Supplementary material for Independent Component Models for Replicated Point Processes

Daniel Gervini
Department of Mathematical Sciences
University of Wisconsin–Milwaukee

March 31, 2016

1 Estimation

1.1 EM algorithm

We derive the EM algorithm for U_ks with Gamma(α_k, β) distributions, so $S = \sum_{k=1}^p U_k$ is Gamma(α, β) with $\alpha = \sum_{k=1}^p \alpha_k$ and $\mathbf{W} = \mathbf{U}/S$ is Dirichlet$(\alpha_1, \ldots, \alpha_p)$, and S and $\mathbf{W}$ are independent. With a slight abuse of notation we are going to identify $x_B = \{t_1, \ldots, t_m\}$ with $(m, \mathbf{t})$ from now on. The joint density of $(m, \mathbf{t})$ and the latent variables $(\mathbf{y}, s, \mathbf{w})$ is

$$\begin{aligned} f(m, \mathbf{t}, \mathbf{y}, s, \mathbf{w}) &= f(\mathbf{t} \mid m, \mathbf{y}) f(m \mid s) f(\mathbf{y} \mid \mathbf{w}) f(s) f(\mathbf{w}) \\ &= \{\prod_{j=1}^m \prod_{k=1}^p \phi_k(t_j)^{y_{jk}}\} \times e^{-s} \frac{s^m}{m!} \times \{\prod_{j=1}^m \prod_{k=1}^p w_k^{y_{jk}}\} \\ &\quad \times \frac{s^{\alpha-1} e^{-s/\beta}}{\Gamma(\alpha)\beta^\alpha} \times \{\frac{\Gamma(\alpha)}{\prod_{k=1}^p \Gamma(\alpha_k)} \prod_{k=1}^p w_k^{\alpha_k - 1}\}. \end{aligned}$$

Since

$$f(m, \mathbf{t}, \mathbf{y}, s, \mathbf{w}) = s^{m+\alpha-1} e^{-s(1+1/\beta)} \times g_1(m, \mathbf{t}, \mathbf{y}, \mathbf{w})$$

it follows that $S \mid (m, \mathbf{t})$ has a Gamma$(m + \alpha, \beta/(\beta + 1))$ distribution. We can also write, using the fact that $\sum_{k=1}^p y_{jk} = 1$ for all j,

$$\begin{aligned} f(m, \mathbf{t}, \mathbf{y}, s, \mathbf{w}) &= \prod_{j=1}^m \prod_{k=1}^p \{w_k \phi_k(t_j)\}^{y_{jk}} \times g_2(m, s, \mathbf{w}) \\ &= \prod_{j=1}^m \prod_{k=1}^p \left\{ \frac{w_k \phi_k(t_j)}{\sum_{l=1}^p w_l \phi_l(t_j)} \right\}^{y_{jk}} \times \prod_{j=1}^m \{\sum_{l=1}^p w_l \phi_l(t_j)\} \times g_2(m, s, \mathbf{w}) \\ &= \prod_{j=1}^m \prod_{k=1}^p \xi_{jk}^{y_{jk}} \times g_3(m, \mathbf{t}, s, \mathbf{w}) \end{aligned}$$

with $\xi_{jk} = w_k\phi_k(t_j)/\sum_{l=1}^p w_l\phi_l(t_j)$. Then the $\mathbf{Y}_j$s are independent conditionally on $(m, \mathbf{t}, \mathbf{w})$, and each $\mathbf{Y}_j \mid (m, \mathbf{t}, \mathbf{w})$ is Multinomial$(1, \boldsymbol{\xi}_j)$. Finally, if $y_{\cdot k} = \sum_{j=1}^m y_{jk}$, we have

$$f(m, \mathbf{t}, \mathbf{y}, s, \mathbf{w}) = \prod_{k=1}^p w_k^{y_{\cdot k}+\alpha_k-1} \times g_4(m, \mathbf{t}, \mathbf{y}, s)$$

so $\mathbf{W} \mid (m, \mathbf{t}, \mathbf{y})$ has a Dirichlet$(y_{\cdot 1} + \alpha_1, \ldots, y_{\cdot p} + \alpha_p)$ distribution.

Treating the random effects as "missing data", the "complete loglikelihood" is

$$\begin{aligned}
\tilde{\ell}(\boldsymbol{\theta}) &= \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{k=1}^p y_{ijk} \log \phi_k(t_{ij}) + \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{k=1}^p y_{ijk} \log w_{ik} \\
&\quad + \sum_{i=1}^n (-s_i + m_i \log s_i - \log m_i!) \\
&\quad + \sum_{i=1}^n \{(\alpha - 1) \log s_i - \frac{s_i}{\beta} - \log \Gamma(\alpha) - \alpha \log \beta\} \\
&\quad + \sum_{i=1}^n \{\log \Gamma(\alpha) - \sum_{k=1}^p \log \Gamma(\alpha_k) + \sum_{k=1}^p (\alpha_k - 1) w_{ik}\}.
\end{aligned}$$

The objective function maximized at the h-th step of the EM algorithm is

$$Q_{\hat{\boldsymbol{\theta}}^{(h-1)}}(\boldsymbol{\theta}) = n^{-1} \mathbb{E}_{\hat{\boldsymbol{\theta}}^{(h-1)}} \{\tilde{\ell}(\boldsymbol{\theta}) \mid x_{B1}, \ldots, x_{Bn}\} - \zeta P(\boldsymbol{\theta}),$$

subject to the constraints

$$\int_B \phi_k(t) dt = \mathbf{c}_k^T \mathbf{a} = 1, \quad k = 1, \ldots, p,$$

where $\mathbf{a} = \int_B \boldsymbol{\beta}(t) dt$. These can be written in matrix form as $\mathbf{A}\,\mathrm{vec}(\mathbf{C}) = \mathbf{1}_p$, with $\mathbf{A} = \mathbf{I}_p \otimes \mathbf{a}^T$. In addition we have the nonnegativity constraints $\mathbf{c}_k \geq 0$ and the positivity constraints $\alpha_k > 0$ and $\beta > 0$.

