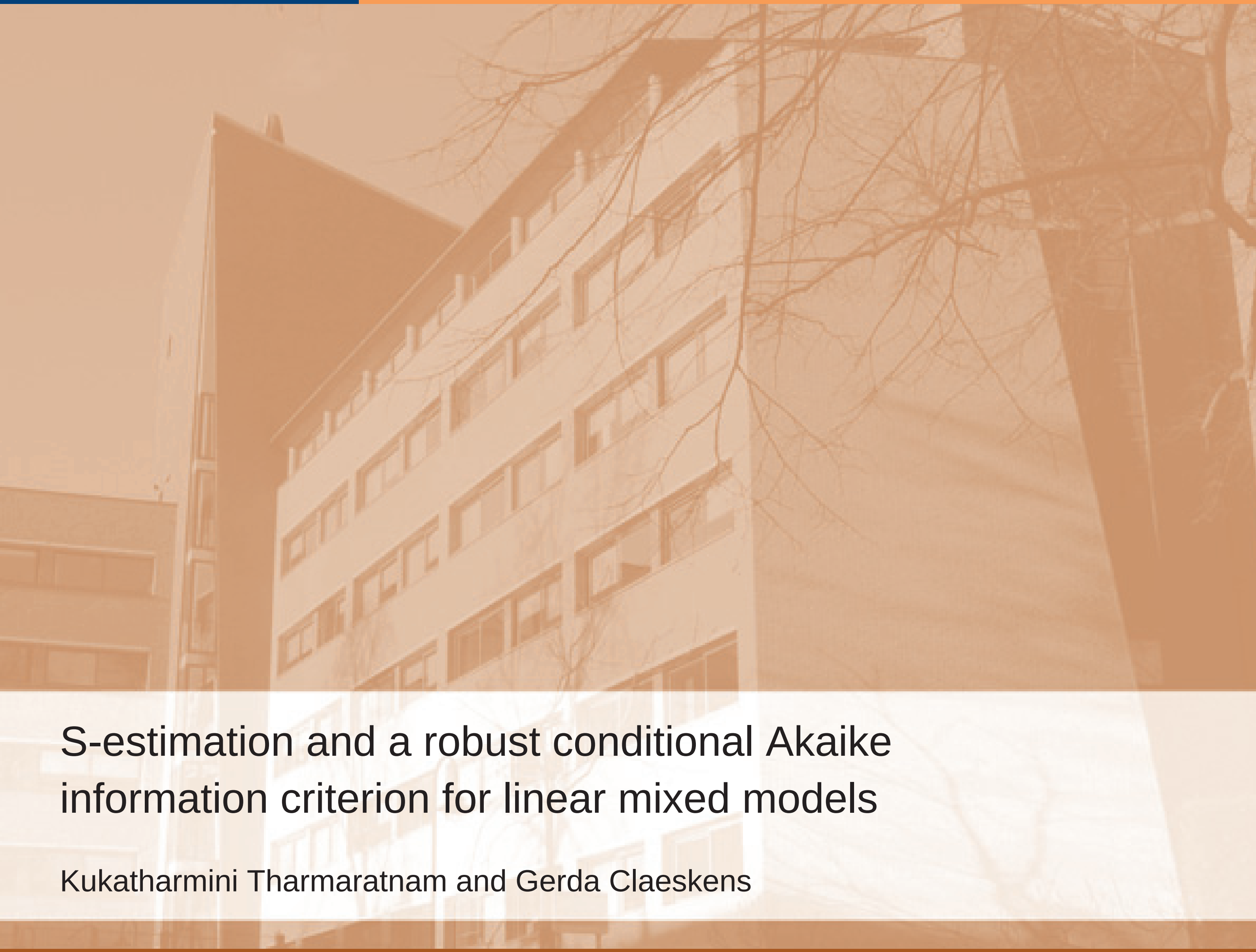
KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics



S-estimation and a robust conditional Akaike information criterion for linear mixed models

Kukatharmini Tharmaratnam and Gerda Claeskens

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1130

S-estimation and a robust conditional Akaike information criterion for linear mixed models

Kukatharmini Tharmaratnam and Gerda Claeskens

OR & Business Statistics and Leuven Statistics Research Center

K.U.Leuven, 3000 Leuven, Belgium

Kukatharmini.Tharmaratnam@econ.kuleuven.be

Gerda.Claeskens@econ.kuleuven.be

August 24, 2011

Abstract

We study estimation and model selection on both the fixed and the random effects in the setting of linear mixed models using outlier robust S-estimators. Robustness aspects on the level of the random effects as well as on the error terms is taken into account. The derived marginal and conditional information criteria are in the style of Akaike's information criterion but avoid the use of a fully specified likelihood by a suitable S-estimation approach that minimizes a scale function. We derive the appropriate penalty terms and provide an implementation using R. The setting of semiparametric additive models fit with penalized regression splines, in a mixed models formulation, fits as a specific application. Simulated data examples illustrate the effectiveness of the proposed criteria.

Some key words: Akaike information criterion; Conditional likelihood; Effective degrees of freedom; Mixed model; Penalized regression spline; S-estimation.

1 Introduction

We consider mixed linear models of the form $Y = X\beta + Zu + \varepsilon$, where u and ε are independent random variables, not necessarily normally distributed. Outlying values may be present in either u or ε. Variable selection in mixed linear models by means of the Akaike information criterion (AIC, Akaike, 1973) which is defined as minus twice the value of the maximized log-likelihood of the model plus twice the number of estimated parameters in the model, may be performed using the marginal log-likelihood of Y. Vaida and Blanchard (2005) have shown that in linear mixed models the resulting marginal AIC is not appropriate for variable selection when the random effects are of interest. They proposed the conditional Akaike information which uses the conditional likelihood of the response Y given the random effects u. The penalty term in the conditional AIC is related to the effective degrees of freedom of a linear mixed model (Hodges and Sargent, 2001). Liang et al. (2008) have proposed a corrected conditional AIC that accounts for the estimation of the variance components. Greven and Kneib (2010) study the theoretical properties of both the marginal and the conditional corrected AIC for the selection of variables in linear mixed models, and they provide a computationally feasible penalty term. Bondell et al. (2010) proposed a joint variable selection for fixed and random effects in linear mixed-effects models. All of the mentioned papers use maximum likelihood or restricted maximum likelihood for estimation.

In this paper we derive a marginal and conditional AIC for linear mixed models that no longer requires likelihood based estimation methods. In particular, we work with robust S-estimators that can accommodate the presence of outliers in (i) the response values, (ii) the random effects. We derive an expression for the penalty term that explicitly takes the estimation of the variance components into account and that can be computed in a straightforward way.

2 S-estimation in linear mixed models

We model the vector of observations for the ith subject, $i = 1, \ldots, n$, as

$$Y_i = X_i\beta + \sum_{j=1}^{r} Z_{ij}u_{ij} + \varepsilon_i, \tag{1}$$

where Y_i has length m_i, X_i is a $m_i \times p$ design matrix of fixed effects, Z_{ij} is a $m_i \times q_j$ design matrix for the random effects. The p-vector β is fixed but unknown, while the q_j-vectors u_{ij} are random with mean zero and variance matrix G_j. The random error ε_i has mean zero, and its variance matrix is denoted by R_i. The total number of observations is equal to $N = \sum_{i=1}^{n} m_i$, resulting in vectors Y and ε of length N, a $N \times p$ design matrix $X = (X_1, \ldots, X_n)^t$ for the fixed effects, a $m_i \times q$ design matrix $Z_i = (Z_{i1}, \ldots, Z_{ir})$ for the random effects, $u_i = (u_{i1}^t, \ldots, u_{ir}^t)^t$ is a $q \times 1$ vector. We denote $Z = \text{diag}(Z_1, \ldots, Z_n)$, Z is a $N \times nq$ vector, $u = (u_1^t, \ldots, u_n^t)^t$ is a $nq \times 1$ vector, $G_i = \text{diag}(G_1, \ldots, G_r)$, $G = \text{diag}(G_1, \ldots, G_n)$, and let $q = \sum_{j=1}^{r} q_j$. We assume that the set of random effects $\{u_{ij}; i = 1, \ldots, n, j = 1, \ldots, r\}$ and the set of error terms $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent, that $\text{Var}(u_{ij}) = G_j = \sigma_j^2 I_{q_j}$ and that $\text{Var}(\varepsilon) = R = \sigma_0^2 I_N$, with I_N the identity matrix with N rows. We define $R_i = \sigma_0^2 I_{m_i}$ and $V = \text{Var}(Y) = R + ZGZ^t$. In the balanced case where all $m_i = m$, we define the $m \times m$ matrices $\text{Var}(Y_i) = V_0$, and $\text{Var}(\varepsilon_i) = R_0 = \sigma_0^2 I_m$, for $i = 1, \ldots, n$, $j = 1, \ldots, r$.

The most frequent assumption in linear mixed models is that both the errors ε and the random effects u have Gaussian distributions. Outliers, extreme observations that are unlike most of the other observations in the sample, may occur for any of the observed random effects as well as for the observed error terms. Consequently, in such case the distributions of the errors and/or random effects may be non-Gaussian.

Welsh and Richardson (1997) present several approaches and give an overview on how to robustly estimate parameters in linear mixed models. We use the high-breakdown S-estimators of Copt and Victoria-Feser (2006) for both the parameters of the mean as well as for the variance components. For the purpose of developing a conditional AIC, we need in addition the predictions of the random effects, for which we develop an estimation scheme.

Copt and Victoria-Feser (2006) work with the *marginal* likelihood in the linear mixed model where all $m_i = m$, and define the S-estimator for the vector β and the variance components $\sigma^2 = (\sigma_0^2, \dots, \sigma_r^2)$ as the values for β and σ^2 that minimize $\det(V_0)$ subject to the constraint

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1[\{(Y_i - X_i\beta)^t V_0^{-1}(Y_i - X_i\beta)\}^{1/2}] = b_1. \tag{2}$$

An appropriate choice of the function ρ_1 and of the value of b_1 will lead to robust estimators with a high breakdown point.

