On the Behaviour of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models

by Sonja Greven and Thomas Kneib

Supplementary Material

This supplementary material, containing detailed proofs and additional simulation and application results, is available as a web appendix.

The package also contains the R package `cAIC`, implementing the analytic representation of the corrected conditional Akaike Information Criterion for linear mixed models.

Contents

A Proofs and Remarks

A.1 Proof of Theorem 1

Let $\psi_K = (\widetilde{\beta}^T, \widetilde{\sigma}^2, \widetilde{\lambda})$ be the true value of ψ and denote by $f_\psi(\cdot) = f(\cdot \mid \psi)$ the marginal likelihood,

$$2\log(f_\psi(y)) = -n\log(2\pi) - n\log(\sigma^2) - \log\{\det(V_\lambda)\} - \frac{(y - X\beta)^T V_\lambda^{-1}(y - X\beta)}{\sigma^2}$$

with $V_\lambda = V_* = I_n + \lambda Z\Sigma Z^T$. Write

$$2\left[\log\{f_{\widehat{\psi}(y)}(y)\} - \log\{f_{\psi_K}(y)\}\right] = 2\left[\log\{f_{\widehat{\psi}_0(y)}(y)\} - \log\{f_{\psi_K}(y)\}\right] + 2\left[\log\{f_{\widehat{\psi}(y)}(y)\} - \log\{f_{\widehat{\psi}_0(y)}(y)\}\right],$$

where $\widehat{\psi}(y) = (\widehat{\beta}^T, \widehat{\sigma}^2, \widehat{\lambda})^T$ is the maximum likelihood estimator, and $\widehat{\psi}_0(y) = (\widetilde{\beta}^T, \widehat{\sigma}_0^2, \widetilde{\lambda})$ with $\widehat{\sigma}_0^2 = (y - X\widetilde{\beta})^T V_{\widetilde{\lambda}}^{-1}(y - X\widetilde{\beta})/n$.

The first term is the contribution from σ^2. Using a Taylor expansion around 1, we can write it as

$$-n\log\left(\frac{\widehat{\sigma}_0^2}{\widetilde{\sigma}^2}\right) - n + n\frac{\widehat{\sigma}_0^2}{\widetilde{\sigma}^2} = \left(\frac{\widehat{\sigma}_0^2/\widetilde{\sigma}^2 - 1}{\sqrt{2/n}}\right)^2 + o_P(1),$$

which converges in distribution to a χ_1^2 variable, as it does in the general linear model. Thus, the expectation of the first term is asymptotically equal to one.

The second term is studied by (Crainiceanu and Ruppert, 2004, Theorems 1 - 3), who show that it is the sum of two terms, where one term converges to a χ_p^2 variable (the contribution from β) with expectation p asymptotically. The other term (the contribution from λ) has a point mass at zero for $\widetilde{\lambda} = 0$, and a second mixture component smaller or equal to χ_1^2 (see also Self and Liang, 1987; Stram and Lee, 1994). For $\widetilde{\lambda} = 0$, the point mass at zero is non-vanishing and between 0.5 and 1, depending on the setting (Crainiceanu et al., 2003). The expectation is then smaller than 1 even asymptotically. For $\widetilde{\lambda} > 0$, the boundary effect decreases with n, but vanishes very slowly for small values of $\widetilde{\lambda}$. If y can be subdivided into K independent subvectors, with $K \to \infty$, this term converges under regularity conditions to a χ_1^2 variable with expectation 1, compare Self and Liang (1987); Stram and Lee (1994); Giampaoli and Singer (2009). Crainiceanu and Ruppert (2004) show that this is not necessarily the case if this subdivision with $K \to \infty$ does not hold. In either case, the overall expectation depends on the true $\widetilde{\lambda}$, and is smaller than $(p + 2)$ asymptotically if $\widetilde{\lambda} = 0$.

Similarly, we have

$$\begin{aligned}
&2\,\mathrm{E}_z\left[\log\{f_{\psi_K}(z)\} - \log\{f_{\widehat{\psi}(y)}(z)\}\right] \\
&= n\log\left(\frac{\widehat{\sigma}^2}{\widetilde{\sigma}^2}\right) + \log\frac{\det(V_{\widehat{\lambda}})}{\det(V_{\widetilde{\lambda}})} + \mathrm{E}_z\left[\frac{1}{\widehat{\sigma}^2}(z - X\widehat{\beta})^T V_{\widehat{\lambda}}^{-1}(z - X\widehat{\beta}) - \frac{1}{\widetilde{\sigma}^2}(z - X\widetilde{\beta})^T V_{\widetilde{\lambda}}^{-1}(z - X\widetilde{\beta})\right] \\
&= \left\{n\log\left(\frac{\widehat{\sigma}^2}{\widetilde{\sigma}^2}\right) - n + n\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2}\right\} + \frac{1}{\widehat{\sigma}^2}(\widehat{\beta} - \widetilde{\beta})^T X^T V_{\widehat{\lambda}}^{-1} X(\widehat{\beta} - \widetilde{\beta}) + \left[\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2}(\mathrm{tr}[V_{\widehat{\lambda}}^{-1}V_{\widetilde{\lambda}}] - n) + \log\left\{\frac{\det(V_{\widehat{\lambda}})}{\det(V_{\widetilde{\lambda}})}\right\}\right].
\end{aligned}$$

As before, the first and second group of terms are the contributions from σ^2 and β, and converge in distribution to χ^2_1 and χ^2_p variables with expectations 1 and p, respectively. For $\widetilde{\lambda} = 0$, the last group of terms again has the same non-vanishing point mass at zero of between 0.5 and 1. For $\widetilde{\lambda} > 0$, the boundary effect again decreases with n. As before, the expectation with respect to y both depends on $\widetilde{\lambda}$ and is less than 1 asymptotically if $\widetilde{\lambda} = 0$.

The result follows from the definition of the AIC given in Section 2.2. □

A.2 Proof of Lemma 1

Profiling σ^2 out of (4) using

$$\widehat{\sigma}^2 = \frac{(y - X\widehat{\beta})^T V_*^{-1}(y - X\widehat{\beta})}{n-p} = \frac{(y - X\widehat{\beta})^T (y - X\widehat{\beta} - Z\widehat{b})}{n-p}, \tag{A.1}$$

consider the profile restricted log-likelihood for $\lambda = \tau^2/\sigma^2$,

$$\begin{aligned} \ell(\lambda) &= \log f(A^T y \mid \theta_*, \widehat{\sigma}^2(\theta_*)) \\ &= \text{const} - \frac{1}{2}\log(\det(V_*)) - \frac{1}{2}\log(\det(X^T V_*^{-1} X)) - \frac{1}{2}(n-p)\log(\widehat{\sigma}^2) - \frac{1}{2}(n-p). \end{aligned}$$

We have either $\widehat{\lambda} = 0$, or

$$\frac{\partial}{\partial\lambda}\ell(\lambda)\Big|_{\lambda=\widehat{\lambda}} = 0 \quad \Rightarrow \quad (n-p)\frac{y^T \widehat{P_*}^T \widehat{V_*}^{-1} Z\Sigma Z^T \widehat{V_*}^{-1} \widehat{P_*} y}{y^T \widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*} y} = \operatorname{tr}(\widehat{P_*} Z\Sigma Z^T \widehat{V_*}^{-1}),$$

where hat-notation again indicates dependence on the estimated parameter $\widehat{\theta_*} = \widehat{\lambda}$. Multiplying both sides by $\widehat{\lambda}$, we obtain for $\widehat{\lambda} \geq 0$

$$(n-p)\frac{y^T \widehat{P_*}^T \widehat{V_*}^{-1} Z\widehat{D_*} Z^T \widehat{V_*}^{-1} \widehat{P_*} y}{y^T \widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*} y} = \operatorname{tr}(\widehat{P_*} Z\widehat{D_*} Z^T \widehat{V_*}^{-1}). \tag{A.2}$$

Consider now the conditional log-likelihood. Then, we have

$$\begin{aligned} -2\log f(y|\widehat{\beta}, \widehat{b}, \widehat{\theta}) &= n\log(2\pi) + n\log(\widehat{\sigma}^2) + \frac{(y - X\widehat{\beta} - Z\widehat{b})^T (y - X\widehat{\beta} - Z\widehat{b})}{\widehat{\sigma}^2} \\ &\overset{(2),(A.1)}{=} n\log(2\pi) + n\log(\widehat{\sigma}^2) + (n-p) - (n-p)\frac{y^T \widehat{P_*}^T \widehat{V_*}^{-1} Z\widehat{D_*} Z^T \widehat{V_*}^{-1} \widehat{P_*} y}{y^T \widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*} y} \\ &\overset{(A.1),(A.2)}{=} n\log(2\pi) + n\log\left(\frac{y^T \widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*} y}{n-p}\right) + (n-p) - \operatorname{tr}(\widehat{P_*} Z\widehat{D_*} Z^T \widehat{V_*}^{-1}) \\ &= n\log(2\pi) + n\log\left(\frac{y^T \widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*} y}{n-p}\right) + \operatorname{tr}(\widehat{P_*}^T \widehat{V_*}^{-1} \widehat{P_*}). \end{aligned}$$

For ML estimation, the result follows analogously using the profile log-likelihood (3). □

A.3 Proof of Theorem 2

We first consider REML estimation. If $\widehat{\lambda} = 0$, equality of the cAICs follows from that of the conditional log-likelihoods and the estimated degrees of freedom (definition (10) with estimated D_*).