To simplify notation, from now on we denote $\mathbb{E}_{\hat{\boldsymbol{\theta}}^{(h-1)}}(s_i \mid x_{B1}, \ldots, x_{Bn})$ by $\hat{s}_i$, or $\hat{s}_i^{(h-1)}$ if necessary, and similarly for the other random effects.

Since the penalty $P(\boldsymbol{\theta})$ does not involve the α_ks and β, we have

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial \beta} = \frac{1}{n} \sum_{i=1}^n (\frac{\hat{s}_i}{\beta^2} - \frac{\alpha}{\beta})$$

so

$$\hat{\beta}^{(h)} = \frac{1}{\hat{\alpha}^{(h)}} \frac{1}{n} \sum_{i=1}^n \hat{s}_i^{(h-1)},$$

although for simplicity we use $\hat{\alpha}^{(h-1)}$ instead of $\hat{\alpha}^{(h)}$ in the updating formula. Since $S_i \mid (m_i, \mathbf{t}_i)$ is Gamma$(m_i + \alpha, \beta/(\beta + 1))$, we have the explicit expression

$$\hat{s}_i^{(h-1)} = \frac{(m_i + \hat{\alpha}^{(h-1)})\hat{\beta}^{(h-1)}}{1 + \hat{\beta}^{(h-1)}}, \tag{1}$$

so

$$\hat{\beta}^{(h)} = \frac{(\bar{m}_\cdot + \hat{\alpha}^{(h-1)})\hat{\beta}^{(h-1)}}{\hat{\alpha}^{(h-1)}(1+\hat{\beta}^{(h-1)})}. \tag{2}$$

For the α_ks we have

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial \alpha_k} = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{\log s_i}^{(h-1)} - \log\beta - \psi(\alpha_k) + \widehat{\log w_{ik}}^{(h-1)}\}$$

where $\psi(\alpha_k) = \Gamma'(\alpha_k)/\Gamma(\alpha_k)$ is the digamma function. Then

$$\hat{\alpha}_k^{(h)} = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\log s_i}^{(h-1)} + \frac{1}{n}\sum_{i=1}^{n}\widehat{\log w_{ik}}^{(h-1)} - \log\hat{\beta}^{(h)}\right),$$

although as before we use $\hat{\beta}^{(h-1)}$ instead of $\hat{\beta}^{(h)}$ for simplicity. For $\widehat{\log s_i}^{(h-1)}$ there are explicit formulas: since $S_i \mid (m_i, \mathbf{t}_i)$ is Gamma$(m_i+\alpha, \beta/(\beta+1))$,

$$\widehat{\log s_i}^{(h-1)} = \psi(m_i + \hat{\alpha}^{(h-1)}) + \log\left\{\frac{\hat{\beta}^{(h-1)}}{1+\hat{\beta}^{(h-1)}}\right\}. \tag{3}$$

For $\widehat{\log w_{ik}}^{(h-1)}$, we know that $\mathbf{W}_i \mid (m_i, \mathbf{t}_i, \mathbf{y}_i)$ has a Dirichlet$(y_{i\cdot 1}+\alpha_1, \ldots, y_{i\cdot p}+\alpha_p)$ distribution, so

$$\begin{aligned}
\mathbb{E}\left(\log W_{ik} \mid m_i, \mathbf{t}_i\right) &= \mathbb{E}\left\{\mathbb{E}\left(\log W_{ik} \mid m_i, \mathbf{t}_i, \mathbf{y}_i\right) \mid m_i, \mathbf{t}_i\right\} \\
&= \mathbb{E}\left\{\psi(Y_{i\cdot k}+\alpha_k) - \psi(y_{i\cdot\cdot}+\alpha) \mid m_i, \mathbf{t}_i\right\} \\
&= \mathbb{E}\left\{\psi(Y_{i\cdot k}+\alpha_k) \mid m_i, \mathbf{t}_i\right\} - \psi(m_i+\alpha)
\end{aligned}$$

since $\sum_{k=1}^{p} y_{i\cdot k} = \sum_{j=1}^{m_i}\sum_{k=1}^{p} y_{ijk} = \sum_{j=1}^{m_i} 1 = m_i$. There is no closed expression for the last conditional expectation, but as an approximation we can plug in the current $\hat{y}_{i\cdot k}$s (see 7 below) and then

$$\widehat{\log w_{ik}}^{(h-1)} = \psi(\hat{y}_{i\cdot k}^{(h-1)} + \hat{\alpha}_k^{(h-1)}) - \psi(m_i + \hat{\alpha}^{(h-1)}). \tag{4}$$

After some simplifications we obtain

$$\hat{\alpha}_k^{(h)} = \psi^{-1}\left\{-\log(1+\hat{\beta}^{(h-1)}) + \frac{1}{n}\sum_{i=1}^{n}\psi(\hat{y}_{i\cdot k}^{(h-1)} + \hat{\alpha}_k^{(h-1)})\right\}. \tag{5}$$

