Appendix to "Conditional logistic regression with longitudinal follow up and individual-level random coefficients: A stable and efficient two-step estimation method"

Radu V. Craiu

Department of Statistics, University of Toronto

Thierry Duchesne

Département de mathématiques et de statistique, Université Laval

Daniel Fortin

Centre d'étude de la forêt and Département de biologie, Université Laval

Sophie Baillargeon

Département de mathématiques et de statistique, Université Laval

February 2, 2011

A General description of the proposed EM-REML algorithm

Let us stack the $\{\hat{\boldsymbol{\beta}}_c, c = 1, \ldots, K\}$ obtained in the first step in a column vector $\widehat{\boldsymbol{\beta}}_{1st} = (\hat{\boldsymbol{\beta}}_1^\top, \ldots, \hat{\boldsymbol{\beta}}_K^\top)^\top$ of length Kp. Let $\mathbf{D}$ denote the between-cluster variancecovariance matrix of the K random effect vectors: $\mathbf{D} = Var[(\mathbf{b}_1^\top, \dots, \mathbf{b}_K^\top)^\top]$. Thus $\mathbf{D}$ is block diagonal with K identical blocks, each equal to $\boldsymbol{\Sigma}$ and the parameters $\boldsymbol{\theta}$ in $F(\mathbf{b}; \boldsymbol{\theta})$ are the distinct elements of $\boldsymbol{\Sigma}$.

Now let $\hat{\mathbf{D}}$ be an estimate of $\mathbf{D}$, $\mathbf{Q}$ be the $p \times Kp$ matrix given by $\mathbf{1}_K^\top \otimes \mathbf{I}_p$, with $\mathbf{1}_h$, $\mathbf{I}_h$ and $\otimes$ respectively denoting a vector of 1's of length h, the $h \times h$ identity matrix and the Kronecker product. Set $\hat{\mathbf{R}} = \text{diag}\{\hat{\mathbf{R}}_1, \dots, \hat{\mathbf{R}}_K\}$. Then $\boldsymbol{\beta}$ is estimated by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Q}\hat{\mathbf{V}}^{-1}\mathbf{Q}^\top\right)^{-1} \mathbf{Q}\hat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\beta}}_{1st}, \tag{A-1}$$

where $\hat{\mathbf{V}} = \hat{\mathbf{D}} + \hat{\mathbf{R}}$. The variance of $\hat{\boldsymbol{\beta}}$ given by (A-1) can be estimated by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{Q}\hat{\mathbf{V}}^{-1}\mathbf{Q}^\top\right)^{-1}. \tag{A-2}$$

All quantities in (A-1) are obtained in the first step except $\hat{\mathbf{D}}$ which is obtained in the second step as follows. Let $\tilde{\boldsymbol{\beta}}$ denote the subset of dimension q of $\boldsymbol{\beta}$ that corresponds to the random regression coefficients and $\widehat{\tilde{\boldsymbol{\beta}}}_c$ be the corresponding cluster-level first-step estimates that are stacked in the Kq vector $\widehat{\tilde{\boldsymbol{\beta}}}$. We define $\{(\tilde{\mathbf{b}}_c, \tilde{\mathbf{R}}_c),\ c = 1 \dots K\}$, $\tilde{\mathbf{D}}$, $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\mathbf{R}}$ in similar fashion. Define $\boldsymbol{\phi} = (\tilde{\mathbf{b}}_1^\top, \dots, \tilde{\mathbf{b}}_K^\top)^\top$ and put $U_{cj} = \widehat{\tilde{\beta}}_{cj}$. Then $\boldsymbol{\phi} \sim N_{Kq}(\mathbf{0}, \tilde{\mathbf{D}})$, with $\tilde{\mathbf{D}}$ depending on a vector of parameters, say $\boldsymbol{\theta}$. Under the considered scenario, given the vectors $\tilde{\mathbf{b}}_c$, we have the following linear mixed model for the regression coefficient estimates:

$$\mathbf{U} = \mathbf{W_1}\tilde{\boldsymbol{\beta}} + \mathbf{W_2}\boldsymbol{\phi} + \boldsymbol{\varepsilon}, \tag{A-3}$$