The loss function ρ_1 is a function that is even, continuously differentiable, non-decreasing on $[0, \infty)$, satisfies that $\rho_1(0) = 0$ and is bounded for above by 1, that is, $\sup_{\varepsilon \in \mathbb{R}} \rho_1(u) = 1$. We define $b_1 = E_{F_0}[\rho_1(\varepsilon)]$ to ensure consistency of the scale estimator under the central model F_0 and assume that $\epsilon_0 < b_1 < 1 - \epsilon_0$, here F_0 is the cumulative distribution function of ε. The notation E_{F_0} means that the expectation is computed with respect to F_0. When $\rho_1(x) = x^2$, the estimation method boils down to maximum likelihood estimation. A translated Tukey biweight function is proposed by Rocke (1996) and is used as ρ_1 in this paper, see also Copt and Victoria-Feser (2006),

$$\rho_1(d; c.M) = \begin{cases} \frac{d^2}{2}, & 0 \leq d \leq M \\ \rho_{M \leq d \leq M+c}(d; c, M), & M \leq d \leq M + c \\ \frac{M^2}{2} + \frac{c(5c+16M)}{30}, & d > M + c, \end{cases}$$

with $M + c < \infty$ and

$$\begin{aligned}\rho_{M \le d \le M+c}(d; c, M) &= \frac{M^2}{2} - \frac{M^2(M^4 - 5M^2c^2 + 15c^4)}{30c^4} + d^2\left(0.5 + \frac{M^4}{2c^4} - \frac{M^2}{c^2}\right) \\ &\quad + d^3\left(\frac{4M}{3c^2} - \frac{4M^3}{3c^4}\right) + d^4\left(\frac{3M^2}{2c^4} - \frac{1}{2c^2}\right) - \frac{4Md^5}{5c^4} + \frac{d^6}{6c^4},\end{aligned}$$

where the constants c and M can be chosen to achieve the desired breakdown point and asymptotic rejection probability.

We consider a *conditional* model for $Y|u$. In a first setting we assume that the random effects have a normal distribution $u_j \sim N(0, G_j)$. The conditional S-estimator (predictor) for the vectors β, u and the variance σ_0^2 are those parameter values that minimize det(R_0)= $|R_0|$ subject to the constraint

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1[\{(Y_i - X_i\beta - Z_i u)^t R_0^{-1}(Y_i - X_i\beta - Z_i u)\}^{1/2}] = b_1. \tag{3}$$

By following the idea of Henderson (1973) we provide an iterative system that gives in addition to estimators of (β, σ^2) the predictions of the random effects. In a likelihood setting the Henderson approach starts by phrasing the joint likelihood of (Y, u) as the product of the likelihood of $Y|u$ and the likelihood of u. In our context this leads to the following joint Lagrangian function, the maximization of which leads to estimators and predictors simultaneously,

$$\mathrm{L}_{\mathrm{joint}}(\beta, u, \sigma^2) = \log|R_0| + \frac{\tau_1}{n}\sum_{i=1}^{n}\{\rho_1(d_i) - b_1\} + \log|G| + u^t G^{-1} u, \tag{4}$$

where $d_i = d_i(\beta, u, R_0) = \{(Y_i - X_i\beta - Z_i u)^t R_0^{-1}(Y_i - X_i\beta - Z_i u)\}^{1/2}$ and τ_1 is a Lagrange multiplier. The estimators of β and σ^2 that result from this procedure are identical to those obtained by Copt and Victoria-Feser (2006) by using a marginal Lagrangian (4) and by omitting the part related to the marginal density of u, which is the reason why that approach does not automatically provide predictions of the random effects. The derivation of the estimators/predictions is given in Appendix A.

Result 1 *The S-estimators $\widehat{\beta}$ and $\widehat{\sigma}^2$ of the fixed effect parameters β and of the vector of variance components σ^2 and the S-predictions $\widehat{u}$ of the random effects u in the linear mixed model (1) obtained by maximizing the joint Lagrangian (4), assuming normality of the random effects, are equivalently obtained by iteratively solving the following set of equations,*

$$
\begin{aligned}
\widehat{\beta} &= (X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\widehat{W}\widehat{V}^{-1}Y && (5)\\
\widehat{u} &= \frac{\widehat{\tau}_1}{2n}\widehat{G}Z^t\widehat{W}\widehat{V}^{-1}(Y-X\widehat{\beta}) && (6)\\
\widehat{\sigma}_0^2 &= (\widehat{d}^t\widehat{W}\widehat{d})^{-1}(Y-X\widehat{\beta}-Z\widehat{u})^t\widehat{W}(Y-X\widehat{\beta}-Z\widehat{u}) && (7)\\
\widehat{\sigma}_j^2 &= \widehat{u}_j^t\widehat{u}_j/q_j, && (8)
\end{aligned}
$$

where $\widehat{W} = diag_{i=1,\dots,n}(\psi_1(\widehat{d}_i)/\widehat{d}_iI_m)$, $\widehat{d}_i = d_i(\widehat{\beta},\widehat{u},\widehat{R}_0)$, $\widehat{d} = (\widehat{d}_1,\dots,\widehat{d}_n)^t$ and the vector $\widehat{u}$ decomposes in components $\widehat{u}_j$ with length q_j, $j = 1,\dots,r$, $\widehat{\tau}_1 = 2mn(\widehat{d}^t\widehat{W}\widehat{d})^{-1}$,

$$
\widehat{V} = \widehat{R} + Z(\frac{\widehat{\tau}_1}{2n}\widehat{G})Z^t\widehat{W}. \qquad (9)
$$

When $\rho_1(t) = t^2$ the S-scale estimator $\widehat{\sigma}_0$ reduces to the sample standard deviation. In this case we have that $\widehat{W} = 2\,I_n$ and that $\widehat{\tau}_1 = n$. Hence, as expected, $\widehat{V} = \widehat{R} + Z\widehat{G}Z^t$ and (5) and (6) correspond to the maximum likelihood fixed and random effects formulae where $\widehat{\beta}_{ML} = (X^t\widehat{V}^{-1}X)^tX^t\widehat{V}^{-1}Y$ and $\widehat{u}_{ML} = \widehat{G}Z^t\widehat{V}^{-1}(Y - X\widehat{\beta}_{ML})$.

To accommodate possible outliers on the random effects we consider robust S-prediction of the random effects simultaneous with S-estimation of the fixed effects and variance components. For this purpose we define a new joint Lagrangian function that is to be optimized over β, u and σ^2,

$$
\mathrm{L}_{\mathrm{joint},2}(\beta,u,\sigma^2) = \log|R_0| + \frac{\tau_1}{n}\sum_{i=1}^{n}\{\rho_1(d_i)-b_1\} + \log|G| + \frac{\tau_2}{r}\sum_{j=1}^{r}\{\rho_2(d_{2j})-b_2\}. \quad (10)
$$

Here $d_{2,j} = (u_j^tG_j^{-1}u_j)^{1/2}$, ρ_1 and ρ_2 are both translated Tukey's bi-square family loss functions, which might be taken to be different functions, b_1 and b_2 are constants associated with the breakdown point of the estimator, defined by $b_k = E\{\rho_k(\sqrt{U})\}$, where U is a chi-squared distributed random variable with m degrees of freedom, where m is the length of the observation vectors Y_i, τ_1 and τ_2 are Lagrange multipliers.

Result 2 *The S-estimators $\widetilde{\beta}$ and $\widetilde{\sigma}^2$ of the fixed effect parameters β and of the variance components σ^2 and the S-predictions $\widetilde{u}$ of the random effects u in the linear mixed model (1) obtained by maximizing the joint Lagrangian (10), without assuming normality, are equivalently obtained by iteratively solving the following set of equations,*

$$\begin{aligned} \widetilde{\beta} &= (X^t\widetilde{W}\widetilde{V}^{-1}X)^{-1}X^t\widetilde{W}\widetilde{V}^{-1}Y \\ \widetilde{u} &= \frac{r\widetilde{\tau}_1}{n\widetilde{\tau}_2}\left(\widetilde{G}^{-1/2}\widetilde{W}_2\widetilde{G}^{-1/2}\right)^{-1} Z^t\widetilde{W}\widetilde{V}^{-1}(Y - X\widetilde{\beta}) \qquad (11) \\ \widetilde{\sigma}_0^2 &= (\widetilde{d}^t\widetilde{W}\widetilde{d})^{-1}(Y - X\widetilde{\beta} - Z\widetilde{u})^t\widetilde{W}(Y - X\widetilde{\beta} - Z\widetilde{u}) \\ \widetilde{\sigma}_j^2 &= \frac{\widetilde{\tau}_2}{2rq_j}\widetilde{u}_j^t\widetilde{W}_{2j}\widetilde{u}_j \qquad (12) \end{aligned}$$

where the matrix of weights $\widetilde{W} = diag_{i=1,\ldots,n}\{\psi_1(\widetilde{d}_i)/\widetilde{d}_iI_m\}$, $\widetilde{d}_i = d_i(\widetilde{\beta},\widetilde{u},\widetilde{R}_0)$, $\widetilde{d} = (\widetilde{d}_{11},\ldots,\widetilde{d}_{1n})^t$, $\widetilde{\tau}_1 = 2nm(\widetilde{d}^t\widetilde{W}\widetilde{d})^{-1}$,

$$\widetilde{V} = \widetilde{R} + Z\left(\frac{r\widetilde{\tau}_1}{n\widetilde{\tau}_2}(\widetilde{G}^{-1/2}\widetilde{W}_2\widetilde{G}^{-1/2})^{-1}\right)Z^t\widetilde{W}, \qquad (13)$$

$\widetilde{d}_{2j} = (\widetilde{u}_j^t\widetilde{G}_j^{-1}\widetilde{u}_j)^{1/2}$, $\widetilde{d}_2 = (\widetilde{d}_{21},\ldots,\widetilde{d}_{2r})^t$, $\widetilde{W}_2 = diag_{j=1,\ldots,r}(\psi_2(\widetilde{d}_{2j})/\widetilde{d}_{2j}I_{q_j})$,
$\widetilde{\tau}_2 = 2rq\left(\widetilde{d}_2^t\widetilde{W}_2\widetilde{d}_2\right)^{-1}$ and ψ_j $(j = 1, 2)$ is the first derivative of ρ_j.