The updating equations for $\mathrm{vec}(\hat{\mathbf{C}})$ are harder to obtain due to both the nature of $Q(\boldsymbol{\theta})$ and the constraints. The easiest way to deal with the linear constraint $\mathbf{A}\,\mathrm{vec}(\mathbf{C}) = \mathbf{1}_p$ is to reparameterize. Let $\boldsymbol{\Gamma}_1 \in \mathbb{R}^{pq\times p}$ be an orthogonal basis for the rows of $\mathbf{A}$, and $\boldsymbol{\Gamma}_2 \in \mathbb{R}^{pq\times(pq-p)}$ its orthogonal complement. Then $\mathrm{vec}(\mathbf{C}) = \boldsymbol{\Gamma}_1\mathbf{v}_1 + \boldsymbol{\Gamma}_2\mathbf{v}_2$ for $\mathbf{v}_1 \in \mathbb{R}^p$ and $\mathbf{v}_2 \in \mathbb{R}^{pq-p}$. Since $\mathbf{A}\,\mathrm{vec}(\mathbf{C}) = \mathbf{A}\boldsymbol{\Gamma}_1\mathbf{v}_1 = \mathbf{1}_p$, it follows that $\mathbf{v}_1 = (\mathbf{A}\boldsymbol{\Gamma}_1)^{-1}\mathbf{1}_p$, whereas $\mathbf{v}_2$ is free. Then we reparameterize $Q_{\hat{\boldsymbol{\theta}}^{(h-1)}}$ in terms of the unconstrained $\mathbf{v}_2$: $\tilde{Q}(\mathbf{v}_2, \boldsymbol{\eta}) := Q(\boldsymbol{\Gamma}_1\mathbf{v}_1 + \boldsymbol{\Gamma}_2\mathbf{v}_2, \boldsymbol{\eta})$, and it follows that $\hat{\mathbf{v}}_2^{(h)}$ satisfies the equation

$$\nabla_{\mathbf{v}_2}\tilde{Q}(\hat{\mathbf{v}}_2^{(h)}, \hat{\boldsymbol{\eta}}^{(h)}) = \boldsymbol{\Gamma}_2^T\nabla_{\mathrm{vec}\,\mathbf{C}}Q(\hat{\boldsymbol{\theta}}^{(h)}) = \mathbf{0}_{pq-p}.$$

The Hessian of $\tilde{Q}$ with respect to $\mathbf{v}_2$ is

$$\nabla^2_{\mathbf{v}_2}\tilde{Q}(\mathbf{v}_2, \boldsymbol{\eta}) = \boldsymbol{\Gamma}_2^T \nabla^2_{\mathrm{vec}\,\mathbf{C}} Q(\boldsymbol{\theta}) \boldsymbol{\Gamma}_2.$$

We cannot solve the estimating equation for $\hat{\mathbf{v}}_2^{(h)}$ in closed form but we can approximate it by a Newton-Raphson step:

$$\hat{\mathbf{v}}_2^{(h)} = \hat{\mathbf{v}}_2^{(h-1)} - \tau\{\nabla^2_{\mathbf{v}_2}\tilde{Q}(\hat{\mathbf{v}}_2^{(h-1)}, \hat{\boldsymbol{\eta}}^{(h-1)})\}^{-1}\nabla_{\mathbf{v}_2}\tilde{Q}(\hat{\mathbf{v}}_2^{(h-1)}, \hat{\boldsymbol{\eta}}^{(h-1)}),$$

where τ is the step size (usually 1, but sometimes chosen smaller in order to guarantee improvement of the objective function). In terms of the original parameter $\mathrm{vec}\,\mathbf{C}$, this comes down to

$$\mathrm{vec}\,\hat{\mathbf{C}}^{(h)} = \mathrm{vec}\,\hat{\mathbf{C}}^{(h-1)} - \tau\boldsymbol{\Gamma}_2\{\boldsymbol{\Gamma}_2^T \nabla^2_{\mathrm{vec}\,\mathbf{C}} Q(\hat{\boldsymbol{\theta}}^{(h-1)})\boldsymbol{\Gamma}_2\}^{-1}\{\boldsymbol{\Gamma}_2^T \nabla_{\mathrm{vec}\,\mathbf{C}} Q(\hat{\boldsymbol{\theta}}^{(h-1)})\}. \tag{6}$$

So far we have ignored the condition $\mathrm{vec}\,\mathbf{C} \geq 0$. We handle this by projecting (6) onto the feasible space, i.e. by setting the negative coordinates of the $\hat{\mathbf{c}}_k^{(h)}$s to zero and re-scaling the remaining coordinates so that $\mathbf{a}^T\hat{\mathbf{c}}_k^{(h)} = 1$ for all k.

The gradient and Hessian of Q with respect to $\mathrm{vec}\,\mathbf{C}$ are

$$\nabla_{\mathrm{vec}\,\mathbf{C}} Q(\boldsymbol{\theta}) = n^{-1}\mathbb{E}_{\hat{\boldsymbol{\theta}}^{(h-1)}}\{\nabla_{\mathrm{vec}\,\mathbf{C}}\ell(\boldsymbol{\theta}) \mid x_{B1}, \ldots, x_{Bn}\} - \zeta\nabla_{\mathrm{vec}\,\mathbf{C}} P(\boldsymbol{\theta})$$

and

$$\nabla^2_{\mathrm{vec}\,\mathbf{C}} Q(\boldsymbol{\theta}) = n^{-1}\mathbb{E}_{\hat{\boldsymbol{\theta}}^{(h-1)}}\{\nabla^2_{\mathrm{vec}\,\mathbf{C}}\ell(\boldsymbol{\theta}) \mid x_{B1}, \ldots, x_{Bn}\} - \zeta\nabla^2_{\mathrm{vec}\,\mathbf{C}} P(\boldsymbol{\theta}).$$