where $\mathbf{U} = (U_{11}, \dots, U_{1q}, \dots, U_{Kq})^\top$, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \dots, \tilde{\beta}_q)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \dots, \varepsilon_{Kq})^\top$, $\mathbf{W_1} = \mathbf{1}_K \otimes \mathbf{I}_q$, $\mathbf{W_2} = \mathbf{I}_{Kq}$, and $\boldsymbol{\varepsilon}^\top = (\varepsilon_{c1}, \dots, \varepsilon_{cq})$, $c = 1, \dots, K$ are independent $N_q(\mathbf{0}, \tilde{\mathbf{R}}_c)$. Hence $\tilde{\mathbf{R}} = Var(\boldsymbol{\varepsilon})$ is the block diagonal matrix $\tilde{\mathbf{R}} = \text{diag}(\tilde{\mathbf{R}}_c, c = 1, \dots, K)$ and $\boldsymbol{\varepsilon} \sim N_{Kq}(\mathbf{0}, \tilde{\mathbf{R}})$.

Now let $\mathbf{m}_1, \dots, \mathbf{m}_d$, $d = Kq - \text{rank}(\mathbf{W}_1) = q(K-1)$, be vectors such that $\mathbf{m}_\ell^\top \mathbf{W}_1 = \mathbf{0}$, $\ell = 1, \dots, d$, and put $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_d]$. Given the specific form of $\mathbf{W}_1$ here, this can be done by setting $\mathbf{m}_\ell$ equal to the ℓth column of $\mathbf{I}_{Kq} - \frac{1}{K}\mathbf{W}_1\mathbf{W}_1^\top$. Then $\boldsymbol{\gamma} = \mathbf{M}^\top \mathbf{U} | \boldsymbol{\phi} \sim N_d(\mathbf{M}^\top \boldsymbol{\phi}, \mathbf{M}^\top \tilde{\mathbf{R}} \mathbf{M})$, with $\boldsymbol{\phi} \sim N_{Kq}(\mathbf{0}, \tilde{\mathbf{D}})$. The corresponding likelihood function is the restricted (or residual) likelihood and it forms the basis for

REML inference about $\boldsymbol{\theta}$. Numerical maximization of the residual likelihood with respect to $\boldsymbol{\theta}$ in our case was easy to implement and stable when using the EM-algorithm defined below. Assume that the "complete data" $(\boldsymbol{\gamma}, \boldsymbol{\phi})$ are observed and recall that at this step $\mathbf{M}$ and $\tilde{\mathbf{R}}$ are considered known. Then the complete data loglikelihood is proportional to

$$l_{com} \propto -\frac{K}{2}\ln\det(\tilde{\boldsymbol{\Sigma}}) - \frac{1}{2}\boldsymbol{\phi}^\top\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}.$$

In the E-step, we must compute the expected value of l_{com} with respect to the distribution of the unobserved $\boldsymbol{\phi}$ given the observed $\boldsymbol{\gamma}$ and a current value $\tilde{\mathbf{D}}^*$ of $\tilde{\mathbf{D}}$:

$$Q(\tilde{\mathbf{D}}|\tilde{\mathbf{D}}^*) = -\frac{K}{2}\ln\det(\tilde{\boldsymbol{\Sigma}}) - \frac{1}{2}E_{\tilde{\mathbf{D}}^*}[\boldsymbol{\phi}^\top\tilde{\mathbf{D}}^{-1}\boldsymbol{\phi}|\boldsymbol{\gamma}].$$

Since $\boldsymbol{\phi}|\boldsymbol{\gamma} \sim N_{Kq}(\boldsymbol{\mu}^{\tilde{\mathbf{D}}}, \boldsymbol{S}^{\tilde{\mathbf{D}}})$ with $\boldsymbol{S}^{\tilde{\mathbf{D}}} = \{\mathbf{M}(\mathbf{M}^\top\tilde{\mathbf{R}}\mathbf{M})^{-1}\mathbf{M}^\top + \tilde{\mathbf{D}}^{-1}\}^{-1}$ and $\boldsymbol{\mu}^{\tilde{\mathbf{D}}} = \boldsymbol{S}^{\tilde{\mathbf{D}}}\mathbf{M}(\mathbf{M}^\top\tilde{\mathbf{R}}\mathbf{M})^{-1}\boldsymbol{\gamma}$, we get