When $\rho_2(x) = x^2$, the estimators presented in Result 2 coincide with those of Result 1.

Copt and Victoria-Feser (2006, end of Sec. 3.2) argue that arguments as in Davies (1987) can be used to obtain the breakdown point of the constrained estimators, similar arguments hold for the case of S-estimation in the conditional model.

The estimators of the variance components that are obtained by iteratively solving the sets of equations as in Results 1 and 2, are by construction non-negative.

3 AIC for S-estimation in linear mixed models

When both the error terms and the random effects are Gaussian,

$$2\log f(Y\,|\,\widehat{\beta},\widehat{u},\widehat{R}) = -N\log(2\pi) - \log|\widehat{R}| - (Y - X\widehat{\beta} - Z\widehat{u})^t\widehat{R}^{-1}(Y - X\widehat{\beta} - Z\widehat{u}),$$

with maximum likelihood or restricted maximum likelihood estimators $\widehat{\beta}$, $\widehat{u}$, $\widehat{\sigma}^2_\varepsilon$.

A marginal AIC follows from an immediate application of the original AIC (Akaike, 1973), it counts the number of estimated parameters to be used in the penalty part of the criterion and it uses the marginal likelihood of Y, with maximum likelihood estimators inserted for the unknown parameters,

$$\text{mAIC} = -2\,\log f_Y(Y;\widehat{\beta},\widehat{V}) + 2(p + r + 1).$$

Vaida and Blanchard (2005) obtain for variable selection when the random effects are of interest a conditional AIC, defined as

$$\text{cAIC} = -2\,\log f_{Y|u}(Y\,|\,\widehat{\beta},\widehat{u},\widehat{R}) + 2(\text{Tr}(H) + 1),$$

where $f_{Y|u}$ is the conditional likelihood for $Y|u$ and $H = C(C^tR^{-1}C + B)^{-1}C^tR^{-1}$, where $C = (X, Z)$ and $B = \text{diag}(0_p, G^{-1})$, where 0_p is a vector of zeros of length p. The added value of 1 in the penalty term reflects the estimation of the error variance σ^2_0.

The boundedness of the functions ρ_1 and ρ_2 for S-estimation has as a consequence that the transformation $\exp(-\rho_k)$, $k = 1, 2$, does not lead to a density function since its integral will be infinite. Hence a substitution of the model's density f by $\exp(-\rho_k)$ in expressions for the AIC is not valid when working with S-estimators, in contrast to the case of M-estimation where the unbounded ρ functions lead to valid density functions. Motivated by the m-variate normal likelihood with mean function μ and variance matrix Σ, a cAIC expression for M-estimation (Ronchetti, 1997) would replace the sum of the Mahalanobis distances by $\sum_{i=1}^{n} \rho(y_i; \mu, \Sigma)$. For S-estimation this, however, is the constant number nb. Indeed, the marginal multivariate S-estimator of (β, u, V) is defined by the minimization of $|V_0|$ subject to the constraint (2), while the conditional multivariate S-estimator of (β, u, R) is defined by the minimization of $|R_0|$ subject to the constraint (3). S-estimation requires a different approach towards defining the AIC. Following Tharmaratnam and Claeskens (2011), we come to the definition of a marginal and conditional AIC for use with S-estimation as

$$\text{mAIC.S1} = 2\log|\widehat{V}| + 2\,(p+q+1),\ \text{cAIC.S1} = 2\log|\widehat{R}| + 2\,\text{Tr}(\widehat{H}_S + 1),$$

where, from application of Result 1, the matrix $\widehat{H}_S = (I_N - \widehat{R}\widehat{V}^{-1}\widehat{P})$, with $\widehat{P} = I_N - X(X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\widehat{W}\widehat{V}^{-1}$.

When robustness in both ε and u is considered we use instead the matrices $\widetilde{R}, \widetilde{V}, \widetilde{W}$ (see Result 2), with the corresponding matrices $\widetilde{H}_S$ and $\widetilde{P}$, leading to

$$\text{mAIC.S2} = 2\log|\widetilde{V}| + 2\,(p+r+1),\ \text{cAIC.S2} = 2\log|\widetilde{R}| + 2\,\text{Tr}(\widetilde{H}_S + 1).$$

Wager et al. (2007) compare the use of marginal and conditional AIC values for selecting a model amongst penalized spline additive mixed models with hierarchical smooth terms. One of their findings is that cAIC performs better for more complex models.

Liang et al. (2008) obtain that $\Phi_0 = \text{Tr}\{\partial\widehat{Y}/(\partial Y)\}$ is a better penalty term than $\text{Tr}(H) + 1$, since it takes the effect of the estimation of the variance components into account. This is further studied and explicitly computed by Greven and Kneib (2010, Thm. 3) for (restricted) maximum likelihood estimation. A large part of the difficulty in arriving at computable expressions is that the estimators $(\widehat{\beta}, \widehat{u}, \widehat{\sigma}^2)$ depend on Y. The corrected conditional AIC from Greven and Kneib (2010) is

$$\text{ccAIC} = -2\,\log f_{Y|u}(Y\,|\,\widehat{\beta}, \widehat{u}, \widehat{R}) + 2\,\Phi_0. \tag{14}$$

For the case of S-estimation we explicitly obtain the generalized degrees of freedom for both situations with one or two levels of robustness. In these calculations we always consider $\sigma_\varepsilon^2 > 0$ to be unknown, and hence we do not need any additional adjustments in the penalty Φ_0 to account for the estimation of the error variance.

Theorem 1 *The generalized degrees of freedom* $\Phi_{S_1} = Tr\{\partial\widehat{Y}/(\partial Y)\}$ *when the estimators are obtained via the joint Lagrangian* (4) *are computed as:*

$$\Phi_{S1} = Tr(I_N - \widehat{R}\widehat{V}^{-1}\widehat{P} - B) \tag{15}$$

where

$$B \;=\; \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial Y}Y = \left(\frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial Y_1}Y, \ldots, \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial Y_N}Y\right)$$

of which the kth column ($k = 1, 2, \ldots, N$) *equals*

$$B_k \;=\; \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial \sigma_0^2}Y\frac{\partial\widehat{\sigma}_0^2}{\partial Y_k} + \sum_{j=1}^{r}\frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial \sigma_j^2}Y\frac{\partial\widehat{\sigma}_j^2}{\partial Y_k} = D_1 D_{2k} + \sum_{j=1}^{r} D_{3j}D_{4jk}.$$

Here, $D_1 = [I_N - \widehat{R}\widehat{V}^{-1}\{(I_N - \widehat{P})D_{v\sigma_0} - D_{w\sigma_0}\}]\widehat{V}^{-1}\widehat{P}$, $D_{2k} = -H_{\sigma_0}^{-1}H_{\sigma_0 Y_k}$, $D_{3j} = -\tau_1/(2n)\widehat{R}\widehat{V}^{-1}\widehat{P}ZZ^t\widehat{W}\widehat{V}^{-1}\widehat{P}$, $D_{4jk} = -H_{\sigma_j}^{-1}H_{\sigma_j Y_k}$, *with* $\widehat{V}$, $\widehat{W}$ *and* $\widehat{\tau}_1$ *as in Result 1,* $\widehat{A}_j = \widehat{\tau}_1/(2n)Z_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}Y$, $D_{v\sigma_0} = \partial\widehat{V}/\partial\sigma_0^2$, $D_{w\sigma_0} = \partial\widehat{W}/\partial\sigma_0^2$, $D_{\tau_1\sigma_0} = \partial\widehat{\tau}_1/\partial\sigma_0^2$, $D_{vY_k} = \partial\widehat{V}/\partial Y_k$, $D_{wY_k} = \partial\widehat{W}/\partial Y_k$, $D_{\tau_1 Y_k} = \partial\widehat{\tau}_1/\partial Y_k$, $D_{(V^{-1}P_k)Y_k} = \partial(\widehat{V}^{-1}\widehat{P}_k)/\partial Y_k$, P_k *is the kth column of the matrix* P. *In the above formulae we have used* $H_{\sigma_0} = -N/\sigma_0^4 - N^{-1}Y^t\widehat{P}^t\widehat{V}^{-1}[\widehat{W}\widehat{V}^{-1}\widehat{P}YD_{\tau_1\sigma_0} - 2\,\tau_1\widehat{W}Y\widehat{V}^{-1}\{(I_N - \widehat{P})D_{v\sigma_0} - D_{w\sigma_0}\}\widehat{V}^{-1}\widehat{P} + \tau_1 D_{w\sigma_0}\widehat{V}^{-1}\widehat{P}Y]$, $H_{\sigma_j} = -q_j/\sigma_j^4 - 2\,\widehat{A}_j^t\{\tau_1/(2n)Z_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}Z_j\}\widehat{A}_j$, $H_{\sigma_0 Y_k} = -n^{-1}Y_k^t\widehat{P}_k^t \times\widehat{V}^{-1}\{2\,\tau_1\widehat{W}\widehat{V}^{-1}\widehat{P}_k + D_{\tau_1 Y_k}\;\widehat{W}\widehat{V}^{-1}\widehat{P}_kY_k + 2\,\tau_1\widehat{W}D_{(V^{-1}P_k)Y_k}Y_k + \tau_1 D_{wY_k}\;\widehat{V}^{-1}\widehat{P}_kY_k\}$ *and* $H_{\sigma_j Y_k} = -n^{-1}\widehat{A}_j^t\{\tau_1 Z_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}_k + \tau_1 Z_j^tD_{wY_k}\widehat{V}^{-1}\widehat{P}_kY_k + \tau_1 Z_j^t\widehat{W}D_{(V^{-1}P_k)Y_k}Y_k + D_{\tau_1 Y_k}Z_j^t \times\widehat{W}\widehat{V}^{-1}\widehat{P}_kY_k\}$.