The derivatives of the penalty terms are computed in the next subsections. The derivatives involving $\ell(\boldsymbol{\theta})$ are:

$$\frac{\partial\tilde{\ell}(\boldsymbol{\theta})}{\partial\mathbf{c}_k} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ijk}\frac{\boldsymbol{\beta}(t_{ij})}{\phi_k(t_{ij})},$$

$$\frac{\partial^2\tilde{\ell}(\boldsymbol{\theta})}{\partial\mathbf{c}_k\partial\mathbf{c}_k^T} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ijk}\frac{\boldsymbol{\beta}(t_{ij})\boldsymbol{\beta}(t_{ij})^T}{\phi_k^2(t_{ij})},$$

and

$$\frac{\partial^2\tilde{\ell}(\boldsymbol{\theta})}{\partial\mathbf{c}_k\partial\mathbf{c}_l^T} = \mathbf{O} \text{ if } k \neq l.$$

Then, if we define

$$\mathbf{g}_k^{(h-1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{y}_{ijk}^{(h-1)}\frac{\boldsymbol{\beta}(t_{ij})}{\hat{\phi}_k^{(h-1)}(t_{ij})}$$

and

$$\mathbf{H}_k^{(h-1)} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{y}_{ijk}^{(h-1)}\frac{\boldsymbol{\beta}(t_{ij})\boldsymbol{\beta}(t_{ij})^T}{\{\hat{\phi}_k^{(h-1)}(t_{ij})\}^2},$$

we have

$$n^{-1}\mathbb{E}_{\hat{\boldsymbol{\theta}}^{(h-1)}}\{\nabla_{\mathrm{vec}\,\mathbf{C}}\tilde{\ell}(\boldsymbol{\theta}) \mid x_{B1}, \ldots, x_{Bn}\} = \begin{bmatrix} \mathbf{g}_1^{(h-1)} \\ \vdots \\ \mathbf{g}_p^{(h-1)} \end{bmatrix}$$

and

$$n^{-1}\mathbb{E}_{\hat{\boldsymbol{\theta}}^{(h-1)}}\{\nabla^2_{\mathrm{vec}\,\mathbf{C}}\tilde{\ell}(\boldsymbol{\theta}) \mid x_{B1},\ldots,x_{Bn}\} = \begin{bmatrix} \mathbf{H}_1^{(h-1)} & & \\ & \ddots & \\ & & \mathbf{H}_p^{(h-1)} \end{bmatrix}.$$

The previous formulas make use of the $\hat{y}_{ijk}^{(h-1)}$s. To compute the $\hat{y}_{ijk}$s we use the fact that $\mathbf{Y}_{ij} \mid (m_i, \mathbf{t}_i, \mathbf{w}_i)$ is Multinomial$(1, \boldsymbol{\xi}_{ij})$ with $\xi_{ijk} = w_{ik}\phi_k(t_{ij}) / \sum_{l=1}^p w_{il}\phi_l(t_{ij})$. Then

$$\begin{aligned} \mathbb{E}\left(Y_{ijk} \mid m_i, \mathbf{t}_i\right) &= \mathbb{E}\left\{\mathbb{E}\left(Y_{ijk} \mid m_i, \mathbf{t}_i, \mathbf{w}_i\right) \mid m_i, \mathbf{t}_i\right\} \\ &= \mathbb{E}\left\{\frac{W_{ik}\phi_k(t_{ij})}{\sum_{l=1}^p W_{il}\phi_l(t_{ij})} \mid m_i, \mathbf{t}_i\right\}, \end{aligned}$$

for which we do not have a closed form but can approximate by plugging in the previous $\hat{w}_{ik}$s:

$$\hat{y}_{ijk}^{(h)} = \frac{\hat{w}_{ik}^{(h-1)}\hat{\phi}_k^{(h)}(t_{ij})}{\sum_{l=1}^p \hat{w}_{il}^{(h-1)}\hat{\phi}_l^{(h)}(t_{ij})}. \tag{7}$$

Note that the $\hat{y}_{ijk}$s satisfy the condition $\sum_{k=1}^p \hat{y}_{ijk} = 1$ for all i and j.

Next we update the $\hat{w}_{ik}$s. We use the fact that $\mathbf{W}_i \mid (m_i, \mathbf{t}_i, \mathbf{y}_i)$ is Dirichlet$(y_{i\cdot 1} + \alpha_1, \ldots, y_{i\cdot p} + \alpha_p)$, so

$$\begin{aligned} \mathbb{E}\left(W_{ik} \mid m_i, \mathbf{t}_i\right) &= \mathbb{E}\left\{\mathbb{E}\left(W_{ik} \mid m_i, \mathbf{t}_i, \mathbf{y}_i\right) \mid m_i, \mathbf{t}_i\right\} \\ &= \mathbb{E}\left(\frac{Y_{i\cdot k} + \alpha_k}{m_i + \alpha} \mid m_i, \mathbf{t}_i\right), \end{aligned}$$

which we approximate by plugging in the new $\hat{y}_{ijk}$s:

$$\hat{w}_{ik}^{(h)} = \frac{y_{i\cdot k}^{(h)} + \hat{\alpha}_k^{(h)}}{m_i + \hat{\alpha}^{(h)}}. \tag{8}$$

