University of Wollongong
Research Online

Centre for Statistical & Survey Methodology
Working Paper Series

Faculty of Engineering and Information Sciences

2008

Robustness for general design mixed models using the t-distribution

J. Staudenmayer
University of Massachusetts - Amherst

E. E. Lake
Alkermes Inc, Mass.

M. P. Wand
University of Wollongong, mwand@uow.edu.au

Recommended Citation

Staudenmayer, J.; Lake, E. E.; and Wand, M. P., Robustness for general design mixed models using the t-distribution, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 22-08, 2008, 22p.
http://ro.uow.edu.au/cssmwp/19

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

22-08

Robustness for general design mixed models using the *t*-distribution

J. Staudenmayer, E. E. Lake and Matt P. Wand

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Robustness for general design mixed models using the t-distribution

J. STAUDENMAYER[1],

Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003, U.S.A.

E. E. LAKE,

Alkermes Inc., Cambridge, Massachusetts 02139, U.S.A.

AND M. P. WAND

School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522 AUSTRALIA

18th February, 2008

SUMMARY

The t-distribution allows the incorporation of outlier robustness into statistical models while retaining the elegance of likelihood-based inference. In this paper we develop and implement a linear mixed model for the general design of the linear mixed model using the univariate t-distribution. This general design allows a considerably richer class of models to be fit than is possible with existing methods. Included in this class are semi-parametric regression and smoothing, and spatial models.

Keywords: Additive Model; Nonparametric Regression; Random Effects, Semi-parametric Models; Spatial Statistics.

1 Introduction

Mixed models are a flexible extension of ordinary regression models that have proven useful for dealing with repeated measures (e.g. Laird and Ware 1982), spatial correlation (e.g. O'Connell and Wolfinger 1997), and non-linearity through spline-based models (e.g. Wahba 1978; Speed 1991; Lin and Zhang 1999; Kammann and Wand 2002; Ruppert, Wand, and Carroll 2003). These developments and their various combinations allow mixed models to handle a wide variety of problems in a modular framework. It is important to note that in order for the mixed model to be useful for spline-based models, the random effects design matrix cannot be restricted to be of a specific form such as the block diagonal design matrices that are used in repeated

[1]email: `jstauden@math.umass.edu`

measures designs. We use the label *general design* for models that do not make such restrictions.

In more detail, the usual form of the general design linear mixed model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \mathsf{E}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \mathbf{0}, \quad \mathrm{Cov}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}. \tag{1}$$

(e.g. Robinson 1991; McCulloch and Searle 2001, Chapter 6) for an $n \times 1$ vector of response values $\mathbf{y}$, design matrices $\mathbf{X}$ $(n \times p)$ and $\mathbf{Z}$ $(n \times q)$, a vector $\boldsymbol{\beta}$ $(p \times 1)$ of unknown fixed effects, a vector $\mathbf{u}$ $(q \times 1)$ of unobserved random effects, and an $n \times 1$ vector of unobserved errors $\boldsymbol{\varepsilon}$.

Another common version of the linear mixed model is the hierarchical or Laird-Ware model:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad \mathsf{E}\begin{bmatrix}\mathbf{u}_i\\ \boldsymbol{\varepsilon}_i\end{bmatrix} = \mathbf{0}, \quad \mathrm{Cov}\begin{bmatrix}\mathbf{u}_i\\ \boldsymbol{\varepsilon}_i\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}, i = 1, \ldots, n, \tag{2}$$

with $\mathbf{y}_i$ $(m \times 1)$ and $\mathbf{y}_i$ and $\mathbf{y}_j$ marginally uncorrelated when $i \neq j$. Typically, $n > 1$. This model can be gotten from the general design model by using a block diagonal structure for $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{G}$, and $\mathbf{R}$ in (1). In contrast, spline based models and spatial models require non-block diagonal versions of at least some of those matrices. The general design model also is the hierarchical model when $n = 1$. The length of $\mathbf{u}_i$ is typically a small integer in the hierarchical model, but the length of $\mathbf{u}$ usually is at least thirty in a spline or spatial model.

When $\mathbf{u}$ and / or $\boldsymbol{\varepsilon}$ are not normally distributed, the specification of the model though the means and covariances of those terms can be ambiguous. As a result, in this paper we use the following modification of (1) to allow direct incorporation of non-normal distributions for the $\mathbf{u}$ and/or $\boldsymbol{\varepsilon}$ random vectors. First let $\mathbf{G} = (\mathbf{G}^{1/2})^{\mathsf{T}}\mathbf{G}^{1/2}$ be the Cholesky decomposition of $\mathbf{G}$ and $\mathbf{u}^0(q \times 1)$ be a vector containing an independent and identically distributed sample of a zero mean and unit variance distribution. Let $\mathbf{R}^{1/2}$ and $\boldsymbol{\varepsilon}^0(n \times 1)$ be defined similarly. Note that the distribution used to construct $\mathbf{u}^0$ can differ from that used to construct $\boldsymbol{\varepsilon}^0$. The general design linear mixed model that we consider in this paper is then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \mathbf{u} = (\mathbf{G}^{1/2})^{\mathsf{T}}\mathbf{u}^0, \quad \boldsymbol{\varepsilon} = (\mathbf{R}^{1/2})^{\mathsf{T}}\boldsymbol{\varepsilon}^0. \tag{3}$$

The linear mixed model is widely used, and standard estimators of its parameters are based on the multivariate normal likelihood. There also has been a great deal of recent work to relax the assumption of normal $\mathbf{u}$ in favor of a non-parametric or semiparametric specification of the random effect density (eg. Kleinman and Ibrahim 1998; Aitkin 1999; Tao et al. 1999; Zhang and Davidian 2001; Ghidey, Lesaffre, and Eilers 2004). A separate series of papers (Lange, Little, and Taylor 1989; Welsh and Richardson 1997; Gianola and Stranden 1998; van Dyk 2000; and Pinhiero, Liu, and Wu 2001) have replaced the normal assumption with specification of a density that

is outlier robust, such as the t or multivariate-t distribution. Additional work work by Rosa, Gianola, and Padovani (2003) and Rosa, Padovani, and Gianola (2004) consider two additional outlier robust distributions, the slash distribution and the contaminated normal. Work by Jara and Quintana (2006) has used the broader class of skew elliptical distributions of which t-distributions are special cases. Perhaps it is surprising, but all these references address hierarchical, "repeated measures" type mixed models with $n > 1$, i.e. not general design models. While the work above could probably be modified to accomodate the general design model, it is outside the scope of this paper to do so. Instead, our goal is to develop a model that uses a vector univariate t random variables to implement a general design linear mixed model and achieve outlier robustness. We find the simplicity of this model to be appealing.

A seemingly natural and simple approach to making the general design linear mixed model outlier robust would be to specify a multivariate-t distribution of dimension n for $\boldsymbol{\varepsilon}^0$ in (3). The multivariate-t distribution is not unique (e.g. Kotz and Nadarajah 2004, Chapters 4 and 5), but a standard definition is

$$f(\mathbf{x}; \mu\mathbf{1}_n, \psi^2\mathbf{I}_n, \nu) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\nu\pi)^{n/2}\Gamma\left(\frac{\nu}{2}\right)(n\psi^2)^{1/2}\left(1+\frac{(\mathbf{x}-\mu\mathbf{1}_n)^\mathsf{T}(\mathbf{x}-\mu\mathbf{1}_n)}{\psi^2\nu}\right)^{(\nu+n)/2}}, \quad \mathbf{x} \in \mathbb{R}^n,$$

where $\mathbf{1}_n$ is a vector of length n with all ones, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. A perhaps under-appreciated fact is that if a single random n-vector, $\mathbf{x}$, is sampled from that density, then it can be shown easily that the maximum likelihood estimator (MLE) for μ is the non-robust least squares estimator $\mathbf{1}_n^\mathsf{T}\mathbf{x}/n$ (e.g. Breusch, Robertson, and Welsh 1997). That result is in contrast to the situation when $x_i, i = 1, ..., n$ are iid from the univariate density

$$f(t; \mu, \psi^2, \nu) = \psi^{-1}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left\{1+\frac{(t-\mu)^2}{\nu\psi^2}\right\}^{-\left(\frac{\nu+1}{2}\right)},$$

and the maximum likelihood estimate of μ does not have a closed form but is outlier robust. That fact suggests that an outlier robust general design linear mixed model should be based on multiple univariate-t random variables rather than the multivariate-t random vector, and that is our approach.

