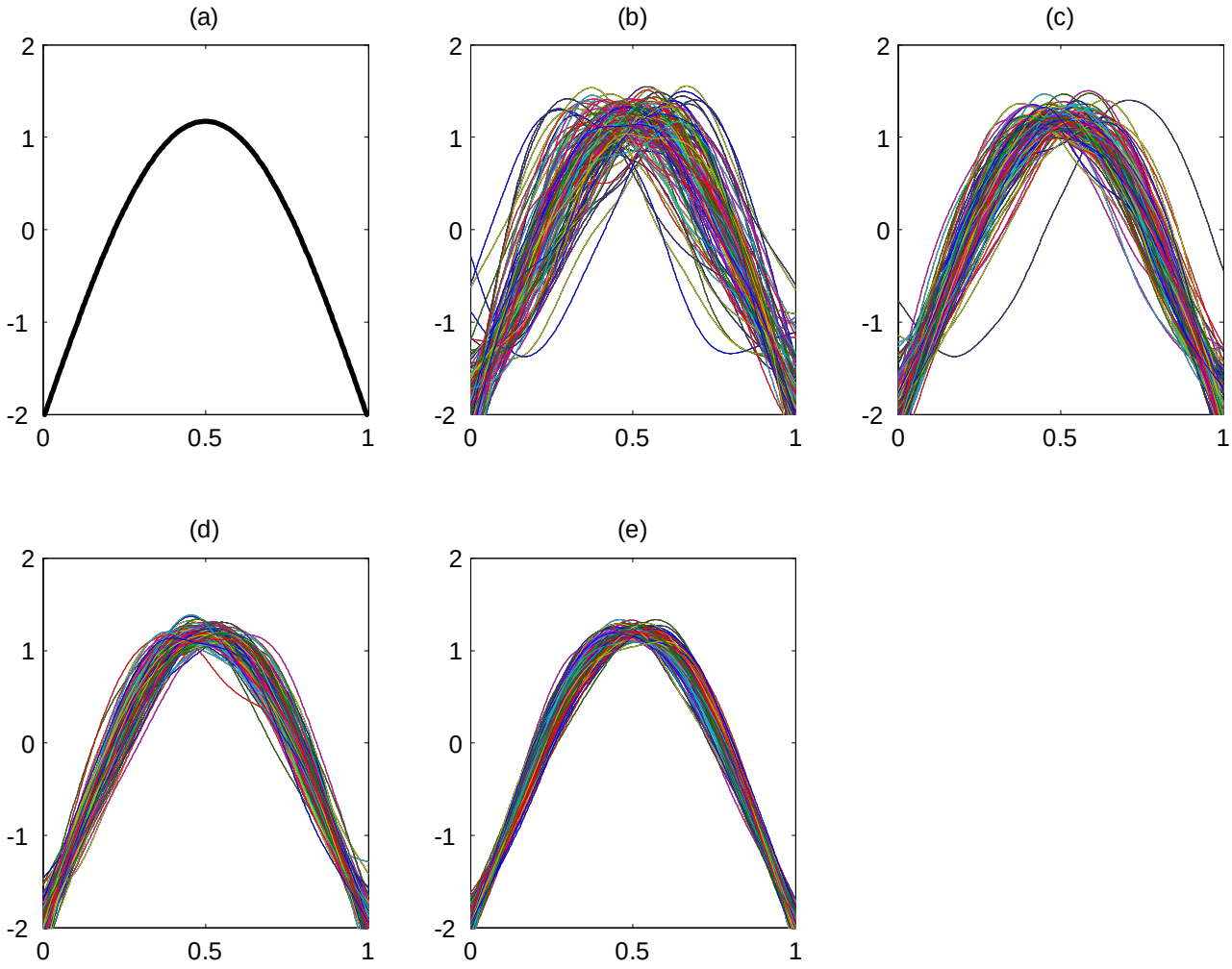

Figure 4: Simulation results. (a) True temporal component ϕ_1 and (b)–(e) simulated estimators $\hat{\phi}_1$ for sample sizes (b) $n = 50$, (c) $n = 100$, (d) $n = 200$ and (e) $n = 400$. Rate parameter $\tau = \log 30$.