Finally, we compute $\hat{s}_i^{(h)}$, $\widehat{\log s_i}^{(h)}$ and $\widehat{\log w_{ik}}^{(h)}$ using (1), (3) and (4), respectively, and update the value of the objective function:

$$Q_{\hat{\boldsymbol{\theta}}^{(h-1)}}(\hat{\boldsymbol{\theta}}^{(h)}) = n^{-1}\mathbb{E}_{\hat{\boldsymbol{\theta}}^{(h-1)}}\{\tilde{\ell}(\hat{\boldsymbol{\theta}}^{(h)}) \mid x_{B1},\ldots,x_{Bn}\} - \zeta P(\hat{\boldsymbol{\theta}}^{(h)}),$$

where the conditional expectation is approximated by plugging in the newly estimated random effects:

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}^{(h-1)}}\{\tilde{\ell}(\hat{\boldsymbol{\theta}}^{(h)}) \mid x_{B1},\ldots,x_{Bn}\} \approx \tilde{\ell}^{(h)}(\hat{\boldsymbol{\theta}}^{(h)}).$$

To summarize, the updating order at each iteration of the algorithm is the following:

1. Update the parameter estimators $\hat{\beta}$, $\{\hat{\alpha}_k\}$ and $\{\hat{\mathbf{c}}_k\}$ using formulas (2), (5) and (6), all of which depend on the previous parameter estimators and on the previous $\hat{y}_{ijk}$s.
2. Update the $\hat{y}_{ijk}$s using (7).
3. Update the $\hat{w}_{ik}$s using (8).

4. Update the rest of the random-effect estimators, $\hat{s}_i$, $\widehat{\log s_i}$ and $\widehat{\log w_{ik}}$, and the objective function $Q_{\hat{\boldsymbol{\theta}}^{(h-1)}}(\hat{\boldsymbol{\theta}}^{(h)})$.

1.2 Initialization

As initial estimators of the EM-algorithm we use the following. First, we choose $\hat{\mathbf{c}}_1^{(0)}, \ldots, \hat{\mathbf{c}}_p^{(0)}$, randomly or otherwise (for example, by splitting the region B into p subregions). Then we assign each t_{ij} to the group k with largest density at t_{ij}:

$$\hat{y}_{ijk}^{(0)} = 1 \iff k = \underset{l=1,\ldots,p}{\operatorname{argmax}} \hat{\phi}_l^{(0)}(t_{ij}).$$

The rest are basically Method of Moments estimators:

$$\hat{w}_{ik}^{(0)} = \overline{\hat{y}_{i\cdot k}^{(0)}},$$

$$\hat{s}_i^{(0)} = m_i,$$

$$\hat{\beta}^{(0)} = \frac{\operatorname{var}(\hat{s}_i^{(0)})}{\operatorname{mean}(\hat{s}_i^{(0)})}, \quad \hat{\alpha}^{(0)} = \frac{\operatorname{mean}(\hat{s}_i^{(0)})}{\hat{\beta}^{(0)}},$$

$$\hat{\alpha}_k^{(0)} = \overline{\hat{w}_{\cdot k}^{(0)}}\hat{\alpha}^{(0)}.$$

1.3 Derivatives of smoothness penalty

For time-dependent processes,

$$\begin{aligned} P(\boldsymbol{\theta}) &= \sum_{k=1}^{p} \int_{B_0} \{\phi_k''(t)\}^2 dt = \sum_{k=1}^{p} \mathbf{c}_k^T \mathbf{J}_2 \mathbf{c}_k \\ &= \operatorname{vec} \mathbf{C}^T \boldsymbol{\Omega} \operatorname{vec} \mathbf{C} \end{aligned}$$

with $\mathbf{J}_2 = \int_{B_0} \ddot{\boldsymbol{\beta}}(t)\ddot{\boldsymbol{\beta}}(t)^T dt$ and $\boldsymbol{\Omega} = \mathbf{I}_p \otimes \mathbf{J}_2$, and for spatial processes,

$$\begin{aligned} P(\boldsymbol{\theta}) &= \sum_{k=1}^{p} \iint_{B_0} \left\{ (\frac{\partial^2 \phi_k}{\partial t_1^2})^2 + 2(\frac{\partial^2 \phi_k}{\partial t_1 \partial t_2})^2 + (\frac{\partial^2 \phi_k}{\partial t_2^2})^2 \right\} dt_1 dt_2 \\ &= \sum_{k=1}^{p} \mathbf{c}_k^T (\mathbf{J}_{2,11} + 2\mathbf{J}_{2,12} + \mathbf{J}_{2,22}) \mathbf{c}_k \\ &= \operatorname{vec} \mathbf{C}^T \boldsymbol{\Omega} \operatorname{vec} \mathbf{C} \end{aligned}$$

as before, with

$$\mathbf{J}_{2,ij} = \iint_{B_0} (\frac{\partial^2 \boldsymbol{\beta}}{\partial t_i \partial t_j})(\frac{\partial^2 \boldsymbol{\beta}}{\partial t_i \partial t_j})^T dt_1 dt_2$$

and $\boldsymbol{\Omega} = \mathbf{I}_p \otimes (\mathbf{J}_{2,11} + 2\mathbf{J}_{2,12} + \mathbf{J}_{2,22})$. So

$$\nabla_{\operatorname{vec} \mathbf{C}} P(\boldsymbol{\theta}) = 2\boldsymbol{\Omega} \operatorname{vec} \mathbf{C} \quad \text{and} \quad \nabla^2_{\operatorname{vec} \mathbf{C}} P(\boldsymbol{\theta}) = 2\boldsymbol{\Omega}$$

in either case.