The contents of this paper are as follows. Section 2 contains the formulation of our robust t-linear mixed model. In Section 3 we describe the estimation algorithm we propose. The computation of standard errors is given in Section 4, and in Section 5 we illustrate our approach with two examples that motivated this work. Conclusions are in Section 6.

2 Model formulation

Common usage of model (3) assumes that the response $\mathbf{y}$ (and hence the errors ε) and the random effects $\mathbf{u}$ follow Gaussian distributions,

$$\begin{aligned} \mathbf{y}|\mathbf{u} &\sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}) \\ \mathbf{u} &\sim \mathrm{N}(\mathbf{0}, \mathbf{G}). \end{aligned} \tag{4}$$

For clarity, this paper's methods are presented for the commonly encountered case when $\mathbf{R} = \sigma_\varepsilon^2 \mathbf{I}_n$, $\mathbf{G} = \text{blockdiag}(\sigma_{u,j}^2 \mathbf{I}_{q_j})$ with $1 \leq j \leq c$, and $\mathbf{u} = (\mathbf{u}_1^\mathsf{T}, \ldots, \mathbf{u}_c^\mathsf{T})^\mathsf{T}$ where u_{jk} is the kth entry of $\mathbf{u}_j$. An Appendix contains the straightforward extensions required by the general variance structures in (2). Using the simpler structure for the variance components, we can rewrite (4) as

$$\begin{aligned} y_i|\mathbf{u} &\overset{\text{ind.}}{\sim} \mathrm{N}\left((\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i, \sigma_\varepsilon^2\right), \quad 1 \leq i \leq n \\ u_{jk} &\overset{\text{ind.}}{\sim} \mathrm{N}\left(0, \sigma_{u,j}^2\right), \quad 1 \leq j \leq c, 1 \leq k \leq q_j. \end{aligned}$$

In many applications, however, the data may contain outliers which negate the plausibility of (4). To reduce the influence of such outliers we propose t-distributed assumptions. Specifically we assume

$$\begin{aligned} y_i|\mathbf{u} &\overset{\text{ind.}}{\sim} t\left((\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i, \sigma_\varepsilon^2, \nu_y\right), \quad 1 \leq i \leq n \\ u_{jk} &\overset{\text{ind.}}{\sim} t\left(0, \sigma_{u,j}^2, \nu_u\right), \quad 1 \leq j \leq c, 1 \leq k \leq q_j \end{aligned} \tag{5}$$

where $t(\mu, \psi^2, \nu)$ denotes the t-distribution with density function $f(t; \mu, \psi^2, \nu)$ defined in the previous section. We refer to (5), a robust version of (3), as the t-linear mixed model.

The fitting alrorithm relies on an equivalent "precision-mixture" formulation which we now describe. First, let $\mathbf{v}_y = (v_{y_1}, \ldots, v_{y_n})$ where $v_{y_i} \overset{\text{i.i.d.}}{\sim} \chi^2_{\nu_y}/\nu_y$. Next, let $\mathbf{v}_u = (\mathbf{v}_{u_1}^\mathsf{T}, \ldots, \mathbf{v}_{u_c}^\mathsf{T})^\mathsf{T}$ where the kth entry of $\mathbf{v}_{u_j}$ is $v_{u_{jk}} \overset{\text{i.i.d.}}{\sim} \chi^2_{\nu_u}/\nu_u$, $1 \leqslant k \leqslant q_j$, $1 \leqslant j \leqslant c$. Adopt the convention that $1/\mathbf{v} = (1/v_1, \ldots, 1/v_n)$ for a general $n \times 1$ vector $\mathbf{v}$. Then (5) can be expressed as:

$$\begin{aligned} \mathbf{y}|\mathbf{u}, \mathbf{v}_y &\sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_\varepsilon^2 \text{diag}\left(1/\mathbf{v}_y\right)\right) \\ \mathbf{u}|\mathbf{v}_u &\sim \mathrm{N}\left(\mathbf{0}, \text{blockdiag}\left(\sigma_{u,j}^2/\mathbf{v}_{u_j}\right)\right). \end{aligned} \tag{6}$$

In subsequent sections we also use the notation $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}_u^2, \sigma_\varepsilon^2)$ where $\boldsymbol{\sigma}_u^2 = (\sigma_{u,1}^2, \ldots, \sigma_{u,c}^2)$.

We present model (5) where both $\mathbf{y}$ conditional on $\mathbf{u}$ and $\mathbf{u}$ follow t-distributions. In the next section we describe a method to estimate the parameters in this model. The applications we consider (Section 5) are well modeled with $\mathbf{y}$ conditional on $\mathbf{u}$ as a t-distribution and $\mathbf{u}$ normally distributed.

3 Estimation

Model (5) provides more resistance to outlying observations than does the Gaussian formulation (4), however, the log-likelihood for (5) has no closed form and maximization in the parameters proves a cumbersome task. A tractable and relatively simple approach to maximum likelihood estimation in the t-distribution formulation is the Monte Carlo Expectation Conditional Maximization (MCECM) algorithm. Other approaches to computing maximum likelihood estimators for the non-Gaussian mixed model are reviewed in McCulloch and Searle (2001), Section 10.3 for instance.

Section 3.1 provides an overview of the EM algorithm and the ECM extension. Section 3.2 provides a summary of the Monte Carlo EM (MCEM) method which uses sampling methods in the expectation step of the EM algorithm. As with the EM algorithm, MCEM can be extended to the ECM case which is also presented in Section 3.2. We present a detailed outline of our proposed algorithm for fitting t-linear mixed models in Section 3.3. Finally, we describe how we select the degrees of freedom corresponding to the t-distributions in Section 3.4.

3.1 EM algorithm and ECM extension

In formulation (6) the responses $\mathbf{y}$ are observed, but the random effects $\mathbf{u}$ and the auxiliary variables $\mathbf{v}_y$ and $\mathbf{v}_u$ are not observed. Such unobserved parameters are referred to as latent data. The Expectation Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) provides a simple iterative approach to finding maximum likelihood estimates in problems involving latent data.

In general, define the complete data $\mathbf{Y}_{comp} = (\mathbf{Y}_{obs}, \mathbf{Y}_{lat})$ to be the observed data $\mathbf{Y}_{obs}$ augmented with the latent data $\mathbf{Y}_{lat}$. Denote the complete and observed data log likelihoods by $l_{comp}(\boldsymbol{\theta})$ and $l_{obs}(\boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta}$ represents the parameter vector to be estimated. The EM algorithm maximizes $l_{obs}(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ by repeatedly applying two steps, an E-step followed by an M-step until convergence. Let $\boldsymbol{\theta}^{(m)}$ denote the current value of $\boldsymbol{\theta}$ after m cycles of the algorithm. The E-step computes the conditional expectation of $l_{comp}(\boldsymbol{\theta})$ given the observed data with a density that uses the parameter estimates from the most recent iteration,

$$\mathbf{Q}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(m)}) = \mathbf{E}[l_{comp}(\boldsymbol{\theta})|\mathbf{Y}_{obs};\, \boldsymbol{\theta}^{(m)}].$$

The M-step maximizes $\mathbf{Q}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(m)})$ as a function of $\boldsymbol{\theta}$, achieving

$$\boldsymbol{\theta}^{(m+1)} = \text{argmax}_{\boldsymbol{\theta}}\, \mathbf{Q}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(m)}).$$

A modification of the M-step results in the ECM generalization of the EM algorithm (Meng and Rubin 1993). This approach maximizes over a subset of the parameters in $\boldsymbol{\theta}$, holding the remainder of $\boldsymbol{\theta}$ fixed (thus the term *conditional maximization*). For instance, if $\boldsymbol{\theta} = (\theta_1, \theta_2)$, the M-step would consist of maximizing $\mathbf{Q}(\theta_1, \theta_2^{(m)}|\theta_1^{(m)}, \theta_2^{(m)})$

over θ_1 to obtain $\theta_1^{(m+1)}$, then maximizing $\mathbf{Q}(\theta_1^{(m+1)}, \theta_2|\theta_1^{(m)}, \theta_2^{(m)})$ over θ_2 to obtain $\theta_2^{(m+1)}$. The algorithm then returns to the E-step described above. The ECM approach may provide a computationally simpler alternative to jointly maximizing over $\boldsymbol{\theta}$. That is, each step is simpler, but the algorithm as a whole may require more iterations than the EM routine. Note that $l_{comp}(\boldsymbol{\theta})$ increases with each iteration and a stationary point (local or global) of the EM (or ECM) algorithm is determined, given regularity conditions (Wu 1983).