Now suppose that $\widehat{\lambda} > 0$. The definition of a REML estimate gives us $\ell(\widehat{\lambda}) \geq \ell(0)$ for the restricted profile log-likelihood $\ell(\lambda) = f(A^T y|\widehat{\sigma}^2(\theta_*), \theta_*)$. The spectral representation of the restricted profile log-likelihood is (Crainiceanu and Ruppert, 2004)

$$2\ell(\lambda) = \text{const} - (n-p)\log\left(\widetilde{\sigma}^2\left\{\sum_{k=1}^{r}\frac{1+\widetilde{\lambda}\mu_{k,n}}{1+\lambda\mu_{k,n}}w_k^2 + \sum_{k=r+1}^{n-p} w_k^2\right\}\right) - \sum_{k=1}^{r}\log\left(1+\lambda\mu_{k,n}\right),$$

where $\widetilde{\sigma}^2$ and $\widetilde{\lambda}$ are the true values of σ^2 and λ, respectively, $\mu_{k,n}, k = 1, \ldots, r$, are the eigenvalues of $\Sigma^{1/2}Z^T(I_n - X(X^TX)^{-1}X^T)Z\Sigma^{1/2}$, and $w_1, \ldots, w_{n-p}$ are independent $N(0,1)$ variables. Thus,

$$\begin{aligned} & (n-p)\log\left(\widetilde{\sigma}^2\left\{\sum_{k=1}^{r}\frac{1+\widetilde{\lambda}\mu_{k,n}}{1+\widehat{\lambda}\mu_{k,n}}w_k^2 + \sum_{k=r+1}^{n-p} w_k^2\right\}\right) + \sum_{k=1}^{r}\log(1+\widehat{\lambda}\mu_{k,n}) \quad \text{(A.3)} \\ \leq \; & (n-p)\log\left(\widetilde{\sigma}^2\left\{\sum_{k=1}^{r}(1+\widetilde{\lambda}\mu_{k,n})w_k^2 + \sum_{k=r+1}^{n-p} w_k^2\right\}\right). \end{aligned}$$

From the representation of $P_*^T V_*^{-1} P_*$ in Crainiceanu and Ruppert (2004),

$$\begin{aligned} \operatorname{tr}(P_*^T V_*^{-1} P_*) &= \sum_{k=1}^{r}\frac{1}{1+\lambda\mu_{k,n}} + (n-p-r) \qquad \text{and} \\ y^T {P_*}^T {V_*}^{-1} P_* y &= \widetilde{\sigma}^2\left\{\sum_{k=1}^{r}\frac{1+\widetilde{\lambda}\mu_{k,n}}{1+\lambda\mu_{k,n}}w_k^2 + \sum_{k=r+1}^{n-p} w_k^2\right\}. \end{aligned}$$

For the effective degrees of freedom, we have (Liang et al., 2008, equation (5))

$$\rho(\lambda) = p + \sum_{k=1}^{r}\frac{\lambda\mu_{k,n}}{1+\lambda\mu_{k,n}}. \quad \text{(A.4)}$$

Note that $\log(x) + 1/x$ is a strictly monotonic increasing function for $x > 1$. As not all $\mu_{k,n}$ are zero, this gives us

$$-\sum_{k=1}^{r}\left\{\frac{1}{1+\widehat{\lambda}\mu_{k,n}} + \log\left(1+\widehat{\lambda}\mu_{k,n}\right)\right\} < -\sum_{k=1}^{r} 1. \quad \text{(A.5)}$$

Putting everything together, we obtain

$$
\begin{aligned}
cAIC(M_2) &\overset{La.1}{=} n\log(2\pi) + n\log\left(\frac{y^T\widehat{P_*}^T\widehat{V_*}^{-1}\widehat{P_*}y}{n-p}\right) + \mathrm{tr}(\widehat{P_*}^T\widehat{V_*}^{-1}\widehat{P_*}) + 2\big(\rho(\widehat{\lambda})+1\big) \\
&\overset{(A.4)}{=} n\log(2\pi) + n\log\left(\widetilde{\sigma}^2\left\{\sum_{k=1}^{r}\frac{1+\widetilde{\lambda}\mu_{k,n}}{1+\widehat{\lambda}\mu_{k,n}}w_k^2 + \sum_{k=r+1}^{n-p}w_k^2\right\}\right) \\
&\quad -n\log(n-p) - \sum_{k=1}^{r}\frac{1}{1+\widehat{\lambda}_R\mu_{k,n}} + (n+p+r) + 2 \\
&\overset{(A.3)}{\leq} n\log(2\pi) + n\log\left(\widetilde{\sigma}^2\left\{\sum_{k=1}^{r}(1+\widetilde{\lambda}\mu_{k,n})w_k^2 + \sum_{k=r+1}^{n-p}w_k^2\right\}\right) - n\log(n-p) \\
&\quad -\frac{n}{n-p}\sum_{k=1}^{r}\log(1+\widehat{\lambda}\mu_{k,n}) - \sum_{k=1}^{r}\frac{1}{1+\widehat{\lambda}\mu_{k,n}} + (n+p+r) + 2 \\
&\overset{(A.5)}{<} n\log(2\pi) + n\log\left(\widetilde{\sigma}^2\left\{\sum_{k=1}^{r}(1+\widetilde{\lambda}\mu_{k,n})w_k^2 + \sum_{k=r+1}^{n-p}w_k^2\right\}\right) \\
&\quad -n\log(n-p) - \sum_{k=1}^{r}1 + (n+p+r) + 2 \\
&= cAIC(M_1).
\end{aligned}
$$

As $\widehat{\lambda} > 0$ iff not $\widehat{\lambda} = 0$, this gives us altogether

$$
\widehat{\lambda} = 0 \;\Leftrightarrow\; cAIC(M_1) = cAIC(M_2) \qquad \text{and} \qquad \widehat{\lambda} > 0 \Leftrightarrow cAIC(M_1) > cAIC(M_2).
$$

For ML estimation, the result follows analogously using the spectral representation of the profile log-likelihood in Crainiceanu and Ruppert (2004),

$$
\text{const} - n\log\left(\widetilde{\sigma}^2\left\{\sum_{k=1}^{r}\frac{1+\widetilde{\lambda}\mu_{k,n}}{1+\lambda\mu_{k,n}}w_k^2 + \sum_{k=r+1}^{n-p}w_k^2\right\}\right) - \sum_{k=1}^{r}\log\Big(1+\lambda\xi_{k,n}\Big),
$$

where $\xi_{k,n}, k = 1,\ldots,r$, are the eigenvalues of $\Sigma^{1/2}Z^TZ\Sigma^{1/2}$. Note that

$$
\mathrm{tr}(V_*^{-1}) = \sum_{k=1}^{r}\frac{1}{1+\lambda\xi_{k,n}} + (n-r)
$$

and $\mu_{k,n} \leq \xi_{k,n}, k = 1,\ldots,r$, for the ordered eigenvalues, due to the positive-semidefiniteness of $\Sigma^{1/2}Z^TX(X^TX)^{-1}X^TZ\Sigma^{1/2}$ (Thompson and Freede, 1971). □

A.4 Remark 1

Results in Lemma 1 and Theorem 2 can also be generalised to more complex models. For example, the representation in Lemma 1 holds as well in the more general case where D is the block diagonal matrix $D = \mathrm{diag}(\tau_1^2\Sigma_1, \ldots, \tau_S^2\Sigma_S)$ with known $\Sigma_s, s = 1, \ldots, S$.

In this case, we can also show the following result. Denote

$$M_1 : y = X\beta + \varepsilon, \quad M_2 : y = X\beta + Zb + \varepsilon, \quad (b, \varepsilon) \sim \mathcal{N}(0, \mathrm{diag}(D, \sigma^2 I_n)).$$

Then,

$$\begin{array}{rcccl} \text{At least one } \widehat{\tau}_s^2 > 0, s = 1, \ldots, S & \Leftrightarrow & cAIC(M_1) & > & cAIC(M_2) \quad \text{and} \\ \widehat{\tau}_s^2 = 0, s = 1, \ldots, S & \Leftrightarrow & cAIC(M_1) & = & cAIC(M_2). \end{array}$$

The analogous result holds using REML estimation. The decision for inclusion or exclusion of a single variance parameter τ_S^2 is complicated by the potential change in the other variance estimates. We can derive simple sufficient conditions, however, for the proposition to carry over. For the case of REML estimation, for instance, consider the condition

$$\mu_{k,n}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_S) \geq \mu_{k,n}(\widehat{\widehat{\lambda}}_1, \ldots, \widehat{\widehat{\lambda}}_{S-1}, 0),$$

where $\mu_{k,n}(\lambda_1, \ldots, \lambda_S)$ are the eigenvalues of $D_*^{1/2} Z^T (I_n - X(X^TX)^{-1}X^T) Z D_*^{1/2}$ and double-hat notation indicates estimation under the constraint $\lambda_S = 0$ or $\tau_S^2 = 0$ (model M_3). This condition is fulfilled in particular if $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{S-1}) = (\widehat{\widehat{\lambda}}_1, \ldots, \widehat{\widehat{\lambda}}_{S-1})$, i.e. the estimates for the first $S-1$ variance components do not change with inclusion of λ_S in the model. Then,

$$\begin{array}{rcccl} \widehat{\tau}_S^2 > 0 & \Leftrightarrow & cAIC(M_3) & > & cAIC(M_2) \quad \text{and} \\ \widehat{\tau}_S^2 = 0 & \Leftrightarrow & cAIC(M_3) & = & cAIC(M_2). \end{array}$$

The case of general D is more involved due to the constraint that $\widehat{D}$ must be positive semidefinite. The geometry of the parameter space thus is more complex (Stram and Lee (1994), using results by Self and Liang (1987)), and is beyond the scope of this paper.