2 Theory

2.1 Model identifiability

The model

$$\Lambda(t) = \sum_{k=1}^{p} U_k \phi_k(t)$$

is identifiable if:

1. the U_ks are independent, non-negative and non-singular random variables (their non-negativity automatically precludes their being Gaussian);
2. the Gram matrix $\mathbf{G}$ with elements $G_{ij} = \int \phi_i \phi_j$ is nonsingular;
3. $\int \phi_k = 1$ for all k;
4. the $V(U_k)$s are decreasing in k (or some other condition that specifies the ordering of the U_ks).

To prove this, suppose $\Lambda(t) = \mathbf{U}^T \boldsymbol{\phi}(t) = \tilde{\mathbf{U}}^T \tilde{\boldsymbol{\phi}}(t)$ for all $t \in B$. Then $\tilde{\mathbf{U}}\Lambda(t) = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T \tilde{\boldsymbol{\phi}}(t)$ and $\tilde{\boldsymbol{\phi}}(t) = E(\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T)^{-1} \times E\{\tilde{\mathbf{U}}\Lambda(t)\}$. The matrix $E(\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T)$ is nonsingular because the $\tilde{U}_k$s are independent and nondegenerate, so there is no $\mathbf{v} \neq \mathbf{0}$ for which $P(\tilde{\mathbf{U}}^T\mathbf{v} = 0) = 1$. On the other hand we have $\Lambda(t) = \mathbf{U}^T\boldsymbol{\phi}(t)$, so $\tilde{\boldsymbol{\phi}}(t) = E(\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T)^{-1}E(\tilde{\mathbf{U}}\mathbf{U}^T)\boldsymbol{\phi}(t)$, that is, $\tilde{\boldsymbol{\phi}}(t) = \mathbf{A}\boldsymbol{\phi}(t)$ for some $\mathbf{A}$.

Then $\mathbf{U}^T\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^T = \tilde{\mathbf{U}}^T\mathbf{A}\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^T$ for all $t \in B_0$, and since the Gram matrix $\int_{B_0} \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^T dt$ is nonsingular, we have $\mathbf{U} = \mathbf{A}^T\tilde{\mathbf{U}}$. Since the U_ks and the $\tilde{U}_k$s are independent and their distributions are neither Gaussian nor singular, Theorem 10 of Comon (1994) implies $\mathbf{A}^T = \mathbf{PD}$ with $\mathbf{P}$ a permutation matrix and $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$. Since $\int \phi_k = \int \tilde{\phi}_k = 1$ for all k, it follows that $d_k = 1$ for all k. If the order of the U_ks is specified, for example that $V(U_k)$ is decreasing in k, then $\mathbf{P} = \mathbf{I}$ and the model is identifiable.

2.2 Consistency

The consistency proof follows the usual steps for maximum likelihood estimators and M-estimators; see examples in Pollard (1984) and Van der Vaart (2000). We show that the asymptotic objective function has a unique maximum at $\boldsymbol{\theta}_0$, that $\{\hat{\boldsymbol{\theta}}_n\}$ is bounded in probability, and then, via the Argmax Theorem, that $\hat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$ in probability.

Lemma 1 *The function $M(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}\{\log f(X_B; \boldsymbol{\theta})\}$ has a unique maximum at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.*

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

$$E_{\boldsymbol{\theta}_0}\left\{\log \frac{f(X_B;\boldsymbol{\theta})}{f(X_B;\boldsymbol{\theta}_0)}\right\} \leq \log E_{\boldsymbol{\theta}_0}\left\{\frac{f(X_B;\boldsymbol{\theta})}{f(X_B;\boldsymbol{\theta}_0)}\right\} = 0 \tag{9}$$

because

$$\begin{aligned} E_{\boldsymbol{\theta}_0}\left\{\frac{f(X_B;\boldsymbol{\theta})}{f(X_B;\boldsymbol{\theta}_0)}\right\} &= \sum_{m=0}^{\infty}\int_B\cdots\int_B \frac{f(\{t_1,\ldots,t_m\};\boldsymbol{\theta})}{f(\{t_1,\ldots,t_m\};\boldsymbol{\theta}_0)}\; f(\{t_1,\ldots,t_m\};\boldsymbol{\theta}_0)\; dt_1\cdots dt_m \\ &= \sum_{m=0}^{\infty}\int_B\cdots\int_B f(\{t_1,\ldots,t_m\};\boldsymbol{\theta})\; dt_1\cdots dt_m \\ &= 1 \end{aligned}$$

for any $\boldsymbol{\theta}$. Moreover, inequality (9) is strict unless $P_{\boldsymbol{\theta}_0}\{f(X_B;\boldsymbol{\theta})/f(X_B;\boldsymbol{\theta}_0) = 1\} = 1$, and by identifiability this happens only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then $E_{\boldsymbol{\theta}_0}\{\log f(X_B;\boldsymbol{\theta})\} < E_{\boldsymbol{\theta}_0}\{\log f(X_B;\boldsymbol{\theta}_0)\}$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. ■

Lemma 2 $\|\hat{\boldsymbol{\theta}}_n\| = O_P(1)$.