The t-linear mixed model (6) we propose in Section 2 has a complete data log likelihood denoted by $l_{comp}(\boldsymbol{\theta}) = l(\boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{\sigma}_u^2, \nu_y, \nu_u; \mathbf{y}, \mathbf{u}, \mathbf{v}_y, \mathbf{v}_u)$ which equals

$$l_1(\boldsymbol{\beta}, \sigma_\varepsilon^2; \mathbf{y}|\mathbf{u}, \mathbf{v}_y) + l_2(\boldsymbol{\sigma}_u^2; \mathbf{u}|\mathbf{v}_u) + l_3(\nu_y; \mathbf{v}_y) + l_4(\nu_u; \mathbf{v}_u) \tag{7}$$

where

$$l_1(\boldsymbol{\beta}, \sigma_\varepsilon^2; \mathbf{y}|\mathbf{u}, \mathbf{v}_y) = -\frac{1}{2}\log\prod_{i=1}^{n}\frac{\sigma_\varepsilon^2}{v_{y_i}} - \frac{1}{2}\sum_{i=1}^{n}\{y_i - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i\}^2\frac{v_{y_i}}{\sigma_\varepsilon^2},$$

$$l_2(\boldsymbol{\sigma}_u^2; \mathbf{u}|\mathbf{v}_u) = -\frac{1}{2}\log\prod_{j=1}^{c}\frac{(\sigma_{u,j}^2)^{q_j}}{\prod_{k=1}^{q_j} v_{u_{jk}}} - \frac{1}{2}\mathbf{u}^{\mathsf{T}}\text{blockdiag}\left(\frac{1}{\sigma_{u,j}^2}\mathbf{I}_{q_j}\mathbf{v}_{u_j}\right)\mathbf{u},$$

$$l_3(\nu_y|\mathbf{v}_y) = \sum_{i=1}^{n}\left[\frac{\nu_y}{2}\left\{\log\left(\frac{\nu_y}{2}\right) + \log(v_{y_i}) - v_{y_i}\right\} - \log(v_{y_i}) - \log\Gamma\left(\frac{\nu_y}{2}\right)\right],$$

$$l_4(\nu_u; \mathbf{v}_u) = \sum_{k=1}^{K}\left[\frac{\nu_u}{2}\left\{\log\left(\frac{\nu_u}{2}\right) + \log(v_{u_k}) - v_{u_k}\right\} - \log(v_{u_k}) - \log\Gamma\left(\frac{\nu_u}{2}\right)\right].$$

The precision-mixture formulation (6) permits straightforward maximum likelihood estimation via an ECM approach. The E-step in such an algorithm requires calculation of conditional expectations (conditional on $\mathbf{y}$) of the four terms in (7). Since none of these expectations have closed forms we use a Monte Carlo version of the E-step.

3.2 MCEM algorithm and MCECM extension

The E-step in an EM algorithm may require computation of intractable integrals. The Monte Carlo EM method (Wei and Tanner 1990) offers an alternative that replaces analytic determination of the required conditional expectations with Monte Carlo estimates.

To implement model (5), the E-step may be carried out by sampling from the joint distribution of $\mathbf{u}, \mathbf{v}_y, \mathbf{v}_u|\mathbf{y}; \boldsymbol{\theta}^{(m)}$ and then computing the MC estimate of the required expectations. To generate the required variates, we use a Gibbs sampler that alternates between samples from the distributions of $\mathbf{u}|\mathbf{y}, \mathbf{v}_y^{(p-1)}, \mathbf{v}_u^{(p-1)}$ (multivariate normal), $\mathbf{v}_y|\mathbf{y}, \mathbf{u}^{(p)}$, and $\mathbf{v}_u|\mathbf{u}^{(p)}$ (scaled chi squared), where $p = 1, \ldots, P_m$ indexes

the generated Gibbs samples and the number of samples (P_m) increases with the EM step (m.) The increase rule that we use is $P_m = P_{m-1} + P_{m-1}/5$ every fifth EM step with $P_0 = 10$ and rounding up. A computationally more efficient rule would be to test whether the Monte Carlo error is large enough relative to the algorithm's progress at each iteration to merit an increased Monte Carlo sample size (e.g. Booth and Hobert 1999; Levine and Casella 2001).

By sequentially sampling from these simpler conditional distributions we obtain samples from the desired joint conditional distribution. The required sampling distributions are

$$\mathbf{u}|\mathbf{y}, \mathbf{v}_y, \mathbf{v}_u \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8}$$

where, for $q = q_1 + \ldots + q_c$,

$$\boldsymbol{\mu} = \mathbf{A}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad \boldsymbol{\Sigma} = \text{blockdiag}\left(\sigma^2_{u,j}\mathbf{I}_{q_j}\frac{1}{\mathbf{v}_{u_j}}\right) - \mathbf{AZ}\,\text{blockdiag}\left(\sigma^2_{u,j}\mathbf{I}_{q_j}\frac{1}{\mathbf{v}_{u_j}}\right),$$

with $\mathbf{A} = \text{blockdiag}\left(\sigma^2_{u,j}\mathbf{I}_{q_j}\frac{1}{\mathbf{v}_{u_j}}\right)\mathbf{Z}^\mathsf{T}\left\{\sigma^2_\varepsilon \text{diag}\left(\frac{1}{\mathbf{v}_y}\right) + \mathbf{Z}\,\text{blockdiag}\left(\sigma^2_{u,j}\mathbf{I}_{q_j}\frac{1}{\mathbf{v}_{u_j}}\right)\mathbf{Z}^\mathsf{T}\right\}^{-1}$, and

$$v_{y_i}|\mathbf{y}, \mathbf{u} \overset{\text{ind.}}{\sim} \frac{\chi^2_{\nu_y+1}}{(\nu_y + \delta^2_{y_i})}, \quad \text{with} \quad \delta^2_{y_i} = \frac{\{y_i - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu})_i\}^2}{\sigma^2_\varepsilon}; \tag{9}$$

$$v_{u_{jk}}|\mathbf{u} \overset{\text{ind.}}{\sim} \frac{\chi^2_{\nu_u+1}}{(\nu_u + \delta^2_{u_{jk}})}, \quad \text{with} \quad \delta^2_{u_{jk}} = \frac{u^2_{jk}}{\sigma^2_{u,j}}. \tag{10}$$

The full algorithm is described in the next section.

3.3 Complete algorithm for fitting t-linear mixed models

1. Either fix or decide to estimate the degrees of freedom corresponding to the t-distributions to be used in fitting the model. If the degrees of freedom are to be estimated, see Section 3.4 for how to modify the following algorithm.
2. Start with initial parameter estimates $\boldsymbol{\theta}^{(0)}$. As is the case for many iterative algorithms, it is safest run the algorithm from several starting points ($\boldsymbol{\theta}^{(0)}$) in an attempt to identify and try to avoid convergence to local optima.
3. The mth, ($m = 1, \ldots, M$) parameter update begins by using the Gibbs sampling scheme discussed in Section 3.2 to generate samples from distributions (8), (9), and (10). This computes the Monte Carlo E-step of the MCECM algorithm, for the mth iteration.
4. Given the Gibbs samples $\left\{\mathbf{u}^{(p)}\right\}_{p=1}^{P_m}$, $\left\{\mathbf{v}_y^{(p)}\right\}_{p=1}^{P_m}$, and $\left\{\mathbf{v}_u^{(p)}\right\}_{p=1}^{P_m}$, the parameter updates are calculated as:

$$\boldsymbol{\beta}^{(m)} = \underset{\boldsymbol{\beta}}{\text{argmax}}\, \text{E}\left[l_1(\boldsymbol{\beta}, \sigma^{2(m-1)}_\varepsilon; \mathbf{y}|\mathbf{u}, \mathbf{v}_y)|\mathbf{y}\right]$$