A.5 Proof of Theorem 3

We have

$$\begin{aligned}\widehat{y} &= X\widehat{\beta} + Z\widehat{b} = X\widehat{\beta} + Z\widehat{D}_* Z^T \widehat{V}_*^{-1}(y - X\widehat{\beta}) = X\widehat{\beta} + (I_n - \widehat{V}_*^{-1})(y - X\widehat{\beta}) \\ &= y - \widehat{V}_*^{-1} y + \widehat{V}_*^{-1} X\widehat{\beta} = y - \widehat{V}_*^{-1}\widehat{P}_* y,\end{aligned}$$

as $V_* = I_n + ZD_*Z^T$, and therefore $ZD_*Z^TV_*^{-1} = I_n - V_*^{-1}$. Thus,

$$\Phi_0 = \operatorname{tr}\Bigg(\underbrace{\frac{\partial \widehat{y}}{\partial y}}_{n\times n}\Bigg) = \operatorname{tr}\left(I_n - \widehat{V}_*^{-1}\widehat{P}_* - \sum_{j=1}^{q} \underbrace{\frac{\partial}{\partial \theta_{*,j}}\left[\widehat{V}_*^{-1}\widehat{P}_*\right]}_{n\times n} \underbrace{y}_{n\times 1} \underbrace{\left[\frac{d}{dy}\widehat{\theta}_{*,j}(y)\right]}_{1\times n}\right).$$

It is

$$\frac{\partial}{\partial \theta_{*,j}}\left[V_*^{-1}P_*\right] = -P_*^T V_*^{-1} W_{*,j} V_*^{-1} P_* = -A_* W_{*,j} A_*, \quad j = 1, \ldots, q.$$

Let $\widetilde{\theta}_*(y)$ be the maximiser of the (restricted) log-likelihood over $\mathbb{R}^q$. Thus, $\widehat{\theta}_*(y)$ is the projection of $\widetilde{\theta}_*(y)$ onto Θ. Then, $\widetilde{\theta}_{*,j}(y) \neq 0$, $j = s+1, \ldots, s+t$, with probability one, as the maximum over the larger parameter space lies on the boundary with probability zero. Therefore, there exists an ε-ball around $(\widetilde{\theta}_{*,s+1}(y), \ldots, \widetilde{\theta}_{*,s+t})(y)$ consisting of points that have 0 as their projection onto $[0,\infty)^t$. Let e_i be the unit vector for component i. As $\widetilde{\theta}_*(y)$ is a continuous function in y, there exists an $\delta > 0$, such that the projection of $(\widetilde{\theta}_{*,s+1}(y + he_i), \ldots, \widetilde{\theta}_{*,s+t}(y + he_i))$ onto $[0,\infty)^t$ is 0 for all $|h| < \delta, j = s+1, \ldots, s+t, i = 1, \ldots, n$. Thus,

$$\frac{\partial}{\partial y_i}\widehat{\theta}_{*,j} = \lim_{h\to 0} \frac{\widehat{\theta}_{*,j}(y + he_i) - \widehat{\theta}_{*,j}(y)}{h} = 0, \quad i = 1, \ldots, n, j = s+1, \ldots, q,$$

and

$$\Phi_0 = n - \operatorname{tr}(\widehat{A}_*) + \sum_{j=1}^{s}\left[\frac{d}{dy}\widehat{\theta}_{*,j}(y)\right]\widehat{A}_*\widehat{W}_{*,j}\widehat{A}_* y.$$

Now, consider first restricted maximum likelihood estimation. Twice the restricted profile log-likelihood for θ_* is given by

$$-\log\{\det(V_*)\} - \log\{\det(X^TV_*^{-1}X)\} - (n-p)\log\{(y - X\widehat{\beta})^T V_*^{-1}(y - X\widehat{\beta})\}.$$

Using the score equation, and as $(\theta_{*,1}, \ldots, \theta_{*,s})$ is in the interior of Θ_s, the restricted MLE of θ_* fulfills

$$0 \equiv h_j(\widehat{\theta}_*(y), y) := \operatorname{tr}(\widehat{P}_*\widehat{W}_{*,j}\widehat{V}_*^{-1}) - (n-p)\frac{y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{W}_{*,j}\widehat{V}_*^{-1}\widehat{P}_* y}{y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{P}_* y}, \quad j = 1, \ldots, s.$$

Consequently,

$$0 \equiv \underbrace{\frac{d}{dy}h_j(\widehat{\theta}_*(y),y)}_{1\times n} = \sum_{l=1}^{s} \underbrace{\frac{\partial}{\partial\theta_{*,l}}h_j(\widehat{\theta}_*(y),y)}_{1\times 1}\underbrace{\frac{d}{dy}\widehat{\theta}_{*,l}(y)}_{1\times n} + \underbrace{\frac{\partial}{\partial y}h_j(\widehat{\theta}_*(y),y)}_{1\times n}, \quad j=1,\ldots,s$$

$$\Rightarrow \underbrace{\frac{d}{dy}\widehat{\theta}_s(y)}_{s\times n} = -\underbrace{\left[\frac{\partial}{\partial\theta_{*,l}}h_j(\widehat{\theta}_*(y),y)\right]^{-1}_{j,l=1,\ldots,s}}_{s\times s}\underbrace{\frac{\partial}{\partial y}h(\widehat{\theta}_*(y),y)}_{s\times n},$$

where $\frac{\partial}{\partial y}h(\widehat{\theta}_*(y),y)$ includes $1\times n$ rows $\frac{\partial}{\partial y}h_j(\widehat{\theta}_*(y),y), j=1,\ldots,s$.

Note that $\left[\frac{\partial}{\partial\theta_{*,l}}h_j(\widehat{\theta}_*(y),y)\right]_{j,l=1,\ldots,s}$ is negative definite (and thus invertible) with probability one as the Hessian in the first s components of the profile restricted log-likelihood evaluated at $\widehat{\theta}_*(y)$. We have

$$\begin{aligned}
-\frac{\partial}{\partial y}h_j(\widehat{\theta}_*(y),y) &= \frac{2(n-p)}{(y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{P}_*y)^2}\{y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{W}_{*,j}\widehat{V}_*^{-1}\widehat{P}_*y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{P}_*y \\
&\qquad -y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{W}_{*,j}\widehat{V}_*^{-1}\widehat{P}_*yy^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{P}_*\} \\
&= \frac{2(n-p)}{(y^T\widehat{A}_*y)^2}\{y^T\widehat{A}_*\widehat{W}_{*,j}\widehat{A}_*y^T\widehat{A}_*y - y^T\widehat{A}_*\widehat{W}_{*,j}\widehat{A}_*yy^T\widehat{A}_*\}, \quad j=1,\ldots,s,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\theta_{*,l}}h_j(\widehat{\theta}_*(y),y) &= \mathrm{tr}\{\widehat{V}_*^{-1}\widehat{P}_*\widehat{U}_{*,jl} - \widehat{V}_*^{-1}\widehat{P}_*\widehat{W}_{*,l}\widehat{V}_*^{-1}\widehat{P}_*\widehat{W}_{*,j}\} \qquad \text{(A.6)} \\
&\quad -(n-p)\frac{y^T(\widehat{V}_*^{-1}\widehat{P}_*\widehat{U}_{*,jl}\widehat{V}_*^{-1}\widehat{P}_* - \widehat{V}_*^{-1}\widehat{P}_*\widehat{W}_{*,l}\widehat{V}_*^{-1}\widehat{P}_*\widehat{W}_{*,j}\widehat{V}_*^{-1}\widehat{P}_*)y}{y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{P}_*y} \\
&\quad +(n-p)\frac{y^T(\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{W}_{*,j}\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{W}_{*,l}\widehat{V}_*^{-1}\widehat{P}_*)y}{y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{P}_*y} \\
&\quad -(n-p)\frac{y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{W}_{*,j}\widehat{V}_*^{-1}\widehat{P}_*yy^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{W}_{*,l}\widehat{V}_*^{-1}\widehat{P}_*y}{(y^T\widehat{P}_*^T\widehat{V}_*^{-1}\widehat{P}_*y)^2} \\
&= \mathrm{tr}\{\widehat{A}_*\widehat{U}_{*,jl} - \widehat{A}_*\widehat{W}_{*,l}\widehat{A}_*\widehat{W}_{*,j}\} - (n-p)\frac{y^T\widehat{A}_*\widehat{W}_{*,j}\widehat{A}_*yy^T\widehat{A}_*\widehat{W}_{*,l}\widehat{A}_*y}{(y^T\widehat{A}_*y)^2} \\
&\quad -(n-p)\frac{y^T(\widehat{A}_*\widehat{U}_{*,jl}\widehat{A}_* - 2\widehat{A}_*\widehat{W}_{*,l}\widehat{A}_*\widehat{W}_{*,j}\widehat{A}_*)y}{y^T\widehat{A}_*y} \quad j,l=1,\ldots,s.
\end{aligned}$$

For maximum likelihood estimation, using twice the profile log-likelihood for θ_*,

$$-\log\{\det(V_*)\} - n\log\{(y-X\widehat{\beta})^TV_*^{-1}(y-X\widehat{\beta})\},$$

the derivation follows analogoulsy, with every $(n-p)$ replaced by n, and $\mathrm{tr}\{\widehat{A}_*\widehat{U}_{*,jl} - \widehat{A}_*\widehat{W}_{*,l}\widehat{A}_*\widehat{W}_{*,j}\}$ replaced by $\mathrm{tr}\{\widehat{U}_{*,jl}\widehat{V}_*^{-1} - \widehat{W}_{*,j}\widehat{V}_*^{-1}\widehat{W}_{*,l}\widehat{V}_*^{-1}\}$.