For the situation where we consider robustness aspects for both the random effects and the errors, the following theorem is proven in the Appendix. The main difference is in the partial derivatives of the predictions of the random effects, which has an effect on the derivatives of the fixed effects as well, through the connections shown in Result 2.

Theorem 2 *The generalized degrees of freedom* $\Phi_{S2} = Tr\{\partial\widehat{Y}/(\partial Y)\}$ *when the estimators are obtained via the joint Lagrangian* (10) *are computed as:*

$$\Phi_{S2} = Tr\left(I_N - \widetilde{R}\widetilde{V}^{-1}\widetilde{P} - \widetilde{B}\right) \tag{16}$$

where

$$\widetilde{B} = \frac{\partial(\widetilde{R}\widetilde{V}^{-1}\widetilde{P})}{\partial Y}Y = \left(\frac{\partial\widetilde{R}\widetilde{V}^{-1}\widetilde{P}}{\partial Y_1}Y, \ldots, \frac{\partial\widetilde{R}\widetilde{V}^{-1}\widetilde{P}}{\partial Y_N}Y\right)$$

of which the kth column ($k = 1, 2, \ldots, N$*) equals*

$$\widetilde{B}_k \;=\; \frac{\partial(\widetilde{R}\widetilde{V}^{-1}\widetilde{P})}{\partial\sigma_0^2}Y\frac{\partial\widetilde{\sigma}_0^2}{\partial Y_k} + \sum_{j=1}^{r}\frac{\partial(\widetilde{R}\widetilde{V}^{-1}\widetilde{P})}{\partial\sigma_j^2}Y\frac{\partial\widetilde{\sigma}_j^2}{\partial Y_k} = \widetilde{D}_1\widetilde{D}_{2k} + \sum_{j=1}^{r}\widetilde{D}_{3j}\widetilde{D}_{4jk}.$$

The quantities $\widetilde{D}_1$ *and* $\widetilde{D}_{2k}$ *are the same as in Theorem 1 though use the estimators* $\widetilde{V}$, $\widetilde{W_2}$, $\widetilde{d}_2$ *and* $\widetilde{\tau}_2$ *from Result 2,* $\widetilde{D}_{3j} = -\widetilde{R}\widetilde{V}^{-1}\widetilde{P}\widetilde{D}_{v\sigma_j}\widetilde{V}^{-1}\widetilde{P}$ *and* $\widetilde{D}_{4jk} = -\widetilde{H}_{\sigma_j}^{-1}\widetilde{H}_{\sigma_j Y_k}$*. Here* $\widetilde{D}_{v\sigma_j} = \partial\widetilde{V}/\partial\sigma_j^2$, $\widetilde{D}_{\tau_2\sigma_j} = \partial\widetilde{\tau}_2/\partial\sigma_j^2$, $\widetilde{D}_{d_{2j}\sigma_j} = \partial\widetilde{d}_{2j}/\partial\sigma_j^2$, $\widetilde{D}_{W_2\sigma_j} = \partial\widetilde{W_2}/\partial\sigma_j^2$, $\widetilde{D}_{\tau_2 Y_k} = \partial\widetilde{\tau}_2/\partial Y_k$, $\widetilde{D}_{d_{2j}Y_k} = \partial\widetilde{d}_{2j}/\partial Y_k$, $\widetilde{D}_{W_{2j}Y_k} = \partial\widetilde{W_{2j}}/\partial Y_k$*. Further, the derivatives* $\widetilde{H}_{\sigma_j} = -q_j/\widetilde{\sigma}_j^4 + 1/(2r\widetilde{\sigma}_j^2)\widetilde{d}_{2j}^t\ \{(\widetilde{\tau}_2/\widetilde{\sigma}_j^2)\widetilde{W_{2j}}\widetilde{d}_{2j} - \widetilde{W_{2j}}\widetilde{d}_{2j}\widetilde{D}_{\tau_2\sigma_j} - 2\widetilde{\tau}_2\widetilde{W_{2j}}\widetilde{D}_{d_{2j}\sigma_j} - \widetilde{\tau}_2\widetilde{D}_{W_{2j}\sigma_j}\widetilde{d}_{2j}\}$, $\widetilde{H}_{\sigma_j Y_k} = -1/(2r\widetilde{\sigma}_j^2)\widetilde{d}_{2j}^t\{\widetilde{W_{2j}}\widetilde{d}_{2j}\widetilde{D}_{\tau_2 Y_k} + 2\widetilde{\tau}_2\widetilde{W_{2j}}\widetilde{D}_{d_{2j}Y_k} + \widetilde{\tau}_2\widetilde{D}_{W_{2j}Y_k}\widetilde{d}_{2j}\}$.

The generalized degrees of freedom from Theorems 1 and 2 lead to corrected versions of the conditional AIC,

$$\text{ccAIC.S1} = 2\log|\widehat{R}| + 2\,\Phi_{S1},\ \ \text{ccAIC.S2} = 2\log|\widetilde{R}| + 2\,\Phi_{S2}. \tag{17}$$

While the expressions for the generalized degrees of freedom Φ_{S1} and Φ_{S2} might look formidable, they are straightforward to program in any matrix-language software (e.g. `R`). Our code is available from `https://perswww.kuleuven.be/Gerda_Claeskens/software/functionsRCAIC.R`.

4 Numerical results

4.1 Algorithm

An iterative procedure is required to compute the S-estimators in Results 1 and 2, as is the case for other S-estimation schemes, e.g. in linear regression models. The algorithm to obtain the estimators from Result 1 is described in the following steps,

Step 1: Let $\widehat{\beta}^{(0)}$, $\widehat{u}^{(0)}$, $(\widehat{\sigma}_0^2)^{(0)}$ and $(\widehat{\sigma}_j^2)^{(0)}$ be the initial values, for which we use Minimum Covariance Determinant (MCD) estimators from `covMcd` function in the `R` library `robustbase`.

Step 2: Set $k = 0$. Iterate the following steps until convergence:

(i) Compute the $\widehat{d}_i^{(1)}$, weights $\widehat{W}^{(1)}$ and $\widehat{\tau}^{(1)}$ as in Result 1.

(ii) Compute $\widehat{\beta}^{(1)}$ and $\widehat{u}^{(1)}$ by substituting $(\widehat{\sigma}_0^2)^{(0)}$, $(\widehat{\sigma}_j^2)^{(0)}$, $\widehat{d}_i^{(1)}$, $\widehat{W}^{(1)}$ and $\widehat{\tau}^{(1)}$ in the equations (5) and (6).

(iii) Compute $(\widehat{\sigma}_0^2)^{(1)}$, $(\widehat{\sigma}_j^2)^{(1)}$ by substituting $\widehat{\beta}^{(1)}$, $\widehat{u}^{(1)}$, $\widehat{d}_i^{(1)}$, $\widehat{W}^{(1)}$ and $\widehat{\tau}^{(1)}$ in the equations (7) and (8).

(iv) If either $k = maxit$ (i.e., the maximum number of iterations) or $|\widehat{\beta}^{(k)} - \widehat{\beta}^{(k+1)}| < \epsilon\,|\widehat{\beta}^{(k)}|$ where $\epsilon > 0$ is a fixed small constant (the tolerance level), then set $\widehat{\beta}^F = \widehat{\beta}^{(k)}$ and stop.

Step 3: Compute the final estimators $(\widehat{\sigma}_0^2)^{(F)}$, $(\widehat{\sigma}_j^2)^{(F)}$ by substituting $\widehat{\beta}^{(F)}$, $\widehat{u}^{(F)}$, $\widehat{d}_i^{(F)}$, $\widehat{W}^{(F)}$ and $\widehat{\tau}^{(F)}$ in the equations (7) and (8).

We used a similar algorithm for obtaining the estimates from Result 2. We have coded the above algorithm in `R`. In our experience the above algorithm converges without problems in the majority of the cases. The algorithm with $\epsilon = 10^{-6}$ and

$maxit = 500$ converges generally in less than 200 iterations. For all of our simulation experiments, we have never encountered a situation where the algorithm diverged.

4.2 Simulation results – S-estimators

To investigate the performance of the S1 and S2-estimators we fit mixed models as in (1) with generated outliers (i) only in the errors ε, (ii) in both ε and u. In a balanced design we take $n = 20$, $m = 4$, $r = 2$, $p = 7$, $q = 2$ and $N = 80$. A comparison is also made with the non-robust maximum likelihood estimators, using the function `lme` from the R library `nlme`. For the robust estimation methods, we used our own implementation of the algorithm in Section 4.1. We simulated 1000 samples for all simulation settings. We used a translated Tukey biweight function with 50% breakdown point and constant $c = 5.41$ to compute the S1- and S2-estimators in all cases.

Case 1: We consider a true model $Y = m_1(x, u) + \varepsilon$, with $x = (x_1, \ldots, x_6)$, and $m_1(x, u) = (\beta_0 + u_{1j}) + (\beta_1 + u2j)x_1 + \beta_2 x_2 + \beta_3 x_3$. We fit linear mixed models with six covariates in all simulation settings. The covariates are generated from a multivariate normal distribution with mean vector $\mu = (1, 1, \ldots, 1)$ and covariance matrix $\Sigma = I_6$, while the errors ε and random components each come from a $N(0, 1)$ distribution. We took $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 5$. To investigate the robustness of the estimation method against outliers, we generated different percentages of outliers (0%, 10%, 20% and 30%) from a $N(100, 0.5^2)$ distribution for the error terms and 10% outlying random components were generated from a student t-distribution with, respectively, 3 and 5 degrees of freedom for u_1 and u_2.