$$
\begin{aligned}
&= \left[\frac{1}{P_m}\left\{\sum_{p=1}^{P_m}\frac{1}{\sigma_\varepsilon^{(m-1)}}\mathbf{X}^\mathsf{T}\mathrm{diag}\left(\mathbf{v}_y^{(p)}\right)\mathbf{X}\right\}\right]^{-1} \\
&\qquad \times\frac{1}{P_m}\left\{\sum_{p=1}^{P_m}\frac{1}{\sigma_\varepsilon^{(m-1)}}\mathbf{X}^\mathsf{T}\mathrm{diag}\left(\mathbf{v}_y^{(p)}\right)\left(\mathbf{y}-\mathbf{Z}\mathbf{u}^{(p)}\right)\right\}, \\
\sigma_\varepsilon^{2(m)} &= \underset{\sigma_\varepsilon^2}{\mathrm{argmax}}\,\mathrm{E}\left[l_1(\boldsymbol{\beta}^{(m)},\sigma_\varepsilon^2;\mathbf{y}|\mathbf{u},\mathbf{v}_y)|\mathbf{y}\right] \\
&= \frac{1}{P_m}\sum_{p=1}^{P_m}\frac{1}{n}\left[\sum_{i=1}^{n}\left\{y_i-(\mathbf{X}\boldsymbol{\beta}^{(m)}+\mathbf{Z}\mathbf{u}^{(p)})_i\right\}^2\mathbf{v}_{y_i}^{(p)}\right], \quad \text{and} \\
\sigma_{u,j}^{2(m)} &= \underset{\boldsymbol{\sigma}_u^2}{\mathrm{argmax}}\,\mathrm{E}[l_2(\boldsymbol{\sigma}_u^2;\mathbf{u}|\mathbf{v}_u)|\mathbf{y}] \\
&= \frac{1}{P_m}\sum_{p=1}^{P_m}\frac{1}{q_j}\left(\sum_{k=1}^{q_j}u_{jk}^{(p)^2}v_{u_{jk}}^{(p)}\right).
\end{aligned}
$$

This step might be termed penalized reweighted least squares, but unlike the t-regression with independent errors, there are no closed forms for the E-step and the resulting weights.

5. We monitor the progress of the estimates over the EM iteration by computing $\max_j \frac{|\theta_j^{(m)}-\theta_j^{(m-1)}|}{s.e.(\theta_j^{(m)})}$. As discussed in Booth and Hobert (1999) and Caffo et al. (2005) though, we caution not to use this quantity as a stopping rule since small relative changes in parameter estimates can lead to unstable estimates and estimated standard errors. That work also proposes more sophisticated and valid stopping rules and automated Monte Carlo EM type algorithms.

6. Predictions of $\mathbf{u}$ are gotten from the mean of the set of Gibbs samples generated for $\mathbf{u}$ at the final iteration of the MCECM algorithm.

3.4 Degree of freedom selection

A useful feature of the t-linear mixed model is the flexibility afforded by the choice of degrees of freedom. The parameters ν_y and ν_u serve the role of robustness parameters and may be estimated from the data as described below.

We estimate the degrees of freedom ν_y and ν_u corresponding to the t-distribution(s) used to fit model (5) via a profile likelihood approach. After computing Monte Carlo estimates of the observed data log likelihood over a grid of ν_y and ν_u values, we select the $\widehat{\nu}$s corresponding to a maximal estimated log likelihood. In order to make this procedure computationally feasible, parameter estimates from the first fit should be used as starting values for fits at neighboring grid elements and so on over the grid. Since the parameter estimates appear to vary relatively slowly with

small changes in degrees of freedom, each fit after the first requires relatively few EM iterations. The approach is outlined below.

(a) Establish a grid of ν values: $\nu_{y,1}, \ldots, \nu_{y,G}, \nu_{u,1}, \ldots, \nu_{u,H}$.

(b) For each $(g,h) \in \{1,\ldots,G\} \times \{1,\ldots,H\}$ do the following:

   i. Fix the degrees of freedom at $\nu_{y,g}$ and $\nu_{u,h}$ and run the complete algorithm for fitting t-linear mixed models outlined in Section 3.3 to get estimates of the other parameters $\widehat{\boldsymbol{\theta}}_{g,h} = [\widehat{\boldsymbol{\beta}}_{g,h}, \widehat{\boldsymbol{\sigma}}^2_{u,g,h}, \widehat{\sigma}^2_{\varepsilon,g,h}]$, as well as the final Gibbs samples $\{\mathbf{u}\}_{p=1}^{P_M}$, $\{\mathbf{v}_y\}_{p=1}^{P_M}$, and $\{\mathbf{v}_u\}_{p=1}^{P_M}$ (from the iteration at which convergence of the complete MCECM algorithm was achieved).

   ii. Compute a Monte Carlo estimate of the conditional expectation of the complete data log likelihood. Lettting $\boldsymbol{\Sigma}^{(p)} = \sigma^2_\varepsilon \text{diag}\left(\frac{1}{\mathbf{v}_y^{(p)}}\right) + \mathbf{Z}\text{blockdiag}\left(\sigma^2_{u,j}\mathbf{I}_{q_j}\frac{1}{\mathbf{v}_{u_j}^{(p)}}\right)\mathbf{Z}^\mathsf{T}$ and
   $l_{1,2}(\boldsymbol{\beta}, \sigma^2_\varepsilon, \boldsymbol{\sigma}^2_u; \mathbf{y}|\mathbf{v}_y^{(p)}, \mathbf{v}_u^{(p)}) = -\frac{1}{2}\log\left|\boldsymbol{\Sigma}^{(p)}\right| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}\left(\boldsymbol{\Sigma}^{(p)}\right)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$,

$$\begin{aligned}\widehat{l}(\widehat{\boldsymbol{\theta}}_{g,h}, \nu_{y,g}, \nu_{u,h}; \mathbf{y}) &= \mathrm{E}\left\{l_{1,2}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2_\varepsilon, \widehat{\boldsymbol{\sigma}}^2_u; \mathbf{y}|\mathbf{v}_y, \mathbf{v}_u)\right\} \\ &\approx \frac{1}{P_m}\sum_{p=1}^{P_m} l_{1,2}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2_\varepsilon, \widehat{\boldsymbol{\sigma}}^2_u; \mathbf{y}|\mathbf{v}_y^{(p)}, \mathbf{v}_u^{(p)})\end{aligned}$$

(c) Choose $\widehat{\nu}_y, \widehat{\nu}_u = \text{argmax}_{\nu_{y,g},\nu_{u,h}}\left\{l(\widehat{\boldsymbol{\theta}}_{g,h}, \nu_{y,g}, \nu_{u,h}; \mathbf{y})\right\}_{g=1,h=1}^{G,H}$.

4 Standard error estimation

Estimates of interest for mixed models often consist of linear combinations of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$. Henderson (1975) argued that for the Gaussian linear mixed model with $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2_\varepsilon\mathbf{I}_n$ and $\text{Cov}(\mathbf{u}) = \sigma^2_u\mathbf{I}_q$, the covariance of these estimates can be estimated with:

$$\text{Cov}\begin{bmatrix}\widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} - \mathbf{u}\end{bmatrix} = \widehat{\sigma}^2_\varepsilon\begin{pmatrix}\mathbf{X}^\mathsf{T}\mathbf{X} & \mathbf{X}^\mathsf{T}\mathbf{Z} \\ \mathbf{Z}^\mathsf{T}\mathbf{X} & \mathbf{Z}^\mathsf{T}\mathbf{Z} + \frac{\widehat{\sigma}^2_\varepsilon}{\widehat{\sigma}^2_u}\mathbf{I}_q\end{pmatrix}^{-1}.$$

This expression is the negative inverse Hessian with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ of the joint Gaussian distribution of $\mathbf{y}$ and $\mathbf{u}$. We use the same method to develop $\text{Cov}\begin{bmatrix}\widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} - \mathbf{u}\end{bmatrix}$ under model (3).