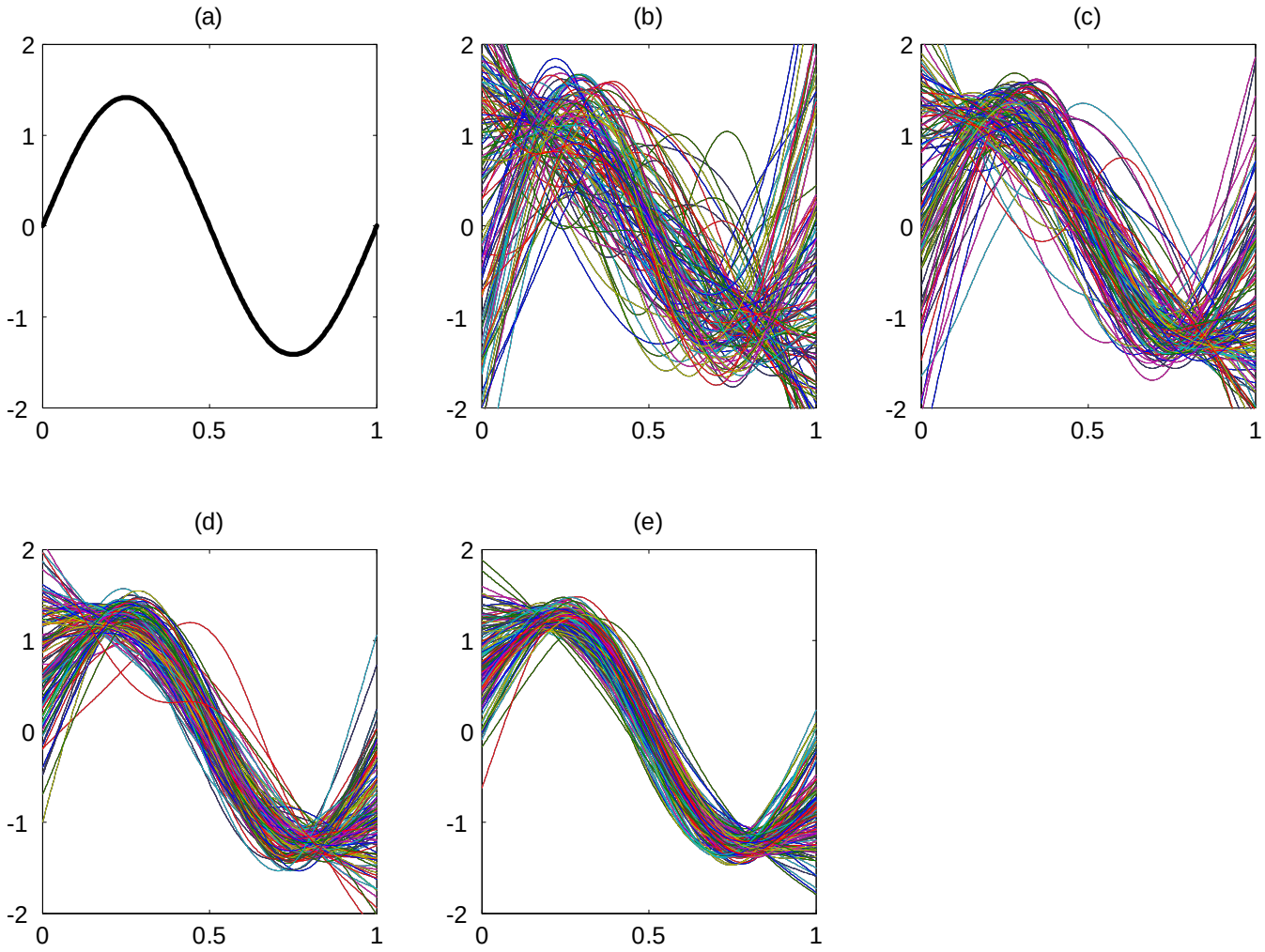

Figure 5: Simulation results. (a) True temporal component ϕ_2 and (b)–(e) simulated estimators $\hat{\phi}_2$ for sample sizes (b) $n = 50$, (c) $n = 100$, (d) $n = 200$ and (e) $n = 400$. Rate parameter $\tau = \log 10$.