$$\begin{aligned}
Q(\tilde{\mathbf{D}}|\tilde{\mathbf{D}}^*) &= -\frac{K}{2}\ln\det(\tilde{\boldsymbol{\Sigma}}) - \frac{1}{2}E_{\tilde{\mathbf{D}}^*}[\mathrm{tr}(\boldsymbol{\phi}^\top\tilde{\mathbf{D}}^{-1}\boldsymbol{\phi})|\boldsymbol{\gamma}] \\
&= -\frac{K}{2}\ln\det(\tilde{\boldsymbol{\Sigma}}) - \frac{1}{2}\mathrm{tr}(E_{\tilde{\mathbf{D}}^*}[\tilde{\mathbf{D}}^{-1}\boldsymbol{\phi}\boldsymbol{\phi}^\top|\boldsymbol{\gamma}]) \\
&= -\frac{K}{2}\ln\det(\tilde{\boldsymbol{\Sigma}}) - \frac{1}{2}\mathrm{tr}\{\tilde{\mathbf{D}}^{-1}(\boldsymbol{S}^{\tilde{\mathbf{D}}^*} + \boldsymbol{\mu}^{\tilde{\mathbf{D}}^*}\boldsymbol{\mu}^{\tilde{\mathbf{D}}^*\top})\}.
\end{aligned}$$

At the M-step, $Q(\tilde{\mathbf{D}}|\tilde{\mathbf{D}}^*)$ must be maximized with respect to $\boldsymbol{\theta}$. The solution to this maximization depends on the particular form of the blocks of $\tilde{\mathbf{D}}$. First we find the maximizer of $Q(\tilde{\mathbf{D}}|\tilde{\mathbf{D}}^*)$ among all block diagonal matrices of the form $\tilde{\mathbf{D}} = \mathrm{diag}(\tilde{\boldsymbol{\Sigma}}, \ldots, \tilde{\boldsymbol{\Sigma}})$. Since $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{D}}^*$ are block diagonal matrices then so is $\boldsymbol{S}^{\tilde{\mathbf{D}}^*}$, say $\boldsymbol{S}^{\tilde{\mathbf{D}}^*} = \mathrm{diag}(\boldsymbol{S}_{11}^{\tilde{\mathbf{D}}^*}, \ldots, \boldsymbol{S}_{KK}^{\tilde{\mathbf{D}}^*})$. The maximization problem can be reformulated as

$$\begin{aligned}
\arg\max_{\boldsymbol{\Sigma}} Q(\tilde{\boldsymbol{\Sigma}}|\tilde{\mathbf{D}}^*) &= -\frac{1}{2}\sum_{c=1}^{K}\left[\ln\det(\tilde{\boldsymbol{\Sigma}}) + \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{S}_{cc}^{\tilde{\mathbf{D}}^*}) + \boldsymbol{\mu}_c^{\tilde{\mathbf{D}}^*\top}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_c^{\tilde{\mathbf{D}}^*}\right], \\
&= -\frac{K}{2}\left[\ln\det(\tilde{\boldsymbol{\Sigma}}) + \mathrm{tr}\left\{\tilde{\boldsymbol{\Sigma}}^{-1}\left(\frac{1}{K}\sum_{c=1}^{K}\boldsymbol{S}_{cc}^{\tilde{\mathbf{D}}^*} + \boldsymbol{\mu}_c^{\tilde{\mathbf{D}}^*}\boldsymbol{\mu}_c^{\tilde{\mathbf{D}}^*\top}\right)\right\}\right]
\end{aligned} \tag{A-4}$$

where $\boldsymbol{\mu}_c^{\tilde{\mathbf{D}}^*} = (\boldsymbol{\mu}_{q(c-1)+1}^{\tilde{\mathbf{D}}^*}, \ldots, \boldsymbol{\mu}_{cq}^{\tilde{\mathbf{D}}^*})$. Following Watson (1964) the maximizer of (A-4) is

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{K}\sum_{c=1}^{K}\left(\boldsymbol{S}_{cc}^{\tilde{\mathbf{D}}^*} + \boldsymbol{\mu}_c^{\tilde{\mathbf{D}}^*}\boldsymbol{\mu}_c^{\tilde{\mathbf{D}}^*\top}\right). \tag{A-5}$$