Let $\mathbf{H}(\boldsymbol{\beta}, \mathbf{u}; \mathbf{y}, \mathbf{v}_u, \mathbf{v}_u)$ be the Hessian and $\mathbf{s}(\boldsymbol{\beta}, \mathbf{u}; \mathbf{y}, \mathbf{v}_u, \mathbf{v}_u)$ be the gradient with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ of the complete data log likelihood:

$$-\mathbf{H}(\boldsymbol{\beta}, \mathbf{u}; \mathbf{y}, \mathbf{v}_y, \mathbf{v}_u) =$$

$$\frac{1}{\sigma_\varepsilon^2}\left\{\begin{array}{ll}\mathbf{X}^\mathsf{T}\text{diag}\left(\mathbf{v}_y\right)\mathbf{X} & \mathbf{X}^\mathsf{T}\text{diag}\left(\mathbf{v}_y\right)\mathbf{Z}, \\ \mathbf{Z}^\mathsf{T}\text{diag}\left(\mathbf{v}_y\right)\mathbf{X} & \mathbf{Z}^\mathsf{T}\text{diag}\left(\mathbf{v}_y\right)\mathbf{Z}+\sigma_\varepsilon^2\text{blockdiag}\left(\frac{\text{diag}_{\mathbf{v}_{u_j}}}{\sigma_{u,j}^2}\right)\end{array}\right\},$$

and

$$\mathbf{s}(\boldsymbol{\beta},\mathbf{u};\mathbf{y},\mathbf{v}_y,\mathbf{v}_u)=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^\mathsf{T}\frac{1}{\sigma_\varepsilon^2}\text{diag}\left(\mathbf{v}_y\right)\left(\mathbf{X}\quad\mathbf{Z}\right).$$

Using these expressions and the method described in Louis (1982), a Monte Carlo estimate of the Hessian with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ of the joint distribution of $\mathbf{y}$ and $\mathbf{u}$ is:

$$\text{Cov}\left[\begin{array}{c}\widehat{\boldsymbol{\beta}}\\ \widehat{\mathbf{u}}-\mathbf{u}\end{array}\right]\approx\left[\frac{1}{P_m}\sum_{p=1}^{P_m}\left\{-\mathbf{H}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}};\mathbf{y},\mathbf{v}_{y_p},\mathbf{v}_{u_p})-\right.\right.$$
$$\left.\left.\mathbf{s}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}};\mathbf{y},\mathbf{v}_{y_p},\mathbf{v}_{u_p})^\mathsf{T}\mathbf{s}(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}};\mathbf{y},\mathbf{v}_{y_p},\mathbf{v}_{u_p})\right\}\right]^{-1} \tag{11}$$

where $\{\mathbf{v}_{y_p}\}$ and $\{\mathbf{v}_{u_p}\}$ are samples from the Gibbs iterations in the final ECM step.

It has been pointed out (e.g. Booth and Hobert 1998) that these sampling variances are approximate since they do not account for the fact that the variance components are estimated. Similar methods (weighted averages of Gaussian based covariances) also could be used to obtain a covariance matrix for estimates of the variance components.

5 Applications

The robust mixed model (5) can be employed in the analysis of data arising in many contexts. We provide two examples that initially motivated this paper. A set of R functions tailored to fitting robust general design mixed models using the t-distribution has been developed by the authors.

5.1 Nonparametric regression

Our first illustration of model (5) involves nonparametric regression (also known as scatterplot smoothing). Nonparametric regression is commonly used to highlight an underlying trend without assuming a function form for the trend. Many scatterplot smoothers exist. Smith and Kohn (1996) present an effective means for robust nonparametric regression modeling within the Bayesian paradigm (along with a software module for implementation). However, their focus is not on parametric mixed models. We instead focus on penalized regression splines (e.g. Ruppert, Wand and Carroll 2003) in a mixed model framework.

Suppose that (x_i, y_i), $i = 1,\ldots,n$, represents measurements on a predictor x and a response variable y. The nonparametric regression model for these data is

$$y_i = f(x_i) + \varepsilon_i \tag{12}$$

where f is a smooth, but otherwise unspecified, function of x. A truncated polynomial spline version of $f(x_i)$ is then

$$f(x_i) = \beta_0 + \beta_x x_i + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ \tag{13}$$

where $\kappa_1, \ldots, \kappa_K$ are knots in the x direction and $(x)_+$ returns x if $x > 0$ and 0 otherwise. If we define

$$\boldsymbol{\beta} = [\beta_0, \beta_1]^\mathsf{T}, \quad \mathbf{u} = [u_1, \ldots, u_K]^\mathsf{T},$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+ & \ldots & (x_1 - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ (x_n - \kappa_1)_+ & \ldots & (x_n - \kappa_K)_+ \end{bmatrix},$$

then we achieve best linear unbiased prediction in the mixed model (3) framework.

We use nonparametric regression to analyze data from a laboratory experiment which gathered data on participants' expiratory flow characteristics when exposed to filtered air. Due to instrumentation error, and an occasional cough or sporadic breath of a subject undergoing evaluation, outlying observations were recorded. Robust fits are therefore preferred as they more accurately convey the average stimulus-response signal under outlier contamination. We present the results of a t-linear mixed model fit.

Nonparametric regression fits for one subject during two separate filtered air experiments appear in Figure 1. Each panel displays log adjusted response, $\mathbf{y}$, (the log of the subject's response for the experiment minus her mean response at baseline prior to the experiment) versus time in seconds, $\mathbf{x}$. Robust t-model fits appear as solid lines and Gaussian model fits appear as dotted lines. In both experiments the t-model fits appear robust to outliers. The degrees of freedom (ν_y) estimated for the t-distributions used to model the response are 3.4, and 5.2 for the panels from left to right. We used a Gaussian assumption for the random effects, $\mathbf{u}$.

The chain was run for 100 EM iterations with the Gaussian model estimates are starting values. The initial P_m was 10, and it was increased by $P_{m-1}/5$ (rounded up) every fifth iteration ($P_{100} = 1107$). The maximum change in the spline fit over a grid of 100 points (evenly distributed over the range of the x_is) was less than 5% of the corresponding standard error for the last five EM iterations. One hundred EM iterations required about 3 minutes on a 2.16 GHz Intel Core Duo Mac Book Pro. In order to assess the progress of the EM algorithm, we also ran the chain for 175 EM iterations which required about 18 hours ($P_{175} = 29519$). Figure 2 shows the progress of the estimate of $f(1250)$ and its estimated standard error for the exhalation measurement at 1250 seconds. (We chose $x = 1250$ because that is where the starting value for the estimate was furthest from the final estimate.) The first column in the figure shows the estimates using a y-axis range that is determined by the range of

estimates over all EM iterations, and the second column uses a y-axis range that is determined by the approximate range of the estimated $f()$. In other words, the second column shows the progress on the scale at which the estimate will be used. As a result, the figure suggests that 100 iterations is sufficient for the purpose of plotting the estimated spline.

5.2 Semi-parametric repeated measures model

Our second example provides a semi-parametric treatment of a repeated measures experiment. Let i index subject ($1 \leqslant i \leqslant n$) and j index observation within subject ($1 \leqslant j \leqslant m$). Suppose the response y_{ij} is measured at time t_{ij}. We model y_{ij} as the sum of three components, an additive subject specific random offset, a smooth function of time, and a within person random error:

$$y_{ij} = u_i + f(t_{ij}) + \varepsilon_{ij}.$$

The coefficients are

$$\boldsymbol{\beta} = [\beta_0, \beta_t]^\mathsf{T}, \quad \mathbf{u} = [u_1, \ldots, u_n, u_1^t, \ldots, u_{K_t}^t]^\mathsf{T},$$

and the design matrices are

$$\mathbf{X} = \begin{bmatrix} 1 & t_{11} \\ \vdots & \vdots \\ 1 & t_{ij} \\ \vdots & \vdots \\ 1 & t_{nm} \end{bmatrix}, \quad \mathbf{Z} = [\mathbf{1}_m \otimes \mathbf{I}_n \quad \mathbf{Z}_{spl} \otimes \mathbf{1}_n]$$

where $\mathbf{Z}_{spl} = [(t_{ij} - \kappa_k)_+]$, as used in the previous example.

A motivating application comes from the application of oral glucose tolerance tests to ten individuals. These data were collected in the Energy Metabolism Laboratory at the University of Massachusetts, Amherst as preliminary data to Hagobian and Braun (2006). After an overnight fast, the ten subjects were each fed 75 grams of glucose, and blood samples were then taken via catheter 0, 30, 60, 90, and 120 minutes later. After data collection, the amount of glucose in each blood sample was measured.

We use a robust t-linear mixed model to model the mean of glucose over time. The degrees of freedom (ν_y) for the t-distribution used in the robust fit were estimated to be 3.1, and we used Gaussian random effects $\mathbf{u}$. Scatterplots of the data from all ten subjects and the estimated spline are displayed in Figure 2. A standard Gaussian mixed model fit is also shown for comparison. This example illustrates that the t-linear mixed model appears to be robust to the two apparent outliers. In the

actual data analysis, it may be important to investigate why outliers occur as well. In this case, it was due to a data transcription error.

Again, the starting values came from the Gaussian model, and the other computational parameters were the same for this example as for the last. In this case, the maximum change in the spline fit over a grid of 100 points was less than 7% of the corresponding standard error for the last five EM iterations, and the computations required about 4 minutes on a 2.16 GHz Intel Core Duo Mac Book Pro. Again, the estimated spline and pointwise confidence interval in Figure 3 appeared identical up to the plotting resolution over the last five iterations of the EM algorithm. Similarly to Figure 2, Figure 4 shows the progress of the estimated $f(30)$ and its standard error when the algorithm is run for 175 EM iterations over 24 hours.