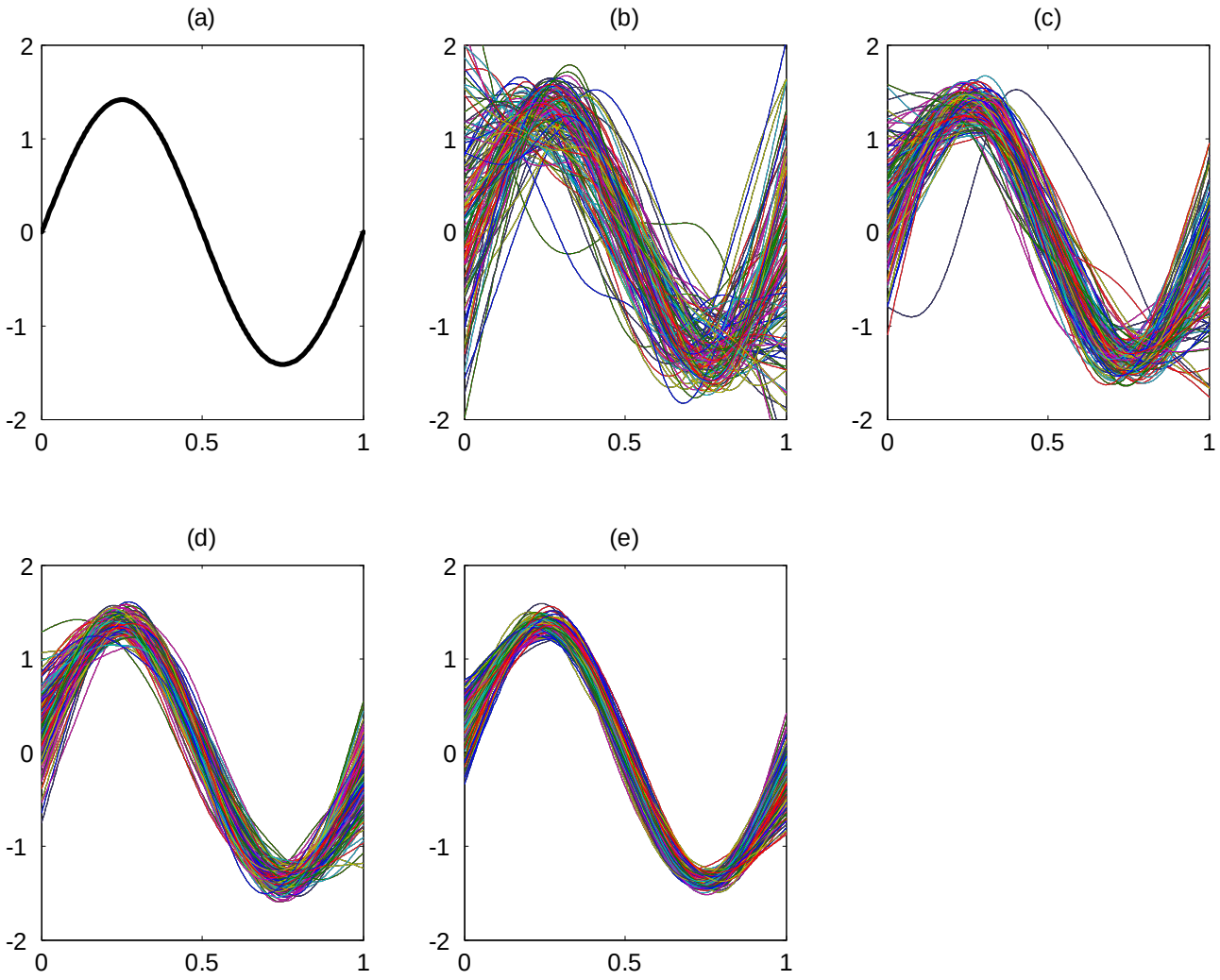

Figure 6: Simulation results. (a) True temporal component ϕ_2 and (b)–(e) simulated estimators $\hat{\phi}_2$ for sample sizes (b) $n = 50$, (c) $n = 100$, (d) $n = 200$ and (e) $n = 400$. Rate parameter $\tau = \log 30$.