Putting everything together, we obtain $\Phi_0 = n - \mathrm{tr}(\widehat{A}_*) + \sum_{j=1}^{s} e_j^T\widehat{B}_*^{-1}\widehat{G}_*\widehat{A}_*\widehat{W}_{*,j}\widehat{A}_*y$. □

B Simulations

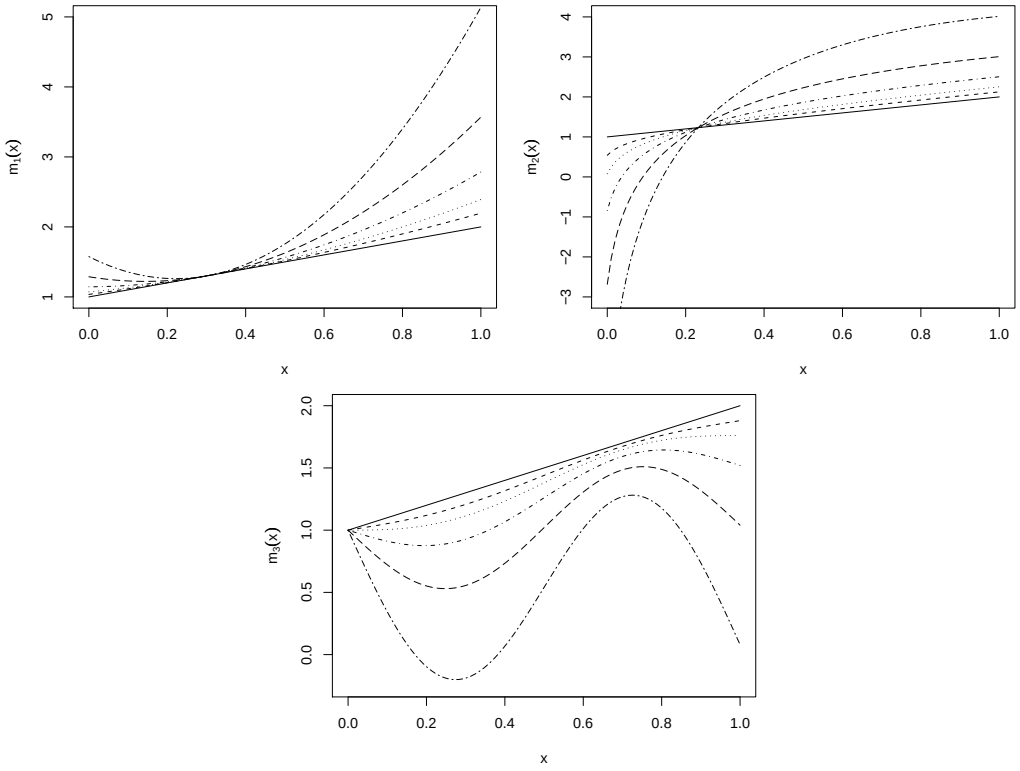


Figure 1: Functions estimated nonparametrically in the simulation study on penalised spline smoothing for varying values of the non-linearity parameter d.

R-code used in the simulations can be found in supplement_RE.R for the random intercept (ANOVA) model, and in supplement_splines.R for the penalised spline model. R-functions used in both cases are in `fcts.R`.

The complete simulation results are presented in the next two subsections. Functions $m(\cdot)$ used in the simulations in Section 5.1 are depicted in Figure 1.

B.1 Penalised Spline Smoothing

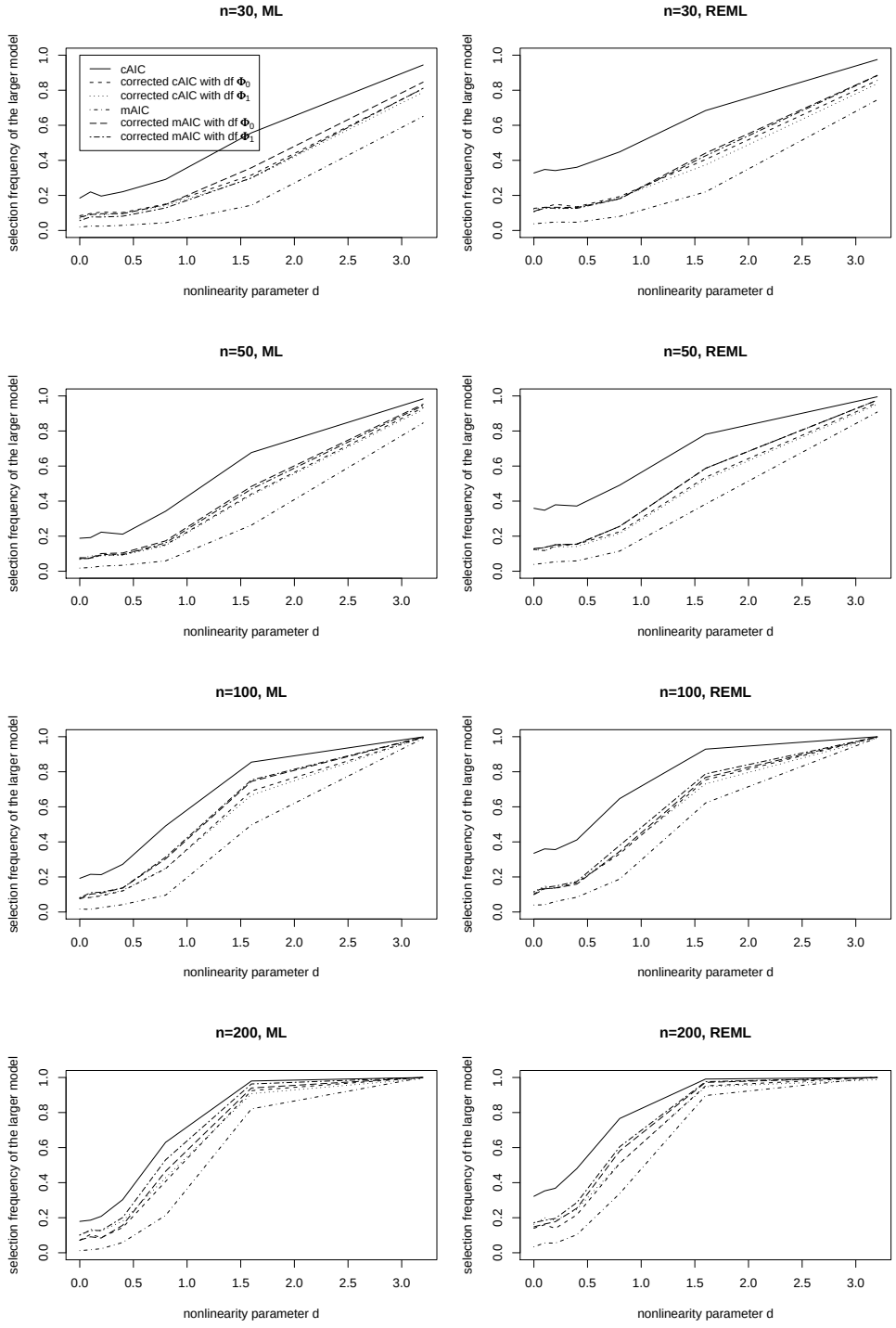


Figure 2: Proportion of simulation replications where the more complex, non-linear model was favored by the AIC for function $m_1(x)$.

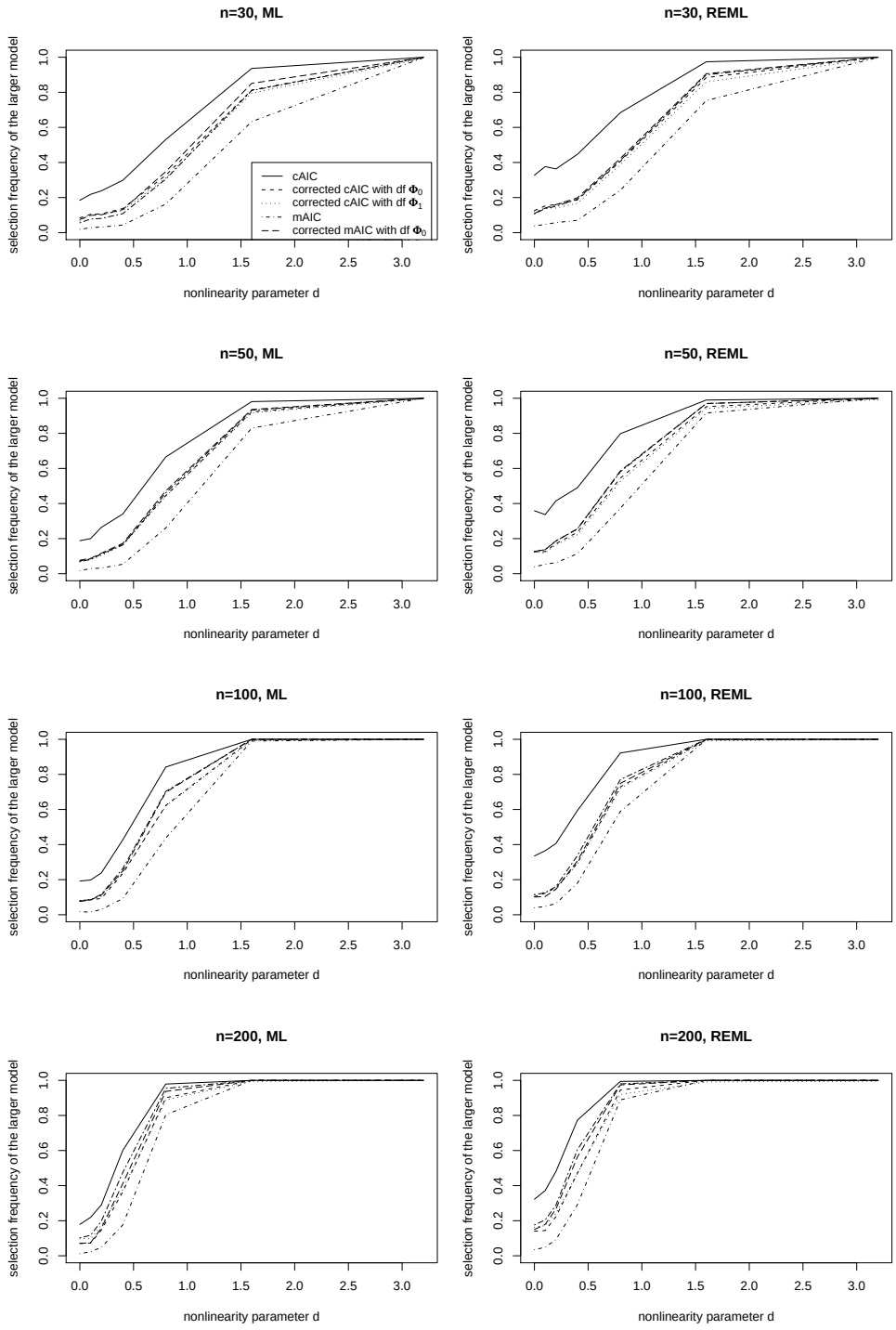


Figure 3: Proportion of simulation replications where the more complex, non-linear model was favored by the AIC for function $m_2(x)$.