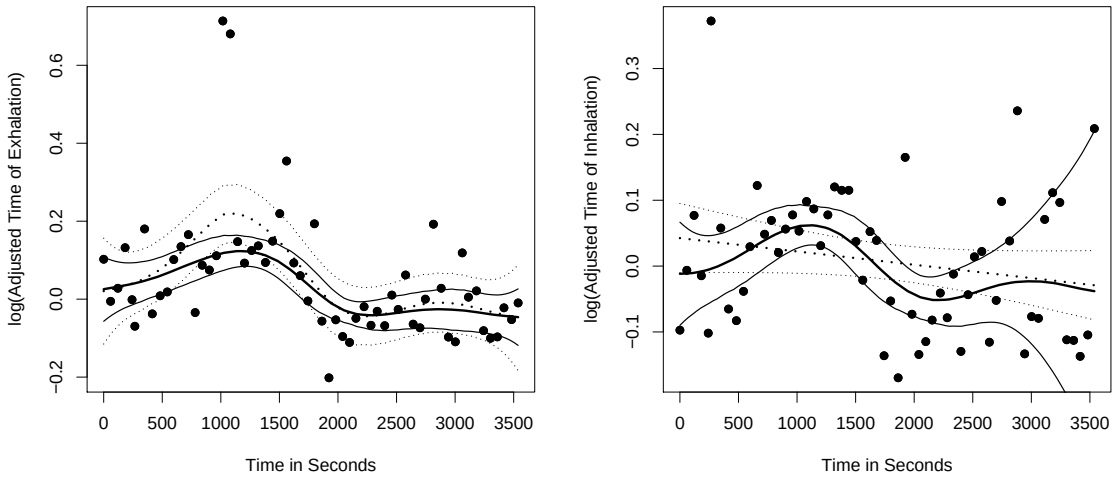

Figure 1: *Non-parametric model fits applied to respiratory data from one subject in the experiment described in the text. Each panel displays adjusted response over time in seconds. Solid lines correspond to the robust t-model fits and $\pm$ 1.96 (pointwise) standard errors; dotted lines are Gaussian model fits.*

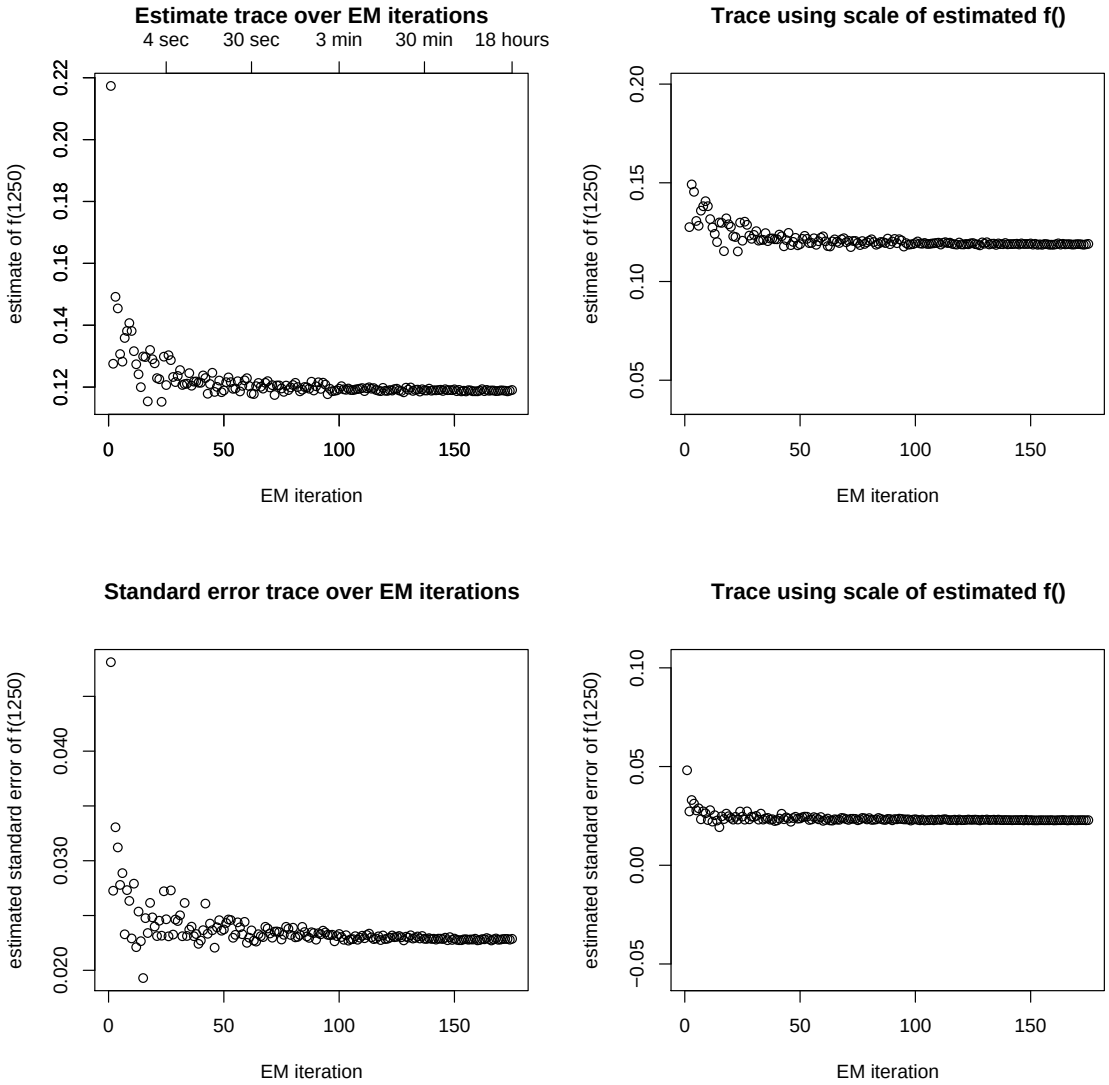

Figure 2: *Estimates of $f()$ and its standard error plotted as a function of EM iteration for the exhalation data. The left column of panels uses a y-axis scale determined by the range of the estimates over the EM iterations. The right column shows the estimates on the scale at which they will be used; the y-axis range is determined by the range of the estimated $f()$. The top axis of the upper left panel shows the approximate cumulative computing time.*

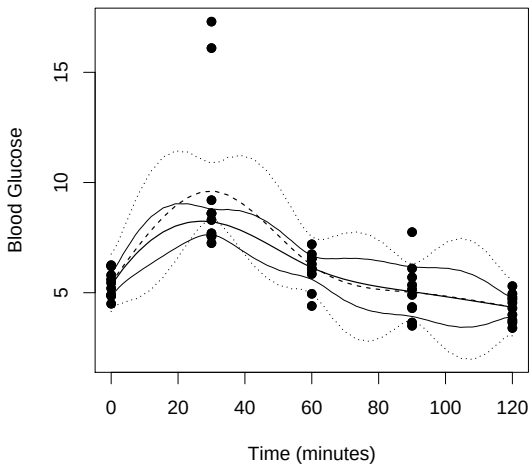

Figure 3: *Semi-parametric mixed model fits applied to the glucose tolerance test data from ten subjects described in the text. Solid lines show the estimated spline part of the model using robust t-model, and dotted lines show the fit using a fully Gaussian model.*