The fit of the estimated models is measured via the median squared prediction error (MSPE). Denoting $\widehat{m}_r(x)$ the estimated value of $m(x)$ for simulation run r, ($r =$

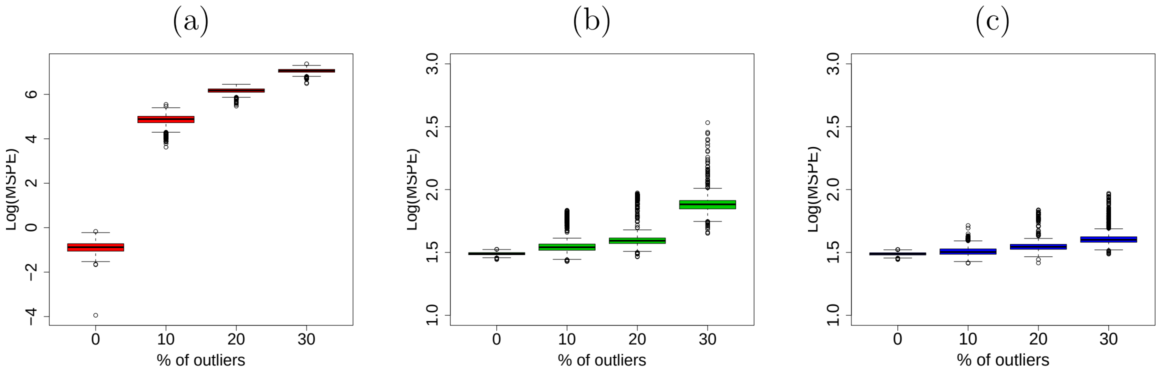


Figure 1: Case 1, outliers in both ε and u. Box plots of the log median squared prediction error using (a) ML, (b) S1 and (c) S2-estimation for samples with mean structure $m_1(x, u)$. S-estimators are computed with 50% breakdown point.

$1, 2, \ldots, 1000)$, the MSPE for the rth simulation run is defined by

$$\text{MSPE}_r = \text{median}\{[m(x_i) - \widehat{m}_r(x_i)]^2, i = 1, \ldots, n\}.$$

boxplots on the log scale of the MSPE values visualize the variability of the obtained estimates, see Figure 1. It is observed that the MSPEs of the S1- and the S2-estimators remain stable as the proportion of contamination increases. The ML-estimator's MSPEs increase in the presence of outliers, even with only 10% of outliers. The S2-estimators perform better than the S1-estimators in the case outliers are present in both the errors and the random components.

4.3 Simulation results – Variable selection

We compare the different versions of the AIC, see Section 3. In each case, the largest model contains six covariates, some of them are redundant. The simulation results are summarized by reporting the proportions of selected models that are (C) a correct fit – the true model only, (O) overfit – models containing all the variables in the true model plus some more that are redundant, (U) underfit – models with only a strict subset of the variables in the true model, (W) wrong fit – all other models. These are the models where some of the relevant variables might be present (though not all of them) in addition to some of the redundant variables.

For case 1 we select amongst the fixed and random components of the model and fit all possible subsets of the largest model to the data. Again, outliers on the response variable are generated from a $N(100, 0.5^2)$ distribution in different percentages (10%, 20% and 30%). From table 1 is clearly observed that the performance of the three marginal AICs is inferior to those of the conditional AICs, which is to be expected since in this setting we select both the fixed and the random components in the model. The conditional S1 and S2-methods have a good performance in the sense of having larger percentages of correctly selected models and smaller percentages of wrong and underfit models, also in the case that no outliers are present in the data, while these methods are preferred in the case of outliers. Similar results were obtained (not shown) in case only the errors ε contain outliers. In that situation, S1 and S2 gave practically the same results, while S2 is the preferred method for the situation of outliers in the random effects.

We considered two other simulation settings. Case 2 is taken from Greven and Kneib (2010), where $m_2(x) = 1+x+2d(0.3-x)^2$. The covariate values x are generated from a uniform distribution on the interval $[0, 1]$. In the model, d is a constant and increasing values of d correspond to the increased non-linearity. We generate 11 different models corresponding to $d = (0, 5, 10, \ldots, 50)$. The model is linear in x when $d = 0$. In the case of no outliers, the error terms ε have a standard normal distribution. We fit a cubic thin plate regression spline model

$$Y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k |x_i - \kappa_k|^3 + \varepsilon_i,$$

using maximum likelihood estimation and the S1-estimation method. In the mixed

Table 1: Case 1. Outliers in u and ε. Proportion of selected models for the true function $m_1(x, u)$ for $p = 6$, error terms and random components from a $N(0, 1)$ distribution, and for sample size $n = 100$. We consider 10%, 20% and 30% of outliers in ε generated from $\varepsilon \sim N(100, 0.5^2)$ and 10% random component outliers generated from student t-distributions with 3 and 5 degrees of freedom for, respectively, u_1 and u_2. S-estimators are computed with a 50% breakdown point.

		AIC			AIC.S1			AIC.S2		
ε %		m	c	cc	m	c	cc	m	c	cc
0	C	0.425	0.536	0.532	0.378	0.484	0.487	0.351	0.499	0.544
	O	0.345	0.456	0.445	0.324	0.403	0.458	0.357	0.420	0.406
	U	0.230	0.000	0.000	0.224	0.094	0.048	0.213	0.049	0.023
	W	0.000	0.008	0.023	0.074	0.019	0.007	0.079	0.032	0.027
10	C	0.010	0.017	0.074	0.345	0.498	0.512	0.379	0.498	0.525
	O	0.009	0.069	0.081	0.342	0.424	0.421	0.326	0.399	0.422
	U	0.574	0.611	0.608	0.234	0.044	0.046	0.201	0.051	0.046
	W	0.407	0.303	0.237	0.079	0.034	0.021	0.094	0.052	0.007
20	C	0.009	0.089	0.021	0.324	0.428	0.473	0.348	0.437	0.527
	O	0.018	0.058	0.024	0.231	0.395	0.374	0.297	0.408	0.430
	U	0.554	0.522	0.532	0.219	0.134	0.058	0.250	0.131	0.011
	W	0.419	0.331	0.423	0.226	0.043	0.095	0.105	0.024	0.032
30	C	0.008	0.001	0.023	0.327	0.475	0.411	0.354	0.487	0.501
	O	0.014	0.075	0.089	0.264	0.309	0.335	0.274	0.392	0.412
	U	0.594	0.658	0.674	0.321	0.214	0.226	0.353	0.114	0.035
	W	0.384	0.266	0.214	0.088	0.002	0.028	0.019	0.007	0.052

model formulation the u_k are random variables with mean zero and variance σ_u^2. We placed the knots according to the quantiles of the data, for sample size $n = 100$ there were 24 knots. For the non-robust estimation methods we have used the `R` library `SemiPar`, function `spm`,

For each value of the constant d, for each simulated data set, we use the mAIC, ccAIC, mAIC.S1 and ccAIC.S1 to decide on either the linear model (with $d = 0$) or the more complex model (with the given value of d). To assess the performance of the marginal and the conditional AIC in distinguishing the linear and non-linear models, we compute the frequency of selecting the nonlinear model for each d value. We use 1000 simulated data sets for both cases with (a) no outliers and (b) 20% outliers on the error terms, generated from a $N(100, 0.5^2)$ distribution for the sample size $n = 100$. From Figure 2 we observe that the corrected conditional AIC selects a larger proportion of nonlinear models than the marginal AIC (which is the true model when $d \neq 0$). This holds for both maximum likelihood estimators and S1-estimators. In these penalized spline models, the random effects correspond to the spline coefficients. The conditional AIC is better suited to decide on the inclusion of random effects (i.e. nonlinear effects in this setup) than the marginal AIC. The results do not change much for different values of d. The significant effect of the robust methods is clearly visible in the case that outliers are present.

For case 3 we consider fitting semiparametric additive models Case 3: $m_3(x) = 1 + 2d_1 \cos(\pi x_1) + d_2 \sin((0.5 - x_2)^2) + x_3$, with $d_1 = 15, d_2 = 25$. The covariates $x_1, \ldots, x_6$ are generated from a multivariate normal distribution which is the same as in case 1. The full model that is fit to the data is

$$Y_i = \beta_0 + \sum_{j=1}^{6} \beta_j x_{ji} + \sum_{k=1}^{K} u_{1k} |x_{1i} - \kappa_k|^3 + \sum_{k=1}^{K} u_{2k} |x_{2i} - \kappa_k|^3 + \varepsilon_i,$$

that is, cubic thin plate splines are used to model smooth functions of x_1, x_2, while

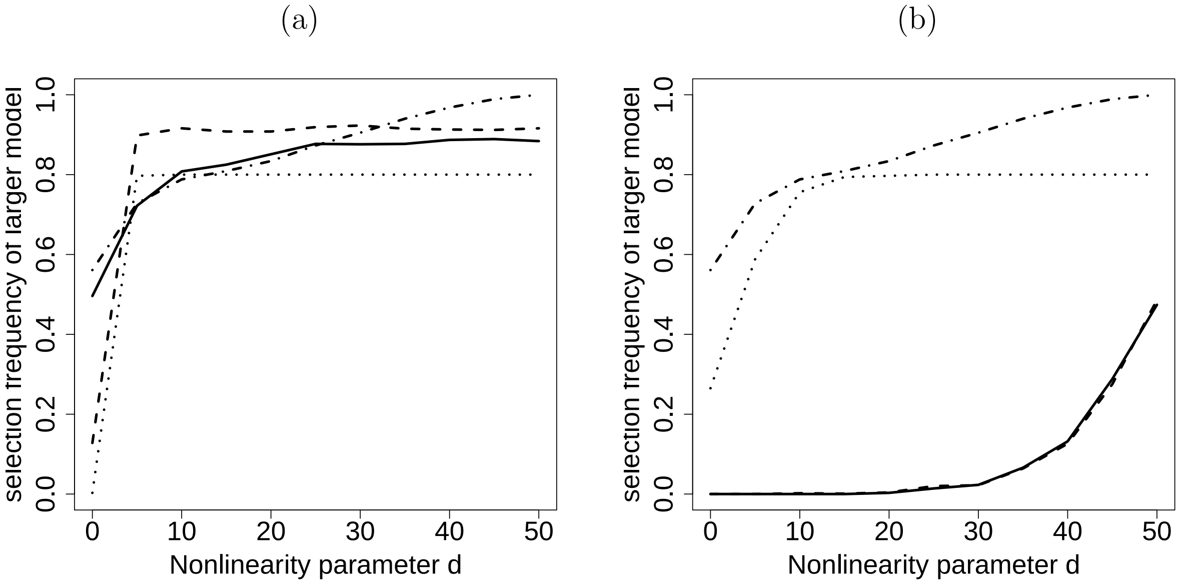


Figure 2: Case 2. Proportion of selected larger models from mAIC (solid line), ccAIC (dashed line), mAIC.S1 (dotted line) and ccAIC.S1 (dot-dashed line) with mean function $m_2(x)$. (a) no outliers in the data, (b) 20% of outliers in the error variables ε.