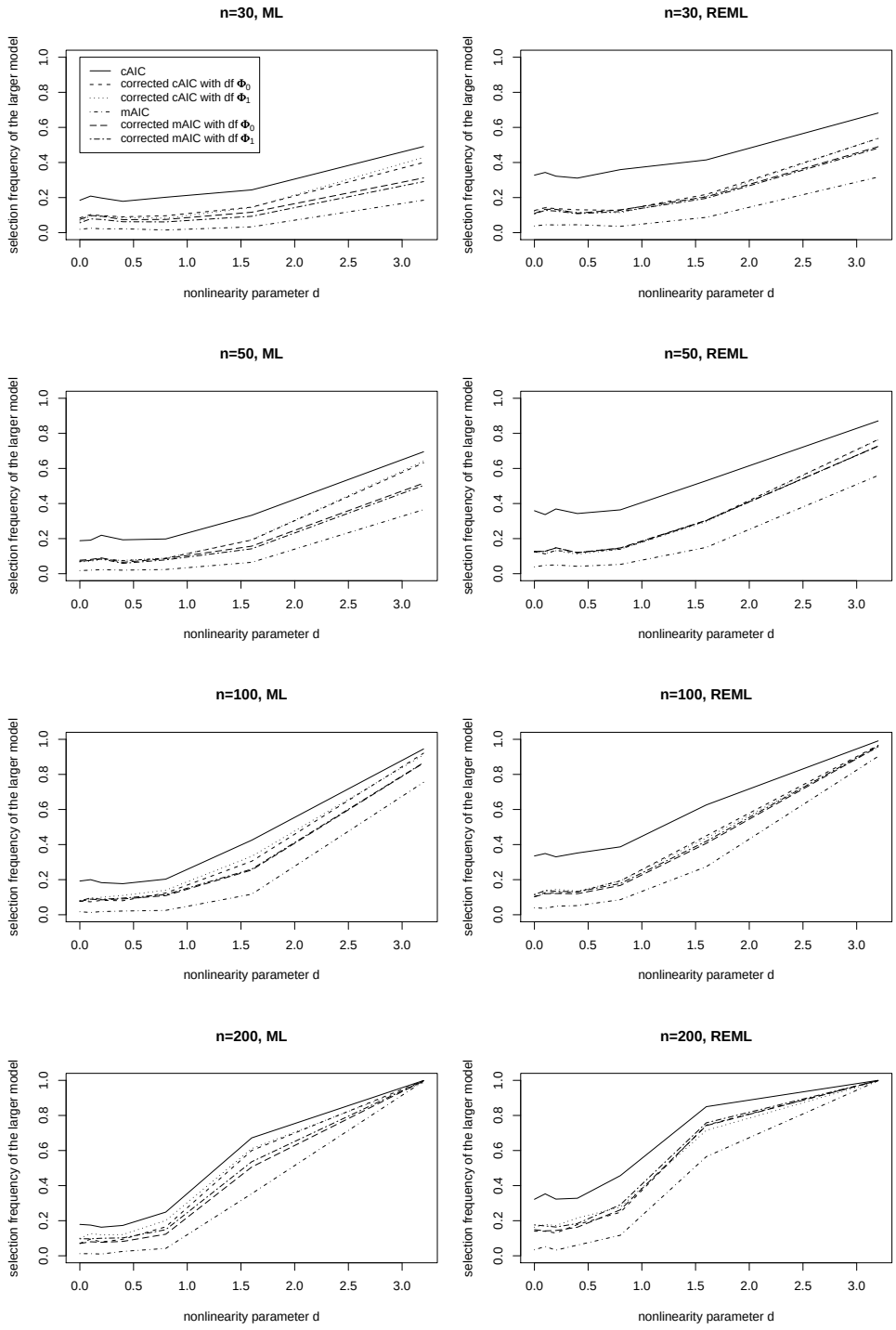


Figure 4: Proportion of simulation replications where the more complex, non-linear model was favored by the AIC for function $m_3(x)$.

B.2 Random Intercept Model

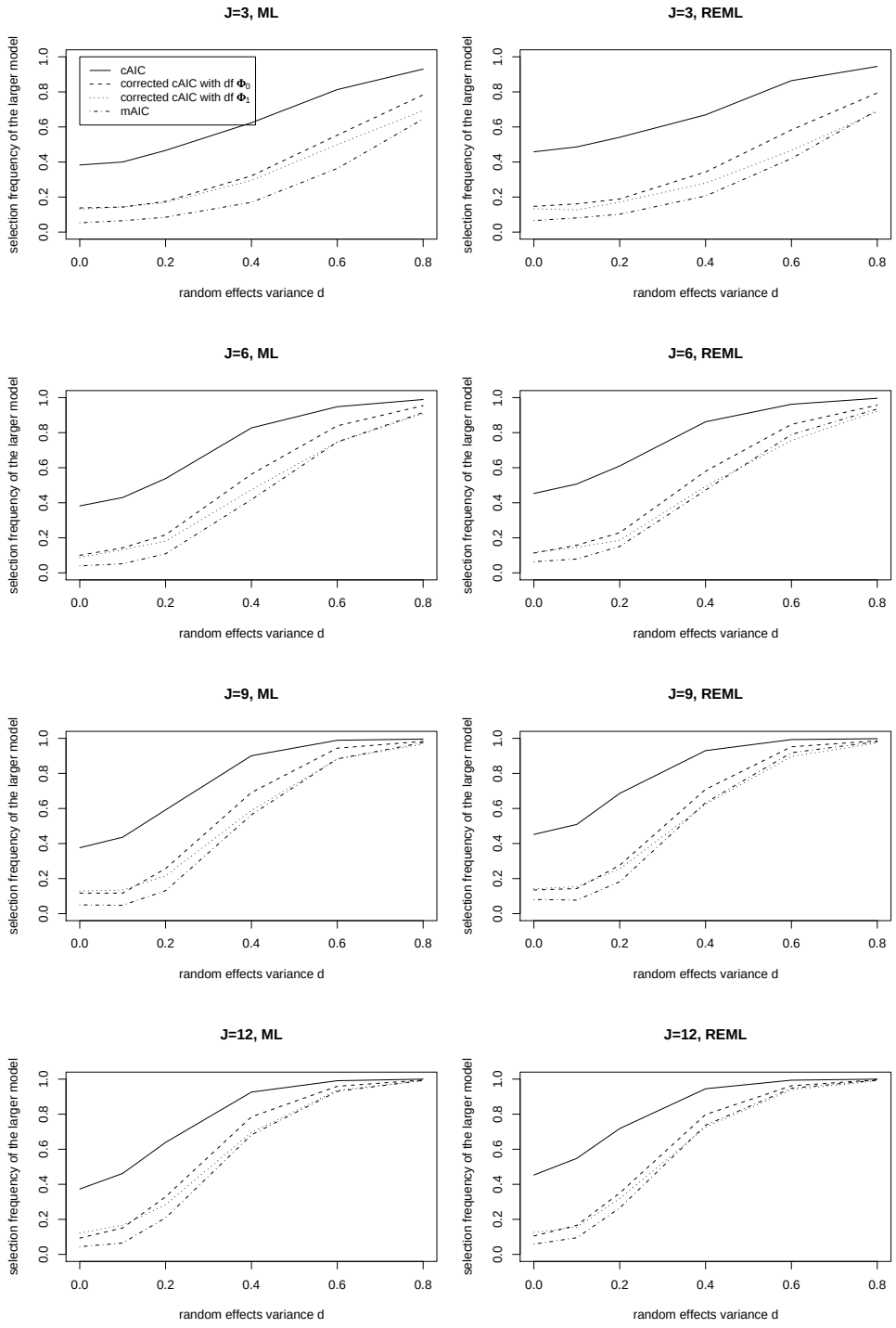


Figure 5: Proportion of simulation replications where the more complex random intercept model was favored by the AIC in the case of ten clusters.