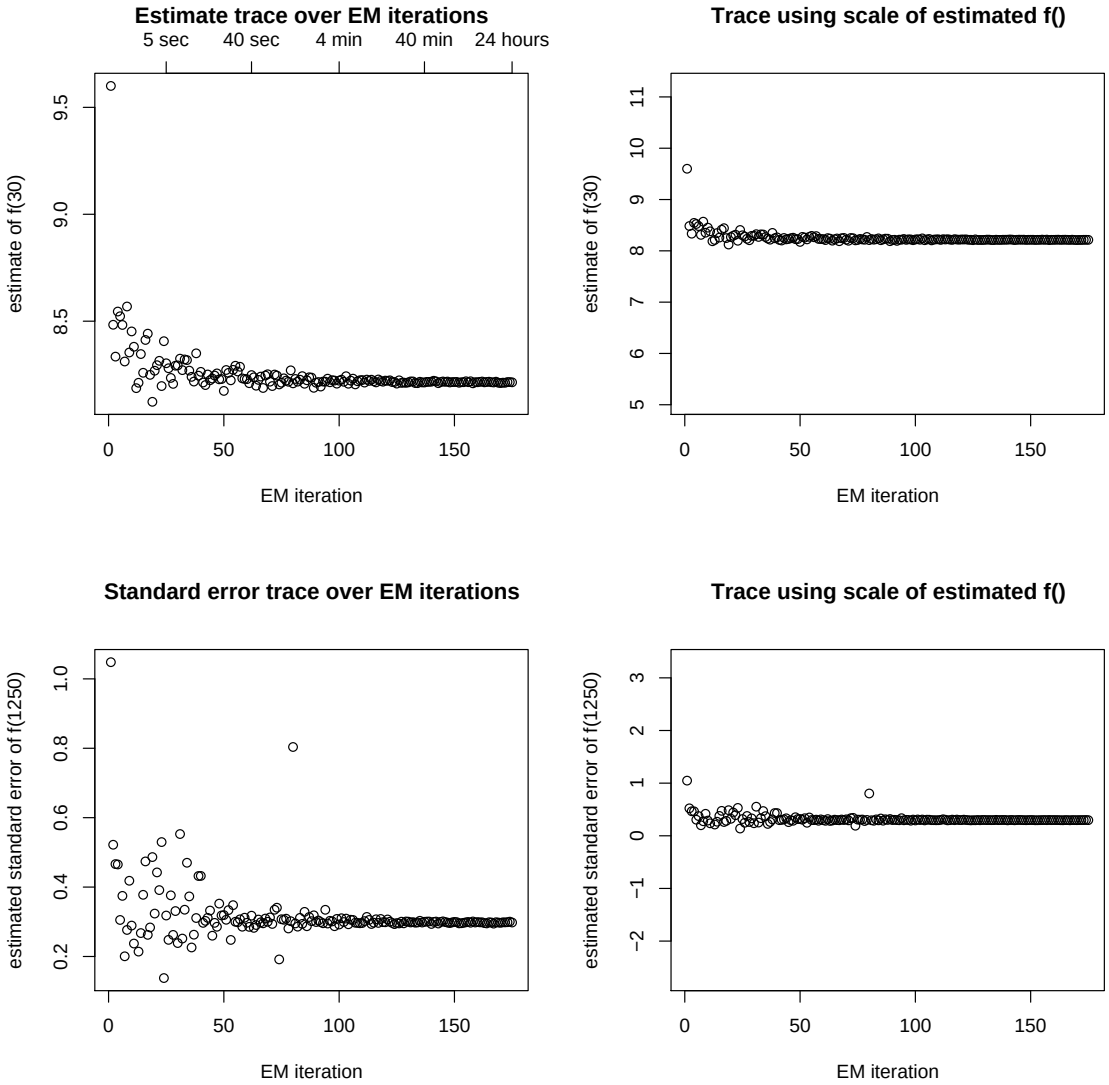

Figure 4: *Estimates of $f()$ and its standard error plotted as a function of EM iteration for the glucose tolerance test data. The left column of panels uses a y-axis scale determined by the range of the estimates over the EM iterations. The right column shows the estimates on the scale at which they will be used; the y-axis range is determined by the range of the estimated $f()$. The top axis of the upper left panel shows the approximate cumulative computing time.*

6 Closing remarks

The general design robust mixed model (5) emerges as an effective approach to modeling data in the presence of outlying observations. The t-model formulation builds in resistance to outliers while retaining the elegance and simplicity of parametric, likelihood based inference and best prediction. Monte Carlo computation is required, but is straightforward to implement (e.g. simple Gibbs sampling from standard distributions).

Model (5) treats the most general matrix structure in (3) and thus accommodates many more situations than do existing robust mixed model methods. Such situations include classical longitudinal settings with multiple hierarchies, spatial statistics, smoothing splines, and crossed random effect designs.

This paper also presents a means for estimating robustness parameters (e.g. degrees of freedom ν_y, ν_u). Such estimates determine the t-distributions used, which ultimately correspond to the robustness of the fit. In addition, large sample estimates of variability are a by-product of the Monte Carlo Expectation Conditional Maximization approach we take.

Of course, there are other methods for robust fitting, but we choose the t-model for its conceptual simplicity, and relatively straightforward implementation. Several simple modifications to the EM aspect of the algorithm may increase its speed, although we found no need for these in our examples. van Dyk (2000) suggests a nesting algorithm and other alternatives such as parameter expansion (Liu, Rubin, and Wu 1998). As a referee pointed out, a Bayesian implementation of our model could be straightforward using the Winbugs software. Winbugs approaches to Gaussian linear mixed models and generalized linear mixed models are described (with code) in Crainiceanu, Ruppert, and Wand (2005) and Zhao et al. (2006) for instance.

Acknowledgments

We are grateful to Russ Hauser for making the filtered air data available for analysis. We would also like to thank Todd Hagobian and Barry Braun for providing the glucose tolerance test data. This paper has was greatly improved by numerous detailed comments from two referees and the Editor. We also benefited from discussions with Brent Coull. This was partially supported by US National Institutes of Health grants T32 ES07142-18 and R01-ES10844-01 and US National Science Foundation grant DMS 0306227.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117128.

Breusch, T. S., Robertson, J. C., and Welsh, A. H. (1997). The emperor's new clothes: a critique of the multivariate t regression model. *Statistica Neerlandica*, **51**, 269–286.

Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistics Society, Series B* , **61**, 265–285.

Caffo, B. S., Jank, W., and Jones, G. L. (2005). Ascent-based Monte Carlo expectation maximization. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **67**, 235251.

Crainiceanu, C., Ruppert, D. and Wand, M.P. (2005). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software*, **14**.

Dempster, A.P. and Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society, Series B* , **39**, 1–22.

Ghidey, W., Lesaffre, E. and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, **60**, 945-953.

Hagobian, T.A. and Braun, B. (2006) Interactions between energy surplus and short-term exercise on glucose and insulin responses in healthy people with induced, mild insulin insensitivity. *Metabolism*, **55**, 402-408.

Hastie, T.J. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.

Jara, A. and Quintana, F. (2006). Linear mixed models with skew-elliptical distributions: A Bayesian approach. *Unpublished manuscript.*

Kammann, E.E and Wand, M.P. (2003). Geoadditive models. To appear in *Applied Statistics*.

Kleinman, K. and Ibrahim, J.G. (1998). A semi-parametric Bayesian approach to the random effects model. *Biometrics*, **54**, 921-938.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge: Cambridge University Press.

Laird, N.L. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989). Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, **84**, 881-896.

Levine, R. A. and Casella, G. (2001). Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, **10**, 422–439.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, **61**, 381–400.

Liu, C. and Rubin. D.B. and Wu, Y.N. (1998). Parameter expansion to accelerate EM: The PX-EX algorithm. *Biometrika*, **85**, 775-770.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226–233.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, linear, and mixed models*. London: Chapman and Hall.

Meng, X.L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

O'Connell, M.A. and Wolfinger, R.D. (1997). Spatial regression models, response surfaces, and process optimization. *Journal of Computational and Graphical Statistics*, **6**, 224–241.

Pinheiro, J.C, Liu, C. and Wu, Y.N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution. *Journal of Computational and Graphical Statistics*, **10**, 249–276.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–51.

Rosa, G.J.M., Gianola, D. and Padovani, C.R. (2004). Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. *Journal of Applied Statistics*, **31**, 855-873.

Rosa, G.J.M., Padovani, C.R. and Gianola, D. (2003). Robust linear mixed models with normal / independent distributions and Bayesian MCMC implementation. *Biometrical Journal*, **45**, 573-590.

Ruppert, D., Wand M.P. and Carroll, R.J. (2003) *Semiparametric Regression Modelling*. Cambridge: Cambridge University Press.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics*, **75**, 317–344.

Speed, T. (1991). Comment on paper by Robinson. *Statistical Science*, **6**, 42–44.

Stranden, I. and Gianola, D. (1998). Mixed effects linear models with t- distributions for quantitative genetic analysis: a Bayesian approach. *Genetics, Selection, Evolution*, **31**, 25-42.

Tao, H., Palta, M., Yandel, B. and Newton, M.A. (1999). An estimation method for the semiparametric mixed effects model. *Biometrics*, **55**, 102-110.

van Dyk, D.A. (2000). Nesting EM algorithms for computational efficiency. *Statistica Sinica*, **10**, 203–225.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B*, **40**, 364–372.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, **85**, 699–704.