$x_3, \ldots, x_6$ enter the model in a linear way. We fit model with all possible combinations of the six covariates, resulting in $(2^6 - 1)$ different models.

For case 3 we conduct selection amongst the parametric and nonparametric (random) components of the model. This results in fitting $2^6 - 1$ different models to the data. Again, outliers on the response variable are generated from a $N(100, 0.5^2)$ distribution in different percentages (10%, 20% and 30%). Based on the results from Table 2 we clearly observe that the performance of the two marginal AICs (mAIC and mAIC.S1) is inferior to that of the conditional AICs, which is to be expected since in this setting we select both the parametric and the nonparametric components in the model. Table 2 shows that the conditional S1-methods have a good performance also in the case that no outliers are present in the data, and these methods are preferred

Table 2: Case 3. Proportion of selected models for data generated with dependent xs, mean $m_3(x)$ for $p = 6$, error terms from a $N(0, 1)$ distribution, and for sample size $n = 100$. We consider different % of outliers on ε, generated from $N(100, 0.5^2)$. S-estimators are computed with a 50% breakdown point.

%		mAIC	cAIC	ccAIC	mAIC.S1	cAIC.S1	ccAIC.S1
0	C	0.383	0.494	0.442	0.270	0.432	0.499
	O	0.307	0.471	0.483	0.210	0.361	0.371
	U	0.231	0.000	0.000	0.364	0.059	0.062
	W	0.079	0.035	0.075	0.156	0.148	0.068
10	C	0.009	0.010	0.011	0.257	0.422	0.474
	O	0.003	0.006	0.008	0.200	0.343	0.352
	U	0.654	0.638	0.685	0.346	0.056	0.059
	W	0.334	0.346	0.296	0.198	0.179	0.115
20	C	0.006	0.008	0.010	0.236	0.409	0.432
	O	0.002	0.004	0.016	0.216	0.337	0.428
	U	0.672	0.683	0.698	0.318	0.052	0.107
	W	0.320	0.305	0.276	0.230	0.202	0.033
30	C	0.002	0.006	0.005	0.283	0.399	0.422
	O	0.001	0.004	0.007	0.491	0.376	0.395
	U	0.710	0.693	0.706	0.084	0.079	0.106
	W	0.287	0.297	0.282	0.142	0.146	0.077

in the case of outliers. Higher proportions of correct and overfit models are obtained when the corrected versions of the conditional AIC are used.

5 Discussion

The need for robust model selection methods in linear mixed models has lead us to develop the generalized degrees of freedom for S-estimation methods. In multilevel models, extreme or outlying observations might occur at any level. Theoretical properties of the proposed estimation and variable selection approach are worth a separate study, which, however, extends beyond the scope of the current paper.

It would be interesting to develop the proposed estimation method and the subsequent generalized degrees of freedom that we have used in a conditional AIC, for other random effect models, such as generalized linear mixed models. Several non robust model selection methods exist for generalized linear mixed models, see for example, Cai et al. (2006), Chen et al. (2003) and Lavergne et al. (2008). Also an extension towards survival-type data is relevant, for which, for example, Ibrahim and Chen (2005), Hjort and Claeskens (2006), Kneib and Fahrmeir (2007) and Liang and Zou (2008) proposed some model selection methods. Müller and Welsh (2009) extend their bootstrap-based model selection method (Müller and Welsh, 2005) which is robust against outliers from linear models to generalized linear models, however, not including random effects. Xu et al. (2009) proposed a semiparametric model selection method with application to proportional hazards mixed models using profile likelihood, however non robust to outliers in the data. Ideas in those papers could be used to propose a robust version of AIC for survival models.

A Appendix. Computation of S-estimators for linear mixed models

A.1 Proof of Result 1

Setting the partial derivatives of L_{joint} in (4) with respect to β, u and the vector σ^2 to zero, and solving for these values, yields estimators $\widehat{\beta}, \widehat{u}, \widehat{\sigma}^2$. We arrive at

$$\begin{aligned} \widehat{\beta} &= (X^t\widehat{W}\widehat{R}^{-1}X)^{-1}X^t\widehat{W}\widehat{R}^{-1}(Y-Z\widehat{u}) && (18)\\ \widehat{u} &= (Z^t\widehat{W}\widehat{R}^{-1}Z+\frac{2n}{\widehat{\tau}_1}\widehat{G}^{-1})^{-1}Z^t\widehat{W}\widehat{R}^{-1}(Y-X\widehat{\beta}). && (19)\end{aligned}$$

Substituting (19) in equation (18) yields (5), while substituting (5) in (19) yields (6). We write $\widehat{V}^{-1}=\widehat{R}^{-1}-\widehat{R}^{-1}Z(Z^t\widehat{W}\widehat{R}^{-1}Z+\frac{2n}{\widehat{\tau}}\widehat{G}^{-1})^{-1}Z^t\widehat{W}\widehat{R}^{-1}$ from which it follows that $\widehat{V}=\widehat{R}+Z(\frac{\widehat{\tau}}{2n}\widehat{G})Z^t\widehat{W}$.

Equating the partial derivative of L_{joint} with respect to σ_0^2 to zero yields, first, by solving for τ_1, that $m=\widehat{\tau}_1/(2n)\sum_{i=1}^n W(\widehat{d}_i)(Y_i-X_i\widehat{\beta}-Z_i\widehat{u})^t\widehat{R}_i^{-1}(Y_i-X_i\widehat{\beta}-Z_i\widehat{u})$, from which follows that $\widehat{\tau}_1=2mn(\widehat{d}^t\widehat{W}\widehat{d})^{-1}$. Second, solving for σ_0^2 yields that

$$\widehat{\sigma}_0^2=\frac{\widehat{\tau}_1}{2mn}(Y-X\widehat{\beta}-Z\widehat{u})^t\widehat{W}(Y-X\widehat{\beta}-Z\widehat{u}),$$

from which (7) follows. The partial derivatives of L_{joint} with respect to σ_j^2 $(j=1,\dots,q)$, which occur only in the matrix G give that $\widehat{\sigma}_j^2=\widehat{u}_j^t\widehat{u}_j/q_j$ in case the true variances are nonzero. The maximizers of the joint Lagrangian are sought either at the values where the first derivative is equal to zero, or at the value zero, which is at the boundary of the parameter space.

A.2 Proof of Result 2

The estimators for β, σ_0^2 and τ_1 are obtained similarly as in Result 1 though now starting from the joint Lagrangian (10). The expressions for the predictors $\widetilde{u}$ and for the variance component estimators are different. After substituting the estimator $\widetilde{\beta}$ in the next equation,

$$\widetilde{u} = (Z^t\widetilde{W}\widetilde{R}^{-1}Z + \frac{n\widetilde{\tau}_2}{q\widetilde{\tau}_1}\widetilde{G}^{-1/2}\widetilde{W}_2\widetilde{G}^{-1/2})^{-1}Z^t\widetilde{W}\widetilde{R}^{-1}(Y - X\widetilde{\beta}),$$

the estimator $\widetilde{u}$ of (11) results. When the true values of the variance components are positive, from

$$\frac{\partial \mathrm{L}_{\mathrm{joint2}}(\widetilde{\beta}, \widetilde{u}, \sigma^2)}{\partial \sigma_j^2}\Big|_{\sigma_j^2 = \widetilde{\sigma}_j^2} = \frac{\partial}{\partial \sigma_j^2}\{\log|G| + \frac{\widetilde{\tau}_2}{r}\sum_{k=1}^{r}\rho_2(\widetilde{d}_{2k})\} = 0, \qquad (20)$$

for all $j = 1, \ldots, r$, (12) follows. Since (20) implies that also the sum over $j = 1, \ldots, r$ of these partial derivatives is equal to zero, $\widetilde{\tau}_2 = 2rq(\sum_{j=1}^{r}\widetilde{u}_j^t\widetilde{W}_{2j}\widetilde{G}_j^{-1}\widetilde{u}_j)^{-1} = 2rq(\widetilde{d}_2^t\widetilde{W}_2\widetilde{d}_2)^{-1}$. When a true variance component is zero, the global maximum is found at the boundary of the parameter space.