This maximization can be simplified if more restrictions are imposed on the form assumed for $\tilde{\mathbf{\Sigma}}$. To illustrate, if $\tilde{\mathbf{\Sigma}} = \mathrm{diag}(\theta_1^2, \ldots, \theta_q^2)$, then $Q(\tilde{\mathbf{D}}|\tilde{\mathbf{D}}^*)$ simplifies to

$$
\begin{aligned}
Q(\tilde{\mathbf{D}}|\tilde{\mathbf{D}}^*) = -\frac{K}{2}\sum_{j=1}^{q} \ln \theta_j^2 \\
& - \frac{1}{2}\mathrm{tr}\{\mathrm{diag}(1/\theta_1^2, \ldots, 1/\theta_q^2, \ldots, 1/\theta_1^2, \ldots, 1/\theta_q^2)(\boldsymbol{S}^{\tilde{\mathbf{D}}^*} + \boldsymbol{\mu}^{\tilde{\mathbf{D}}^*}\boldsymbol{\mu}^{\tilde{\mathbf{D}}^*\top})\}.
\end{aligned}
\tag{A-6}
$$

One can show directly that $Q(\tilde{\mathbf{D}}|\tilde{\mathbf{D}}^*)$ in (A-6) is maximized when

$$
\hat{\theta}_j^2 = \frac{1}{K}\mathrm{tr}\left[\sum_{c=1}^{K} \mathbf{A}^{(cj)}\right] = \frac{1}{K}\sum_{c=1}^{K} \mathbf{A}^{(cj)}_{\mathrm{diag}},
$$

where $\mathbf{A}^{(cj)}$ is a matrix of 0's, except for its $\{(c-1)q+j\}$th line, which is the $\{(c-1)q+j\}$th line of $\mathbf{A} = (\boldsymbol{S}^{\tilde{\mathbf{D}}^*} + \boldsymbol{\mu}^{\tilde{\mathbf{D}}^*}\boldsymbol{\mu}^{\tilde{\mathbf{D}}^*\top})$.

B Additional Simulation Results

We have considered additional simulations in the case in which $\mathbf{\Sigma}$ is assumed diagonal. The number of fixed and random effects vary, $p = q \in \{2, 8\}$ and we vary $\rho \in \{0, 0.6\}$. Note that when $\rho = 0.6$ the model is misspecified. In Table 1 we report the Monte Carlo averages and standard errors based on 1000 replicates for the two-step estimates for β_1, β_2, $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ for different values of ρ, p, q, s. Throughout we use $\beta_1 = 0.75$, $\beta_2 = 1.25$, $K = 30$, $S = 60$, $m = 2$ and $n = 12$.

References

Watson, G. (1964), "A note on the maximum likelihood," *Sankhya A*, 26, 303–304.

Scenario	$\beta_1 = 0.75$	$\beta_2 = 1.25$	$\mathbf{\Sigma}_{11} = s$	$\mathbf{\Sigma}_{22} = s$
$(p = q = 2, \rho = 0, s = 0.2)$	0.746 (0.089)	1.242 (0.092)	0.198 (0.064)	0.193 (0.063)
$(p = q = 2, \rho = 0.6, s = 0.2)$	0.740 (0.092)	1.236 (0.095)	0.197 (0.064)	0.191 (0.063)
$(p = q = 2, \rho = 0, s = 0.5)$	0.746 (0.135)	1.235 (0.135)	0.484 (0.148)	0.481 (0.143)
$(p = q = 2, \rho = 0.6, s = 0.5)$	0.742 (0.132)	1.236 (0.132)	0.479 (0.143)	0.477 (0.141)
$(p = q = 8, \rho = 0, s = 0.2)$	0.752 (0.101)	1.269 (0.102)	0.211 (0.079)	0.214 (0.083)
$(p = q = 8, \rho = 0.6, s = 0.2)$	0.785 (0.098)	1.296 (0.099)	0.213 (0.078)	0.217 (0.085)
$(p = q = 8, \rho = 0, s = 0.5)$	0.749 (0.147)	1.264 (0.150)	0.522 (0.165)	0.518 (0.179)
$(p = q = 8, \rho = 0.6, s = 0.5)$	0.781 (0.145)	1.299 (0.153)	0.546 (0.172)	0.537 (0.172)

Table 1: Simulation results when $\mathbf{\Sigma}$ is assumed diagonal. Throughout $\mathbf{\Sigma}_{11} = \mathbf{\Sigma}_{22} = s$, $\beta_1 = 0.75$ and $\beta_2 = 1, 25$. True values of the parameters p, q, s, ρ are reported in the column "Scenario". Each cell entry shows the Monte Carlo average estimate and the Monte Carlo standard error (between brackets) for the two-step estimator.