Proof. Let

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\log f(X_{iB};\boldsymbol{\theta}).$$

Since $\hat{\boldsymbol{\theta}}_n$ maximizes $\rho_n(\boldsymbol{\theta})$ we have $\rho_n(\hat{\boldsymbol{\theta}}_n) - \rho_n(\boldsymbol{\theta}_0) \geq 0$, or equivalently

$$M_n(\hat{\boldsymbol{\theta}}_n) - M_n(\boldsymbol{\theta}_0) \geq \zeta_n\{P(\hat{\boldsymbol{\theta}}_n) - P(\boldsymbol{\theta}_0)\}.$$

The penalty function is nonnegative, so this implies

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_{iB};\hat{\boldsymbol{\theta}}_n)}{f(X_{iB};\boldsymbol{\theta}_0)} \geq -\zeta_n P(\boldsymbol{\theta}_0), \tag{10}$$

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

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

which satisfy

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

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

$$g(x_B;\boldsymbol{\theta}) \geq \frac{1}{2}\log\frac{f(X_B;\boldsymbol{\theta})}{f(X_B;\boldsymbol{\theta}_0)} + \frac{1}{2}\log(1) = \frac{1}{2}\log\frac{f(X_B;\boldsymbol{\theta})}{f(X_B;\boldsymbol{\theta}_0)},$$

so (10) implies

$$\frac{1}{n}\sum_{i=1}^{n} g(X_{iB};\hat{\boldsymbol{\theta}}_n) \geq \frac{1}{2}\{-\zeta_n P(\boldsymbol{\theta}_0)\}. \tag{11}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} g(X_{iB};\hat{\boldsymbol{\theta}}_n) \leq \frac{1}{n}\sum_{i=1}^{n}\psi(X_{iB}) \tag{12}$$

with

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

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

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

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

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

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

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

Therefore, there exists an $\varepsilon > 0$ and a $K > 0$ such that $E_{\boldsymbol{\theta}_0}\{\psi(X_B)\} < -\varepsilon$. This fact together with (11) and (12) imply that $P(\|\hat{\boldsymbol{\theta}}_n\| \geq K)$ goes to zero as $n \to \infty$. ■

Lemma 3 $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

Proof. By Lemma 2, for any $\varepsilon > 0$ we can choose $K > 0$ such that $P\{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > K\} < \varepsilon/2$ for all n. So let us focus on the set $\{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq K\}$. In this case

$$\hat{\boldsymbol{\theta}}_n = \operatorname*{argmax}_{\Theta \cap \{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq K\}} \rho_n(\boldsymbol{\theta}),$$

with

$$\rho_n(\boldsymbol{\theta}) = M_n(\boldsymbol{\theta}) - \zeta_n P(\boldsymbol{\theta}).$$

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

$$\operatorname*{argmax}_{\Theta \cap \{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq K\}} \rho_n(\boldsymbol{\theta}) \xrightarrow{P} \operatorname*{argmax}_{\Theta \cap \{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq K\}} M(\boldsymbol{\theta}) = \boldsymbol{\theta}_0,$$

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

2.3 Asymptotic normality

Lemma 4 *If* $\zeta_n = O_p(n^{-1/2})$ *then* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_P(n^{-1/2})$.

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

$$\tilde{\rho}_n(\boldsymbol{\theta}) = n\{\rho_n(\boldsymbol{\theta}) - \rho_n(\boldsymbol{\theta}_0)\},$$

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

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

and $M(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0} \{\log f(X_B; \boldsymbol{\theta})\}$ as above. Then

$$\begin{aligned} \check{\rho}_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \nabla \log f(X_{iB}, \boldsymbol{\theta}_0)^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \sum_{i=1}^{n} [r(X_{iB}, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - E_{\boldsymbol{\theta}_0}\{r(X_B, \boldsymbol{\theta}, \boldsymbol{\theta}_0)\}] \\ &\quad + n\{M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}_0)\} - n\zeta_n\{P(\boldsymbol{\theta}) - P(\boldsymbol{\theta}_0)\}. \end{aligned}$$

Note that $E_{\theta_0}\{\nabla \log f(X_B, \boldsymbol{\theta}_0)\} = \nabla M(\boldsymbol{\theta}_0) = \mathbf{0}$ because $f(x_B, \boldsymbol{\theta})$ is a density; the fact that $\boldsymbol{\theta}_0$ maximizes $M(\boldsymbol{\theta})$ does not automatically imply $\nabla M(\boldsymbol{\theta}_0) = \mathbf{0}$ because $\boldsymbol{\theta}_0$ may be on the border of Θ. Let

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

and

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

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

$$\begin{aligned} & \sqrt{n}\mathbf{Z}_n^T(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| R_n(\hat{\boldsymbol{\theta}}_n) - n\zeta_n\{P(\hat{\boldsymbol{\theta}}_n) - P(\boldsymbol{\theta}_0)\} \\ \geq \; & -n\{M(\hat{\boldsymbol{\theta}}_n) - M(\boldsymbol{\theta}_0)\}. \end{aligned} \tag{13}$$

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 $P(\boldsymbol{\theta})$ implies

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

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 (13) that

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

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 $\mathbf{F}_0$,

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

Since $\sqrt{n}\zeta_n = O_P(1)$ we finally arrive at the inequality

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

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

Theorem 5 *If $\sqrt{n}\zeta_n \to \kappa$ as $n \to \infty$, either deterministically or in probability, then $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} \boldsymbol{\delta}(\mathbf{Z})$, the maximizer of*

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

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

Proof. Let $W_n(\boldsymbol{\delta}) = \tilde{\rho}_n(\boldsymbol{\theta}_0 + \boldsymbol{\delta}/\sqrt{n})$ with $\tilde{\rho}_n(\boldsymbol{\theta})$ as above. Then

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

and $\hat{\boldsymbol{\delta}}_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ maximizes $W_n(\boldsymbol{\delta})$ over $\mathscr{D}_n = \sqrt{n}(\Theta - \{\boldsymbol{\theta}_0\})$. 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 $\mathscr{D}_n \cap \{\|\boldsymbol{\delta}\| \leq K\}$. In the limit as $n \to \infty$ we have:

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

(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

$$n\zeta_n\{P(\boldsymbol{\theta}_0 + \boldsymbol{\delta}/\sqrt{n}) - P(\boldsymbol{\theta}_0)\} = \sqrt{n}\zeta_n\{\nabla P(\boldsymbol{\theta}_0) + o_P(1)\}^T\boldsymbol{\delta}.$$