B Appendix. Generalized degrees of freedom for the S-estimators

B.1 Proof of Theorem 1

We start from model (1) and assume that all variance components are unknown. The generalized degrees of freedom is defined by $\Phi_{S1} = \mathrm{Tr}\left(\partial\widehat{Y}/(\partial Y)\right)$. From (5) and (6) it follows that

$$\widehat{Y} \;=\; X\widehat{\beta} + Z\widehat{u} = X\widehat{\beta} + Z\left(\frac{\widehat{\tau}_1}{2n}\widehat{G}Z^t\widehat{W}\widehat{V}^{-1}(Y - X\widehat{\beta})\right).$$

The expression of V from Result 1, see (9), leads to rewriting $Z(\frac{\widehat{\tau}_1}{2n}\widehat{G})Z^t\widehat{W} = \widehat{V} - \widehat{R}$, from which it follows that $\widehat{Y} = X\widehat{\beta} + (I_N - \widehat{R}\widehat{V}^{-1})(Y - X\widehat{\beta}) = Y - \widehat{R}\widehat{V}^{-1}\widehat{P}Y$ where $\widehat{P} = I_N - X(X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\widehat{W}\widehat{V}^{-1}$. Thus Φ_{S1} is as in (15). With $\widehat{R} = \widehat{\sigma}_0^2 I_N$ and

$\widehat{G}_j = \widehat{\sigma}_j^2 I_{q_j}$, $j = 1, \ldots, r$, Y is a vector of length N, Y_k is the kth element of the vector Y. The $N \times N$ matrix B with columns $B_1, \ldots, B_N$ as in Theorem 1, is re-written using the chain rule as follows,

$$B_k \;=\; \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial\sigma_0^2} Y \frac{\partial\widehat{\sigma}_0^2}{\partial Y_k} + \sum_{j=1}^{r} \frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial\sigma_j^2} Y \frac{\partial\widehat{\sigma}_j^2}{\partial Y_k}. \tag{21}$$

A further application of the chain rule leads to $\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})/\partial\sigma_0^2 =$

$$\widehat{V}^{-1}\widehat{P} - \widehat{R}\widehat{V}^{-1}\Big\{\frac{\partial\widehat{V}}{\partial\sigma_0^2} - X(X^t\widehat{W}\widehat{V}^{-1}X)^{-1}X^t\big(\widehat{W}\widehat{V}^{-1}\frac{\partial\widehat{V}}{\partial\sigma_0^2} - \frac{\partial\widehat{W}}{\partial\sigma_0^2}\big)\Big\}\widehat{V}^{-1}\widehat{P}. \tag{22}$$

Starting from (9), $\frac{\partial\widehat{V}}{\partial\sigma_0^2} = I_N + Z(\frac{1}{2n}\widehat{G})Z^t\widehat{W}\frac{\partial\widehat{\tau}_1}{\partial\sigma_0^2} + Z(\frac{\widehat{\tau}_1}{2n}\widehat{G})Z^t\frac{\partial\widehat{W}}{\partial\sigma_0^2}$, where it holds that $\frac{\partial\widehat{\tau}_1}{\partial\sigma_0^2} = -2mn(\widehat{d}^t\widehat{W}\widehat{d})^{-1}\big(2\widehat{d}^t\widehat{W}\frac{\partial\widehat{d}}{\partial\sigma_0^2} + \widehat{d}^t\frac{\partial\widehat{W}}{\partial\sigma_0^2}\widehat{d}\big)(\widehat{d}^t\widehat{W}\widehat{d})^{-1}$, $\frac{\partial\widehat{d}_i}{\partial\sigma_0^2} = \frac{1}{2\widehat{d}_i}(Y_i - X_i\widehat{\beta} - Z_i\widehat{u})^t\widehat{R}_i^{-1}\widehat{R}_i^{-1}(Y_i - X_i\widehat{\beta} - Z_i\widehat{u})$, $\frac{\partial\widehat{W}}{\partial\sigma_0^2} = \mathrm{diag}_{i=1,\ldots,n}\Big[\Big(\frac{\widehat{d}_i\;\psi_1'(\widehat{d}_i) - \psi_1(\widehat{d}_i)}{\widehat{d}_i^2}\Big)\frac{\partial\widehat{d}_i}{\partial\sigma_0^2}I_m\Big]$. Since from (9) it follows that $\frac{\partial\widehat{V}}{\partial\sigma_j^2} = \frac{\widehat{\tau}_1}{2n}Z_jZ_j^t\widehat{W}$ and since $\frac{\partial\widehat{P}}{\partial\sigma_j^2} = \frac{\widehat{\tau}_1}{2n}(I_N - \widehat{P})Z_jZ_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}$, it follows that $\frac{\partial(\widehat{R}\widehat{V}^{-1}\widehat{P})}{\partial\sigma_j^2} = -\frac{\widehat{\tau}_1}{2n}\widehat{R}\widehat{V}^{-1}\widehat{P}Z_jZ_j^t\widehat{W}\widehat{V}^{-1}\widehat{P}$. Define for $j = 0, \ldots, r$ for the case where the variance components are all positive

$$\frac{\partial \mathrm{L}_{\mathrm{joint}}(\widehat{\beta}, \widehat{u}, \sigma^2)}{\partial\sigma_j^2}\Big|_{\sigma_j^2 = \widehat{\sigma}_j^2} = h(\widehat{\sigma}_j^2(Y), Y) = 0. \tag{23}$$

Using the estimators from Result 1, $h(\widehat{\sigma}_0^2(Y), Y) = \frac{m}{\widehat{\sigma}_0^2} - \frac{\widehat{\tau}_1}{n}(Y - X\widehat{\beta} - Z\widehat{u})^t\widehat{R}^{-1}\widehat{W}\widehat{R}^{-1}(Y - X\widehat{\beta} - Z\widehat{u}) = \frac{m}{\widehat{\sigma}_0^2} - \frac{\widehat{\tau}_1}{n}Y^t\widehat{P}^t\widehat{V}^{-1}\widehat{W}\widehat{V}^{-1}\widehat{P}Y$. In this expression $\widehat{\tau}_1$, $\widehat{P}$ and $\widehat{W}$ are a function of Y and $\widehat{\sigma}_0^2$. Take the full differentiation of $h(\widehat{\sigma}_0^2(Y), Y)$ with respect to Y_k,

$$\frac{dh(\widehat{\sigma}_0^2(Y), Y)}{dY_k} = \frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial\sigma_0^2}\frac{d\widehat{\sigma}_0^2}{dY_k} + \frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial Y_k} = 0,$$

to find that $\frac{d\widehat{\sigma}_0^2}{dY_k} = -\Big[\frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial\sigma_0^2}\Big]^{-1}\frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial Y_k}$. From Result 1 it follows that $\frac{\partial h(\widehat{\sigma}_0^2(Y), Y)}{\partial Y_k} = H_{\sigma_0 Y_k}$, with $\frac{\partial\widehat{\tau}_1}{\partial Y_k} = -2mn(\widehat{d}^t\widehat{W}\widehat{d})^{-1}\Big[2\widehat{d}^t\widehat{W}\frac{\partial\widehat{d}}{\partial Y_k} + \widehat{d}^t\frac{\partial\widehat{W}}{\partial Y_k}\widehat{d}\Big](\widehat{d}^t\widehat{W}\widehat{d})^{-1}$, $\frac{\partial\widehat{d}_i}{\partial Y_k} = (Y_i - X_i\widehat{\beta} - Z_i\widehat{u})^t\widehat{R}_i^{-1}/\widehat{d}_i$, $\frac{\partial\widehat{W}}{\partial Y_k} = \mathrm{diag}_{i=1,\ldots,n}\Big[\Big(\{\widehat{d}_i\psi_1'(\widehat{d}_i) - \psi_1(\widehat{d}_i)\}/\widehat{d}_i^2\Big)\frac{\partial\widehat{d}_i}{\partial Y_k}I_m\Big]$. Further it follows from (9) and from matrix differentiation rules that $\frac{\partial(\widehat{V}^{-1}\widehat{P}_k)}{\partial Y_k} = -\widehat{V}^{-1}\frac{\partial\widehat{V}}{\partial Y_k}\widehat{V}^{-1}\widehat{P}_k +$

$\widehat{V}^{-1}\frac{\partial \widehat{P}_k}{\partial Y_k}$, where

$$\begin{aligned}\frac{\partial \widehat{V}}{\partial Y_k} &= \frac{1}{2n} Z \frac{\partial \widehat{\tau}_1}{\partial Y_k} \widehat{G} Z^t \widehat{W} + Z \frac{\widehat{\tau}_1}{2n} \widehat{G} Z^t \frac{\partial \widehat{W}}{\partial Y_k}, \\ \frac{\partial \widehat{P}_k}{\partial Y_k} &= X(X^t \widehat{W} \widehat{V}^{-1} X)^{-1} X^t \left(\widehat{W} \widehat{V}^{-1} \frac{\partial \widehat{V}}{\partial Y_k} - \frac{\partial \widehat{W}}{\partial Y_k} \right) \widehat{V}^{-1} \widehat{P}_k.\end{aligned}$$

With the calculations done so far, we immediately obtain that $\frac{\partial h(\widehat{\sigma}_0^2(Y),Y)}{\partial \sigma_0^2} = H_{\sigma_0}$, where $\frac{\partial(\widehat{V}^{-1}\widehat{P})}{\partial \sigma_0^2} = -\widehat{V}^{-1}\Big\{ \frac{\partial \widehat{V}}{\partial \sigma_0^2} - X(X^t \widehat{W} \widehat{V}^{-1} X)^{-1} X^t \big(\widehat{W} \widehat{V}^{-1} \frac{\partial \widehat{V}}{\partial \sigma_0^2} - \frac{\partial \widehat{W}}{\partial \sigma_0^2} \big) \Big\} \widehat{V}^{-1} \widehat{P}$. We consider next the functions $h(\widehat{\sigma}_j^2(Y), Y)$ for $j = 1, \ldots, r$. Using the expressions from Result 1, $h(\widehat{\sigma}_j^2(Y), Y) = \frac{q_j}{\widehat{\sigma}_j^2} - \widehat{u}_j^t \widehat{G}_j^{-1} \widehat{G}_j^{-1} \widehat{u}_j = \frac{q_j}{\widehat{\sigma}_j^2} - \frac{\widehat{\tau}_1}{2n} Y^t \widehat{P}^t \widehat{V}^{-1} \widehat{W} Z_j \frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P} Y = \frac{q_j}{\widehat{\sigma}_j^2} - \widehat{A}_j^t \widehat{A}_j$, where $\widehat{A}_j = \frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P} Y$. By the full differentiation of h, this further leads to $\frac{d\widehat{\sigma}_j^2}{dY_k} = -\left[\frac{\partial h(\widehat{\sigma}_j^2(Y),Y)}{\partial \sigma_j^2} \right]^{-1} \frac{\partial h(\widehat{\sigma}_j^2(Y),Y)}{\partial Y_k}$, where via similar calculations we arrive at $\frac{\partial h(\widehat{\sigma}_j^2(Y),Y)}{\partial Y_k} = H_{\sigma_j Y_k}$ and

$$\frac{\partial h(\widehat{\sigma}_j^2(Y), Y)}{\partial \sigma_j^2} = -\frac{q_j}{\widehat{\sigma}_j^4} - 2\widehat{A}_j^t \frac{\partial \widehat{A}_j}{\partial \sigma_j^2} = -\frac{q_j}{\widehat{\sigma}_j^4} - 2\widehat{A}_j^t \left(\frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P} Z_j \right) \widehat{A}_j,$$

where it holds that

$$\frac{\partial \widehat{A}_j}{\partial \sigma_j^2} = \frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \frac{\partial(\widehat{V}^{-1}\widehat{P})}{\partial \sigma_j^2} Y = \frac{\widehat{\tau}_1}{2n} Z_j^t \widehat{W} \frac{\widehat{\tau}_1}{2n} \widehat{V}^{-1} \widehat{P} Z_j Z_j^t \widehat{W} \widehat{V}^{-1} \widehat{P} Y.$$

When the variance component is estimated to be zero, by the use of the iterative procedure from Result 1, there exists a sequence of estimators that are positive, for which the above derivatives are defined. We define the derivative at the boundary value by the limit of the derivatives of the estimators in this sequence. Considering the expressions for the derivatives, this will result in a zero contribution to the effective degrees of freedom, see also Greven and Kneib (2010). This proves Theorem 1.