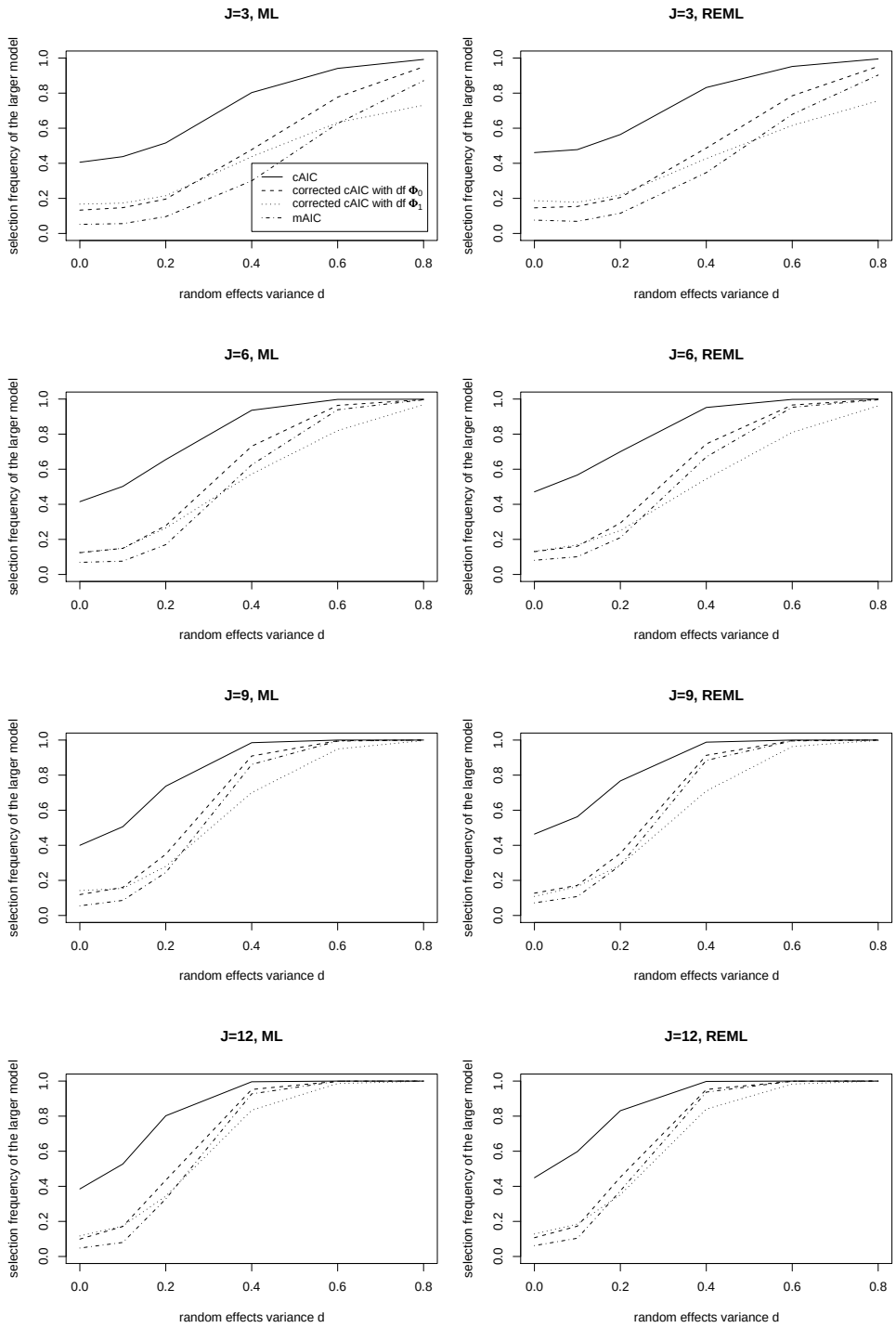


Figure 6: Proportion of simulation replications where the more complex random intercept model was favored by the AIC in the case of twenty clusters.

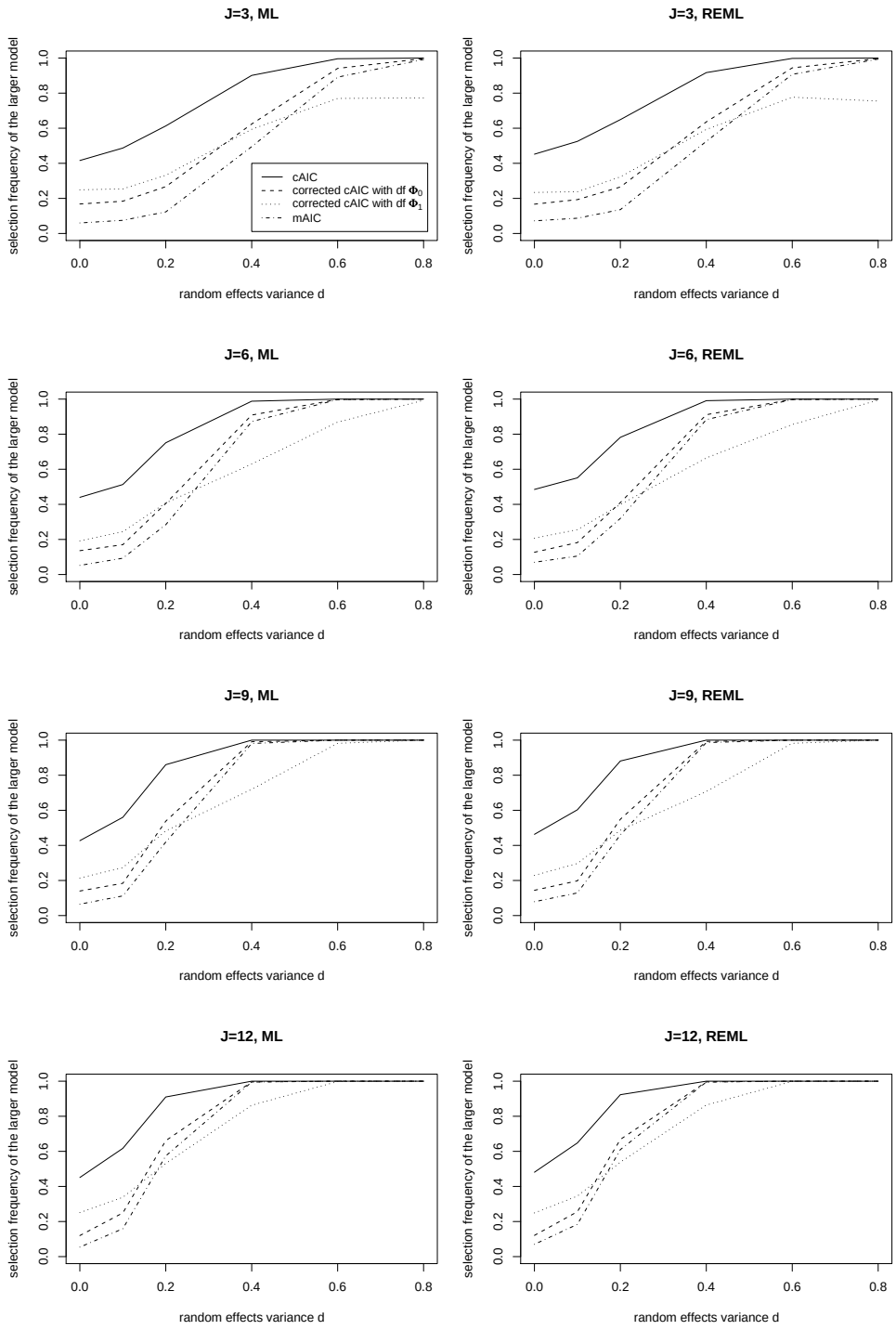


Figure 7: Proportion of simulation replications where the more complex random intercept model was favored by the AIC in the case of forty clusters.

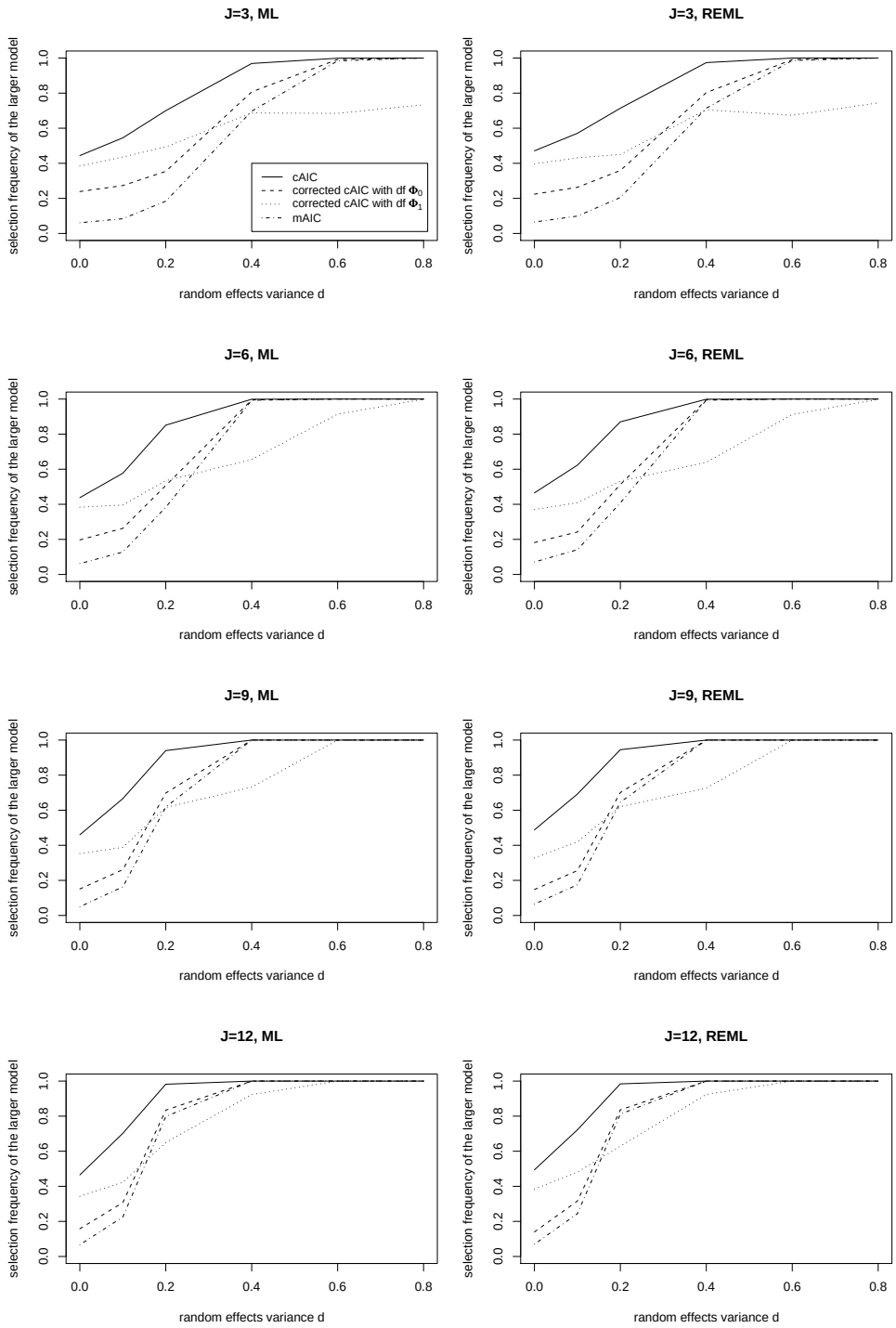


Figure 8: Proportion of simulation replications where the more complex random intercept model was favored by the AIC in the case of eighty clusters.