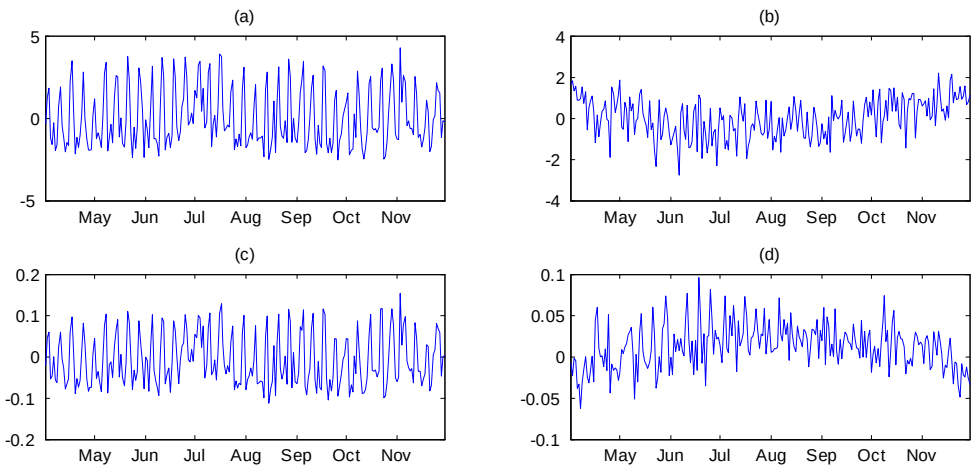

Figure 7: Divvy Data Analysis. Component scores for (a) first temporal component, (b) second temporal component, (c) first spatial component, and (d) second spatial component.

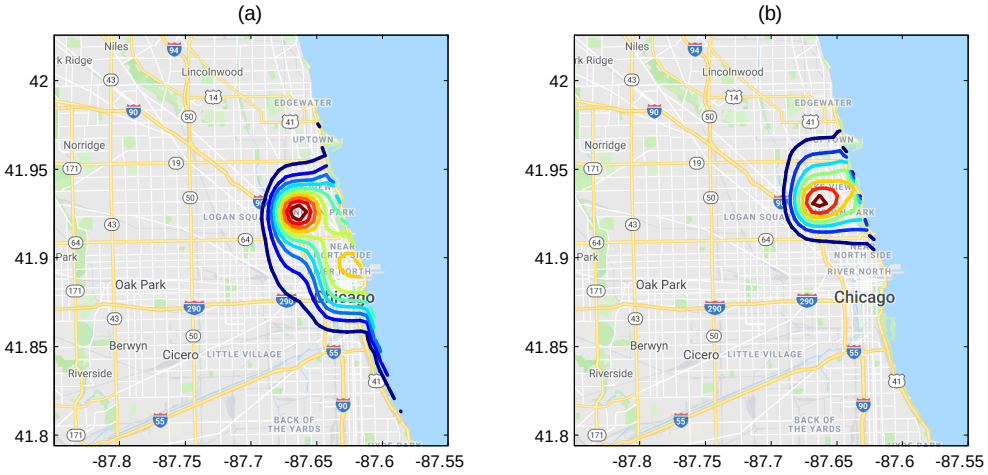

Figure 8: Divvy Data Analysis. Effect of the first spatial component on the baseline intensity. Contour plots show baseline intensity minus [(a)] and plus [(b)] a multiple of the component.

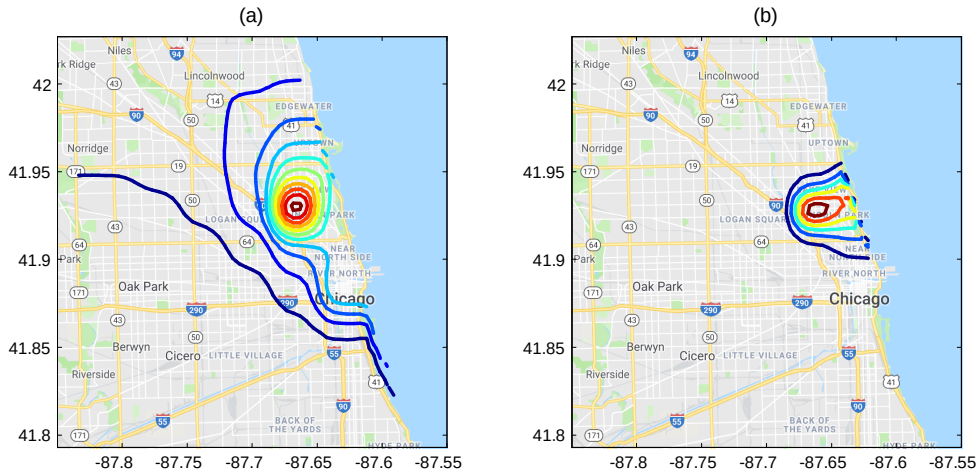

Figure 9: Divvy Data Analysis. Effect of the second spatial component on the baseline intensity. Contour plots show baseline intensity minus [(a)] and plus [(b)] a multiple of the component.

Geyer, C.J. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics* **22** 1993–2010.

Magnus, J.R., and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, New York.

Rockafellar, R.T., and Wets, R.J. (1998). *Variational Analysis.* Springer, New York.

Van der Vaart, A. (2000). *Asymptotic Statistics.* Cambridge University Press, Cambridge, UK.