B.2 Proof of Theorem 2

The proof goes along the same lines as that of Theorem 1, with this difference that we use the estimators from Result 2, and in particular the expressions for (16) and (21) with these estimators, in addition to (23) using now L_{joint2}. This leads to obtaining $\frac{\partial h(\widetilde{\sigma}_0^2(Y),Y)}{\partial Y_k}$ and $\frac{\partial h(\widetilde{\sigma}_0^2(Y),Y)}{\partial \sigma_0^2}$ from which we in a similar way arrive at the estimator $\frac{d\widetilde{\sigma}_0^2}{dY_k} = -\left[\frac{\partial h(\widetilde{\sigma}_0^2(Y),Y)}{\partial \sigma_0^2}\right]^{-1} \frac{\partial h(\widetilde{\sigma}_0^2(Y),Y)}{\partial Y_k}$. The quantities $\widetilde{R}$, $\widetilde{\tau}_1$ and $\widetilde{W}$ do not depend on $\widetilde{\sigma}_j^2; j = 1, \dots, r$. From (13),

$$\begin{aligned}\frac{\partial \widetilde{V}}{\partial \sigma_j^2} &= Z\frac{r\widetilde{\tau}_1}{n\widetilde{\tau}_2}(\widetilde{G}_j^{-1/2}\widetilde{W}_{2j}\widetilde{G}_j^{-1/2})^{-1}\left[\widetilde{G}_j^{-1/2}\big(\widetilde{W}_{2j}\widetilde{G}_j^{-1} + \frac{\partial \widetilde{W}_{2j}}{\partial \sigma_j^2}\big)\widetilde{G}_j^{-1/2}\right.\\ &\qquad \left.\times(\widetilde{G}_j^{-1/2}\widetilde{W}_{2j}\widetilde{G}_j^{-1/2})^{-1} - \frac{1}{\widetilde{\tau}_2}\frac{\partial \widetilde{\tau}_2}{\partial \sigma_j^2}\right] Z^t\widetilde{W}.\end{aligned}$$

With $\partial \widetilde{d}_{2j}/\partial \sigma_j^2 = -\frac{1}{2}\widetilde{G}_j^{-1}\widetilde{d}_{2j}$, and δ_{jk} the Kronecker delta such that $\delta_{jk} = 1$ if and only if $j = k$, and $\delta_{jk} = 0$ otherwise,

$$\begin{aligned}\frac{\partial \widetilde{W}_{2j}}{\partial \sigma_j^2} &= \text{diag}_{k=1,\dots,r}\left[\left(\delta_{jk}\frac{\widetilde{d}_{2k}\psi_1'(\widetilde{d}_{2k}) - \psi_1(\widetilde{d}_{2k})}{\widetilde{d}_{2k}^2}\right)\frac{\partial \widetilde{d}_{2k}}{\partial \sigma_j^2}I_{q_k}\right]\\ \frac{\partial \widetilde{\tau}_2}{\partial \sigma_j^2} &= -2qr(\widetilde{d}_2^t\widetilde{W}_2\widetilde{d}_2)^{-1}\left[2\widetilde{d}_2^t\widetilde{W}_2\frac{\partial \widetilde{d}_2}{\partial \sigma_j^2} + \widetilde{d}_2^t\frac{\partial \widetilde{W}_2}{\partial \sigma_j^2}\widetilde{d}_2\right](\widetilde{d}_2^t\widetilde{W}_2\widetilde{d}_2)^{-1}.\end{aligned}$$

All this taken together gives us $\partial(\widetilde{R}\widetilde{V}^{-1}\widetilde{P})/(\partial\sigma_j^2)$. Defining

$$\frac{\partial \text{L}_{\text{joint2}}(\widetilde{\beta}, \widetilde{u}, \sigma^2)}{\partial \sigma_j^2}|_{\sigma_j^2 = \widetilde{\sigma}_j^2} = 0 = h_2(\widetilde{\sigma}_j^2(Y), Y) = q_j/\sigma_j^2 - \frac{\widetilde{\tau}_2}{r\widetilde{\sigma}_j^2}\widetilde{d}_{2j}^t\widetilde{W}_{2j}\widetilde{d}_{2j},$$

it follows that $\partial h_2(\widetilde{\sigma}_j^2(Y), Y)/\partial Y_k = \widetilde{H}_{\sigma_j Y_k}$, and $\partial h_2(\widetilde{\sigma}_j^2(Y), Y)/\partial \sigma_j^2 = \widetilde{H}_{\sigma_j}$, where $\frac{\partial \widetilde{\tau}_2}{\partial Y_k} = -2qr(\widetilde{d}_{2j}^t\widetilde{W}_{2j}\widetilde{d}_{2j})^{-1}[2\widetilde{d}_{2j}^t\widetilde{W}_{2j}\frac{\partial \widetilde{d}_{2j}}{\partial Y_k} + \widetilde{d}_{2j}^t\frac{\partial \widetilde{W}_{2j}}{\partial Y_k}\widetilde{d}_{2j}](\widetilde{d}_{2j}^t\widetilde{W}_{2j}\widetilde{d}_{2j})^{-1}$, $\frac{\partial \widetilde{d}_{2j}}{\partial Y_k} = \frac{r\widetilde{\tau}_1}{n\widetilde{\tau}_2}\widetilde{G}_j^{-1/2} \times (\widetilde{G}_j^{-1/2}\widetilde{W}_{2j}\widetilde{G}_j^{-1/2})^{-1}Z_j^t\widetilde{W}\widetilde{V}^{-1}\widetilde{P}_kY_k$, $\frac{\partial \widetilde{W}_{2j}}{\partial Y_k} = \text{diag}_{j=1,\dots,r}[(\frac{\widetilde{d}_{2j}\psi_2'(\widetilde{d}_{2j}) - \psi_2(\widetilde{d}_{2j})}{\widetilde{d}_{2j}^2})\frac{\partial \widetilde{d}_{2j}}{\partial Y_k}]$. This leads to $\frac{d\widetilde{\sigma}_j^2}{dY_k} = -[\frac{\partial h_2(\widetilde{\sigma}_j^2(Y),Y)}{\partial \sigma_j^2}]^{-1}\frac{\partial h_2(\widetilde{\sigma}_j^2(Y),Y)}{\partial Y_k}$, from which the stated results follows. The case of zero variance components is handled similarly as in Theorem1.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Bondell, H., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077.

Cai, B., Dunson, D. B., and Gladen, T. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics*, 62:446–457.

Chen, M.-H., Ibrahim, J. G., Shao, Q.-M., and Weiss, R. E. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111(1-2):57 – 76.

Copt, S. and Victoria-Feser, M. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association*, 101(473):292–300.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15(3):1269–1292.

Greven, S. and Kneib, T. (2010). On the behavior of marginal and conditional Akaike information criteria in linear mixed models. *Biometrika*, 97(4):773–789.

Henderson, C. R. (1973). Maximum likelihood estimation of variance components. Technical report, Department of Animal Science, Cornell University.

Hjort, N. and Claeskens, G. (2006). Focussed information criteria and model averaging for Cox's hazard regression model. *Journal of the American Statistical Association*, 101:1449–1464.

Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika*, 88(2):367–379.

Ibrahim, J. G. and Chen, M.-H. (2005). *Bayesian Model Selection in Survival Analysis*. John Wiley & Sons, Ltd.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34(1):207–228.

Lavergne, C., Martinez, M.-J., and Trottier, C. (2008). Empirical model selection in generalized linear mixed effects models. *Computational Statistics*, 23:99–109.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, 95(3):773–778.

Liang, H. and Zou, G. (2008). Improved AIC selection strategy for survival analysis. *Computational Statistics & Data Analysis*, 52(5):2538–2548.

Müller, S. and Welsh, A. (2005). Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, 100:1297–1310.

Müller, S. and Welsh, A. (2009). Robust model selection in generalized linear models. *Statistica Sinica*, 19:1155–1170.

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of statistics*, 24:1327–1345.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica*, 7:327–338.

Tharmaratnam, K. and Claeskens, G. (2011). A comparison of robust versions of the AIC based on M, S and MM-estimators. *Statistics*, in press: DOI:10.1080/02331888.2011.568120.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92:351–370.

Wager, C., Vaida, F., and Kauermann, G. (2007). Model selection for penalized spline smoothing using Akaike information criteria. *Aust. N. Z. J. Stat.*, 49(2):173–190.

Welsh, A. H. and Richardson, A. M. (1997). Approaches to the robust estimation of mixed models. In *Robust inference*, volume 15 of *Handbook of Statistics*, pages 343–384. North-Holland, Amsterdam.

Xu, R., Vaida, F., and Harrington, D. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica*, 19:819–842.