C Childhood Malnutrition in Zambia

Table 1: Explanatory variables in the Zambia data set

variable	description
$csex$	gender of the child (1 = male, 0 = female)
$cfeed$	duration of breastfeeding (in months)
$cage$	age of the child (in months)
$mage$	age of the mother (at birth, in years)
$mheight$	height of the mother (in cm)
$mbmi$	body mass index of the mother
$medu$	education of the mother (1 = no education, 2 = primary school, 3 = elementary school, 4 = higher)
$mwork$	employment status of the mother (1 = employed, 0 = unemployed)
$district$	residential district (54 districts in total)

C.1 Univariate Smoothing

	$cAIC$				$mAIC$			
	ML		$REML$		ML		$REML$	
	M_1	M_2	M_1	M_2	M_1	M_2	M_1	M_2
cfeed	4467.56	**4353.42**	4467.56	**4353.35**	4467.56	**4387.41**	4474.52	**4386.51**
cage	4429.76	**4344.23**	4429.77	**4344.21**	4429.76	**4354.52**	4437.32	**4358.94**
mage	4538.55	**4535.61**	4538.56	**4535.48**	**4538.55**	4539.73	4545.22	**4544.9**
mheight	**4444.39**	**4444.39**	4444.39	**4443.85**	**4444.39**	4446.39	**4450.14**	4451.78
mbmi	**4519.6**	4519.6	**4519.6**	4519.6	**4519.6**	4521.6	**4525.39**	4527.39

Table 2: cAIC and mAIC for non-linear (H_1) and linear (H_0) modelling of single continuous covariate effects in the Zambia data.

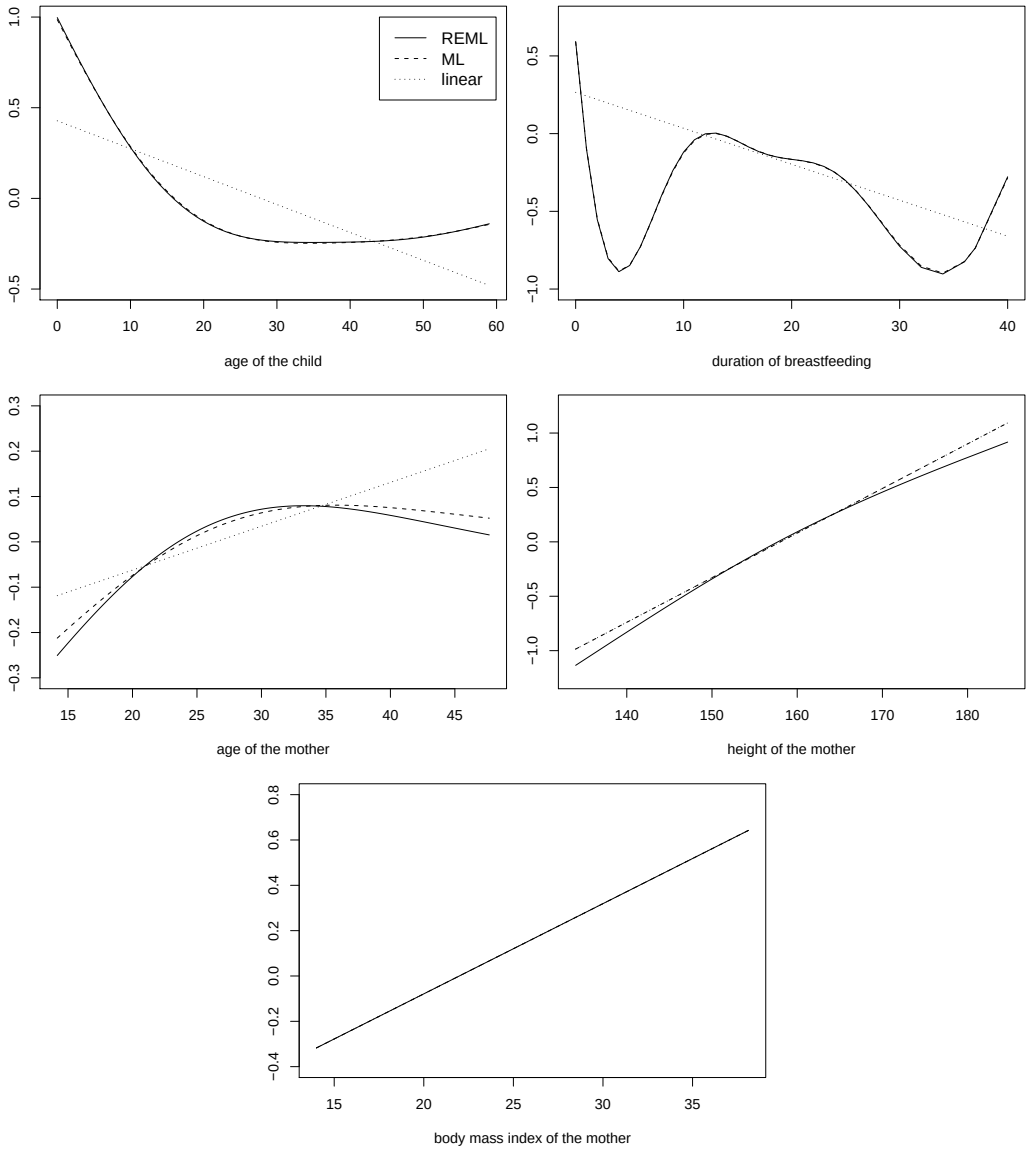


Figure 9: Estimated linear and non-linear effects obtained with ML and REML estimation in the univariate smoothing problem.

C.2 Additive Mixed Model

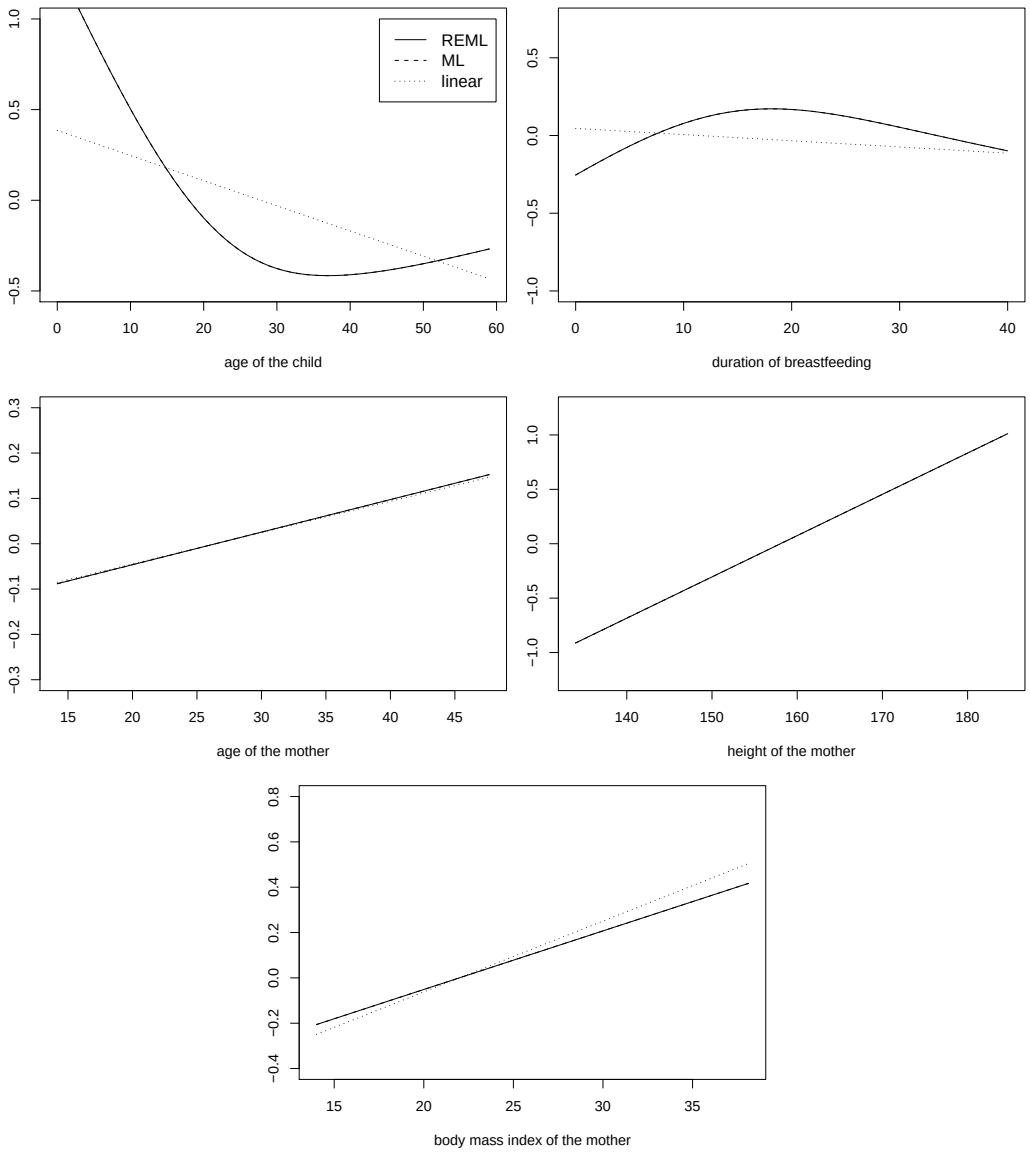
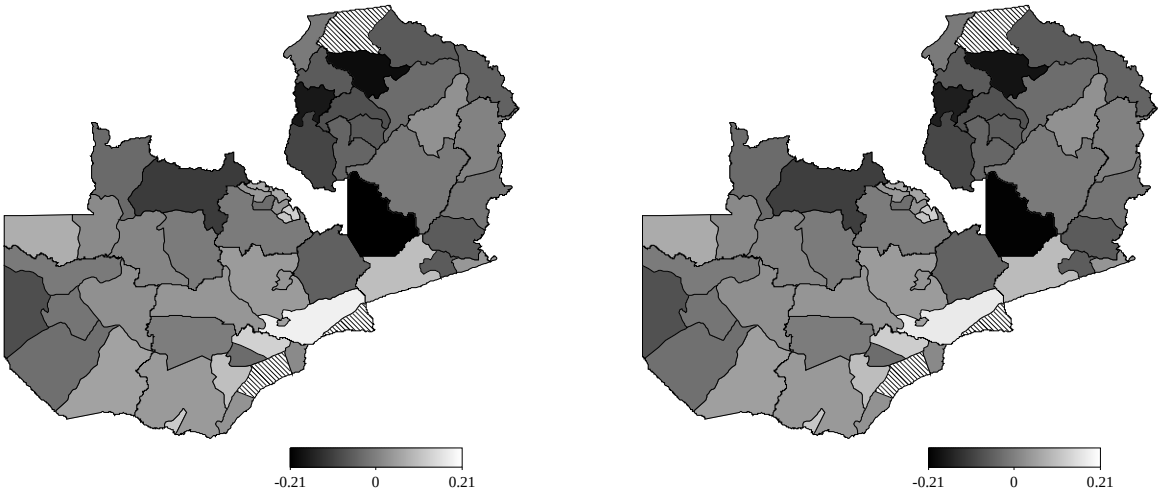