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

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

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

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

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

2.4 Explicit $\nabla \log f(x_B;\boldsymbol{\theta})$ for Gamma scores

Suppose $U_k \sim \text{Gamma}(\alpha_k, \beta)$ for $k = 1, \ldots, p$. Then

$$\begin{aligned} f(x_B) &= \int f(m, \mathbf{t}, \mathbf{u})\ d\mathbf{u} \\ &= \int f(\mathbf{t} \mid m, \mathbf{u}) f(m \mid \mathbf{u}) f(\mathbf{u})\ d\mathbf{u} \end{aligned}$$

with

$$f(\mathbf{t} \mid m, \mathbf{u}) f(m \mid \mathbf{u}) = \exp\{-\int_B \lambda_{\mathbf{u}}(t)dt\} \prod_{j=1}^{m} \lambda_{\mathbf{u}}(t_j),$$

for

$$\lambda_{\mathbf{u}}(t) = \sum_{k=1}^{p} u_k \mathbf{c}_k^T \boldsymbol{\beta}(t),$$

and

$$f(\mathbf{u}) = \prod_{k=1}^{p} \frac{u_k^{\alpha_k - 1} e^{-u_k/\beta}}{\Gamma(\alpha_k)\beta^{\alpha_k}} I_{(0,\infty)}(u_k).$$

Note that

$$\begin{aligned} \nabla \log f(x_B;\boldsymbol{\theta}) &= \frac{1}{f(x_B;\boldsymbol{\theta})} \int \nabla f(m, \mathbf{t}, \mathbf{u})\ d\mathbf{u} \\ &= \frac{1}{f(x_B;\boldsymbol{\theta})} \int \nabla \log f(m, \mathbf{t}, \mathbf{u}) \cdot f(m, \mathbf{t}, \mathbf{u})\ d\mathbf{u} \\ &= \int \nabla \log f(m, \mathbf{t}, \mathbf{u}) \cdot f(\mathbf{u} \mid x_B)\ d\mathbf{u}. \end{aligned}$$

Then, recalling that $\mathbf{a} = \int_B \boldsymbol{\beta}(t)dt$, we have

$$\begin{aligned} \nabla_{\mathbf{c}_k} \log f(m, \mathbf{t}, \mathbf{u}) &= \nabla_{\mathbf{c}_k} \log\{f(\mathbf{t} \mid m, \mathbf{u}) f(m \mid \mathbf{u})\} \\ &= \nabla_{\mathbf{c}_k} \{-\int_B \lambda_{\mathbf{u}}(t)dt + \sum_{j=1}^{m} \lambda_{\mathbf{u}}(t_j)\} \\ &= \nabla_{\mathbf{c}_k} \{-\sum_{k=1}^{p} u_k \mathbf{c}_k^T \mathbf{a} + \sum_{j=1}^{m} \sum_{k=1}^{p} u_k \mathbf{c}_k^T \boldsymbol{\beta}(t_j)\} \\ &= u_k \{-\mathbf{a} + \sum_{j=1}^{m} \boldsymbol{\beta}(t_j)\}, \end{aligned}$$

so

$$\nabla_{\mathbf{c}_k} \log f(x_B;\boldsymbol{\theta}) = \{-\mathbf{a} + \sum_{j=1}^{m} \boldsymbol{\beta}(t_j)\} \mathbb{E}(U_k \mid x_B).$$

Also,

$$\begin{aligned} \nabla_{\alpha_k} \log f(m, \mathbf{t}, \mathbf{u}) &= \nabla_{\alpha_k} \log f(\mathbf{u}) \\ &= \nabla_{\alpha_k} \{\sum_{k=1}^{p} (\alpha_k - 1) \log u_k - \frac{u_k}{\beta} - \log \Gamma(\alpha_k) - \alpha_k \log \beta\} \\ &= \log u_k - \psi(\alpha_k) - \log \beta \end{aligned}$$

and

$$\nabla_\beta \log f(m, \mathbf{t}, \mathbf{u}) = \frac{\sum_{k=1}^p u_k}{\beta^2} - \frac{\sum_{k=1}^p \alpha_k}{\beta},$$

so

$$\nabla_{\alpha_k} \log f(x_B; \boldsymbol{\theta}) = \mathbb{E}(\log U_k \mid x_B) - \psi(\alpha_k) - \log \beta$$

and

$$\nabla_\beta \log f(x_B; \boldsymbol{\theta}) = \frac{\sum_{k=1}^p \mathbb{E}(U_k \mid x_B)}{\beta^2} - \frac{\sum_{k=1}^p \alpha_k}{\beta}.$$

In practice, Fisher's Information matrix $\mathbf{F}_0$ is estimated by

$$\hat{\mathbf{F}} = \frac{1}{n} \sum_{i=1}^n \nabla \log f(x_{Bi}; \hat{\boldsymbol{\theta}}) \nabla \log f(x_{Bi}; \hat{\boldsymbol{\theta}})^T,$$

where $\hat{u}_{ik}$ and $\widehat{\log u_{ik}}$ substitute the conditional expectations $\mathbb{E}(U_k \mid x_B)$ and $\mathbb{E}(\log U_k \mid x_B)$. Since $\hat{u}_{ik}$ and $\widehat{\log u_{ik}}$ are by-products of the EM algorithm, no new computations are needed.

References

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing* **36** 287–314.

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

Pollard, D. (1984). *Convergence of Stochastic Processes.* New York: Springer.

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