Welsh, A.H. and Richardson, A.M. (1997). Approaches to the robust estimation of mixed models. In *Handbook of Statistics, Volume 15* (G.S. Maddala and C.R. Rao, eds.), New York: Elsevier Science B.V.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**, 95–103.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**, 795-802.

Zhao, Y., Staudenmayer, J., Coull, B.A. and Wand, M.P. (2006). General Design Bayesian Generalized Linear Mixed Models. *Statistical Science*, **21**, 35-51.

Appendix

For completeness, this appendix extends the algorithm to the linear mixed model with the variance component matrices $\mathbf{G} = \text{blockdiag}\mathbf{G}_j$ and $\mathbf{R}$, rather than diagonal matrices.

Model

The general formulation of the t-linear mixed model is:

$$\begin{aligned}
\tilde{\varepsilon}_i &\overset{\text{i.i.d.}}{\sim} t(0,1,\nu_y), 1 \leqslant i \leqslant n \\
\widetilde{u}_{jk} &\overset{\text{i.i.d.}}{\sim} t(0,1,\nu_u), 1 \leqslant j \leqslant c, 1 \leqslant k \leqslant q_j \\
\mathbf{u}_j &= \mathbf{G}_j^{1/2}\widetilde{\mathbf{u}}_j, 1 \leqslant j \leqslant c \\
\mathbf{y}|\mathbf{u} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \sigma_\varepsilon \widetilde{\mathbf{R}}^{1/2}\tilde{\varepsilon} \\
&:= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{R}^{1/2}\tilde{\varepsilon}.
\end{aligned}$$

Results and research regarding linear combinations of iid t random variables are discussed and reviewed in Kotz and Nadarajah (2004).

Estimation

The complete data likelihood is: $l_{comp}(\boldsymbol{\theta}) = l(\boldsymbol{\beta}, \mathbf{R}, \mathbf{G}, \nu_y, \nu_u; \mathbf{y}, \mathbf{u}, \mathbf{v}_y, \mathbf{v}_u)$ which equals

$$l_1(\boldsymbol{\beta}, \mathbf{R}; \mathbf{y}|\mathbf{u}, \mathbf{v}_y) + l_2(\mathbf{G}; \mathbf{u}|\mathbf{v}_u) + l_3(\nu_y; \mathbf{v}_y) + l_4(\nu_u; \mathbf{v}_u).$$

Let $\mathbf{V}_y = \text{diag}(\mathbf{v}_y)$ and $\mathbf{V}_u = \text{diag}(\mathbf{v}_u)$. The last two components, $l_3(\nu_y; \mathbf{v}_y)$ and $l_4(\nu_u; \mathbf{v}_u)$ are as in Section 3.1. The first two components are:

$$\begin{aligned}
l_1(\boldsymbol{\beta}, \mathbf{R}; \mathbf{y}|\mathbf{u}, \mathbf{v}_y) = &-\frac{1}{2}\log\left|\mathbf{R}^{1/2}\mathbf{V}_y^{-1}\mathbf{R}^{\mathsf{T}/2}\right| \\
&-\frac{1}{2}\left\{\mathbf{y} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})\right\}^{\mathsf{T}}\mathbf{R}^{-\mathsf{T}/2}\mathbf{V}_y\mathbf{R}^{-1/2}\left\{\mathbf{y} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})\right\}
\end{aligned}$$

$$\text{and } l_2(\mathbf{G}; \mathbf{u}|\mathbf{v}_u) = -\frac{1}{2}\log\left|\mathbf{G}^{1/2}\mathbf{V}_u^{-1}\mathbf{G}^{\mathsf{T}/2}\right| - \frac{1}{2}\mathbf{u}^{\mathsf{T}}\mathbf{G}^{-\mathsf{T}/2}\mathbf{V}_u\mathbf{G}^{-1/2}.$$

The required sampling distributions are

$$\mathbf{u}|\mathbf{y}, \mathbf{v}_y, \mathbf{v}_u \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where,

$$\boldsymbol{\mu} = \mathbf{A}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right), \quad \boldsymbol{\Sigma} = \mathbf{G}^{1/2}\mathbf{V}_u\mathbf{G}^{\mathsf{T}/2} - \mathbf{A}\mathbf{Z}\mathbf{G}^{1/2}\mathbf{V}_u\mathbf{G}^{\mathsf{T}/2}$$

with

$$\mathbf{A} = \mathbf{G}^{1/2}\mathbf{V}_u\mathbf{G}^{\mathsf{T}/2}\left\{\mathbf{R}^{1/2}\mathbf{V}_y\mathbf{R}^{\mathsf{T}/2} + \mathbf{Z}\mathbf{G}^{1/2}\mathbf{V}_u\mathbf{G}^{\mathsf{T}/2}\mathbf{Z}^{\mathsf{T}}\right\}^{-1}.$$

Further, let

$$[r_1, \ldots, r_n] = \left\{\mathbf{y} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})\right\}^{\mathsf{T}}\mathbf{R}^{-\mathsf{T}/2},$$

and

$$[s_{1,1}, \ldots, s_{cq_c}] = \mathbf{u}^{\mathsf{T}}\mathbf{G}^{-\mathsf{T}/2}.$$

With these definitions,

$$v_{y_i}|\mathbf{y}, \mathbf{u} \overset{\text{ind.}}{\sim} \frac{\chi^2_{\nu_y+1}}{(\nu_y + r_i^2)}, \text{ and } v_{u_{jk}}|\mathbf{u} \overset{\text{ind.}}{\sim} \frac{\chi^2_{\nu_u+1}}{(\nu_u + s_{jk}^2)}.$$

Finally, the ensuing parameter updates will depend on the specific form of the variance components, but in general, they are:

$$\begin{aligned}
\boldsymbol{\beta}^{(m)} &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \mathrm{E}\left[l_1(\boldsymbol{\beta}, \mathbf{R}^{(m-1)}; \mathbf{y}|\mathbf{u}, \mathbf{v}_y)|\mathbf{y}\right], \\
\mathbf{R}^{(m)} &= \underset{\mathbf{R}}{\operatorname{argmax}}\, \mathrm{E}\left[l_1(\boldsymbol{\beta}^{(m)}, \mathbf{R}; \mathbf{y}|\mathbf{u}, \mathbf{v}_y)|\mathbf{y}\right], \quad \text{and} \\
\mathbf{G}^{(m)} &= \underset{\mathbf{G}}{\operatorname{argmax}}\, \mathrm{E}[l_2(\mathbf{G}; \mathbf{u}|\mathbf{v}_u)|\mathbf{y}].
\end{aligned}$$

Again, note that in the last two expressions the maximizations are over the variance components in the two matrices. Completely unstructured matrices would not be identifiable for every sample size.

Standard Errors

For the general case,

$$-\mathbf{H}(\boldsymbol{\beta}, \mathbf{u}; \mathbf{y}, \mathbf{v}_y, \mathbf{v}_u) =$$

$$\sigma_\varepsilon^2 \left\{ \begin{array}{ll} \mathbf{X}^{\mathsf{T}}\widetilde{\mathbf{R}}^{-\mathsf{T}/2}\mathbf{V}_y\widetilde{\mathbf{R}}^{-1/2}\mathbf{X} & \mathbf{X}^{\mathsf{T}}\widetilde{\mathbf{R}}^{-\mathsf{T}/2}\mathbf{V}_y\widetilde{\mathbf{R}}^{-1/2}\mathbf{Z}, \\ \mathbf{Z}^{\mathsf{T}}\widetilde{\mathbf{R}}^{-\mathsf{T}/2}\mathbf{V}_y\widetilde{\mathbf{R}}^{-1/2}\mathbf{X} & \mathbf{Z}^{\mathsf{T}}\widetilde{\mathbf{R}}^{-\mathsf{T}/2}\mathbf{V}_y\widetilde{\mathbf{R}}^{-1/2}\mathbf{Z} + \sigma_\varepsilon^2\mathbf{G}^{-\mathsf{T}/2}\mathbf{V}_u\mathbf{G}^{-1/2} \end{array} \right\},$$

and

$$\mathbf{s}(\boldsymbol{\beta}, \mathbf{u}; \mathbf{y}, \mathbf{v}_y, \mathbf{v}_u) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^{\mathsf{T}}\mathbf{R}^{-\mathsf{T}/2}\mathbf{V}_y\mathbf{R}^{-1/2}[\mathbf{X} \quad \mathbf{Z}].$$