Figure 10: Estimated non-linear effects in the full model 64 obtained with ML and REML, and estimated linear effects in the simplest model 1 (only linear effects, no district-specific random intercepts).

							ML		$REML$	
	cfeed	cage	mage	mheight	mbmi	district	$cAIC$	$mAIC$	$cAIC$	$mAIC$
1	−	−	−	−	−	−	4261.258	4261.258	4261.296	4289.287
2	−	−	−	−	−	+	4249.393	4258.190	4249.374	4285.460
3	+	−	−	−	−	−	4178.298	4211.837	4178.289	4232.128
4	+	−	−	−	−	+	4168.932	4210.120	4168.818	4229.724
5	−	+	−	−	−	−	4148.080	4158.453	4148.260	4183.692
6	−	+	−	−	−	+	4134.518	4154.564	4134.644	4179.004
7	−	−	+	−	−	−	4261.258	4263.258	4261.296	4291.287
8	−	−	+	−	−	+	4249.393	4260.190	4249.374	4287.460
9	−	−	−	+	−	−	4261.258	4263.258	4261.116	4291.179
10	−	−	−	+	−	+	4249.394	4260.190	4249.191	4287.346
11	−	−	−	−	+	−	4261.258	4263.258	4261.296	4291.287
12	−	−	−	−	+	+	4249.393	4260.190	4249.374	4287.460
13	+	+	−	−	−	−	4137.779	4154.146	4137.999	4177.238
14	+	+	−	−	−	+	**4125.784**	**4151.102**	4125.942	**4173.468**
15	+	−	+	−	−	−	4178.299	4213.837	4178.289	4234.128
16	+	−	+	−	−	+	4168.932	4212.120	4168.818	4231.724
17	+	−	−	+	−	−	4178.298	4213.837	4178.236	4234.101
18	+	−	−	+	−	+	4168.932	4212.120	4168.818	4231.724
19	+	−	−	−	+	−	4178.299	4213.837	4178.289	4234.128
20	+	−	−	−	+	+	4168.932	4212.120	4168.821	4231.725
21	−	+	+	−	−	−	4148.080	4160.453	4148.260	4185.692
22	−	+	+	−	−	+	4134.518	4156.564	4134.644	4181.004
23	−	+	−	+	−	−	4148.080	4160.453	4148.149	4185.623
24	−	+	−	+	−	+	4134.518	4156.564	4134.546	4180.939
25	−	+	−	−	+	−	4148.080	4160.453	4148.260	4185.692
26	−	+	−	−	+	+	4134.518	4156.564	4134.644	4181.004
27	−	−	+	+	−	−	4261.258	4265.258	4261.116	4293.179
28	−	−	+	+	−	+	4249.393	4262.190	4249.192	4289.346
29	−	−	+	−	+	−	4261.258	4265.258	4261.296	4293.287
30	−	−	+	−	+	+	4249.393	4262.190	4249.374	4289.460
31	−	−	−	+	+	−	4261.258	4265.258	4261.116	4293.179
32	−	−	−	+	+	+	4249.394	4262.190	4249.191	4289.346
33	+	+	+	−	−	−	4137.779	4156.146	4137.999	4179.238
34	+	+	+	−	−	+	**4125.784**	4153.102	4125.942	4175.468
35	+	+	−	+	−	−	4137.779	4156.146	4137.776	4179.109
36	+	+	−	+	−	+	**4125.784**	4153.102	**4125.753**	4175.352
37	+	+	−	−	+	−	4137.779	4156.146	4137.999	4179.238
38	+	+	−	−	+	+	**4125.784**	4153.102	4125.942	4175.468
39	+	−	+	+	−	−	4178.299	4215.837	4178.289	4236.128
40	+	−	+	+	−	+	4168.933	4214.122	4168.818	4233.724
41	+	−	+	−	+	−	4178.299	4215.837	4178.289	4236.128
42	+	−	+	−	+	+	4168.932	4214.120	4168.818	4233.724
43	+	−	−	+	+	−	4178.298	4215.837	4178.290	4236.128
44	+	−	−	+	+	+	4168.932	4214.122	4168.820	4233.725
45	−	+	+	+	−	−	4148.080	4162.453	4148.149	4187.623
46	−	+	+	+	−	+	4134.518	4158.564	4134.546	4182.939
47	−	+	+	−	+	−	4148.080	4162.453	4148.260	4187.692
48	−	+	+	−	+	+	4134.518	4158.564	4134.644	4183.004
49	−	+	−	+	+	−	4148.080	4162.453	4148.149	4187.623
50	−	+	−	+	+	+	4134.518	4158.564	4134.545	4182.939
51	−	−	+	+	+	−	4261.258	4267.258	4261.116	4295.179
52	−	−	+	+	+	+	4249.393	4264.190	4249.191	4291.346
53	+	+	+	+	−	−	4137.779	4158.146	4137.777	4181.109
54	+	+	+	+	−	+	**4125.784**	4155.102	**4125.753**	4177.352
55	+	+	+	−	+	−	4137.779	4158.146	4137.999	4181.238
56	+	+	+	−	+	+	**4125.784**	4155.102	4125.784	4177.720
57	+	+	−	+	+	−	4137.779	4158.146	4137.776	4181.109
58	+	+	−	+	+	+	**4125.784**	4155.102	**4125.753**	4177.352
59	+	−	+	+	+	−	4178.298	4217.837	4178.289	4238.127
60	+	−	+	+	+	+	4168.932	4216.120	4168.817	4235.724
61	−	+	+	+	+	−	4148.080	4164.453	4148.149	4189.623
62	−	+	+	+	+	+	4134.517	4160.564	4134.546	4184.939
63	+	+	+	+	+	−	4137.779	4160.146	4137.776	4183.109
64	+	+	+	+	+	+	**4125.784**	4157.102	**4125.753**	4179.352

Table 3: Conditional and marginal AIC for various specifications of additive mixed models. The first column contains a model identification number, the following six columns indicate non-linear (+) versus linear (−) modelling of continuous covariate effects and presence (+) versus absence (−) of a district-specific random effect. $cAIC$ denotes the conventional cAIC, and $cAIC_c$ the corrected cAIC. In each column, the models with minimal AIC are bolded.

Figure 11: Estimated district-specific random intercepts in the full model 64 obtained with REML (left) and ML (right). Striped regions did not contain any observations.



References

Crainiceanu, C. and D. Ruppert (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B 66*(1), 165–185.

Crainiceanu, C., D. Ruppert, and T. Vogelsang (2003). Some properties of likelihood ratio tests in linear mixed models. Technical report, Department of Statistical Science, Cornell University.
`http://legacy.orie.cornell.edu/~ddavidr/papers/zeroprob_rev01.pdf`.

Giampaoli, V. and J. Singer (2009). Likelihood ratio tests for variance components in linear mixed models. *Journal of Statistical Planning and Inference 139*(4), 1435–1448.

Liang, H., H. Wu, and G. Zou (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika 95*, 773–778.

Self, S. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association 82*(398), 605–610.

Stram, D. and J.-W. Lee (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics 50*(3), 1171–1177.

Thompson, R. and L. Freede (1971). On the eigenvalues of sums of Hermitian matrices. *Linear Algebra and its Applications 4*(4), 369–376.