A General Skew-t Mixed Model That Allows Different Degrees of Freedom for Random Effects and Error Distributions

Pankaj K. Choudhary[1], **Dishari Sengupta**

Department of Mathematical Sciences, FO 35

University of Texas at Dallas

Richardson, TX 75083-0688, USA

Phillip Cassey

School of Earth & Environmental Sciences

G41a Mawson Laboratories

University of Adelaide

North Terrace, SA 5005, Australia

Abstract

This article develops a robust mixed model that assumes a multivariate skew-t distribution for random effects and an independent multivariate t-distribution for errors. It simultaneously captures skewness and heavy tailedness in data, while allowing the random effects and error distributions to have different degrees of freedom. It is fit using an EM-type algorithm. Simulations show that its efficiency for estimating mean response is comparable to that of the recent skew-t mixed model. But it may be considerably more efficient than the latter for estimating variance-covariance parameters when at least one of the random effects distribution or the error distribution has heavy tails, possibly due to outliers. The proposed model is used to analyze a data set consisting of lengths of claws of fiddler crabs (*Uca mjoebergi*).

Keywords: EM algorithm, heavy tailed distribution, method comparison, outlier, robust mixed-effects model, skew-t distribution.

1 Introduction

Linear mixed models are routinely used for analyzing a variety of dependent data, including longitudinal data, repeated measurements data and method comparison data. The popularity of these

[1]Corresponding author. Email: pankaj@utdallas.edu, Tel: +1-972-883-4436, Fax: +1-972-883-6622.

models is primarily due to the fact that they offer considerable flexibility in modeling of within-subject dependence in the data while remaining mathematically tractable and computationally efficient (see, e.g., Pinheiro and Bates, 2000; Jiang, 2007). Besides all major statistical software such as `R` (R Development Core Team, 2012), `SAS` (SAS Institute Inc.) and `SPSS` (IBM Corp.) provide the capability to fit such models, facilitating their usage.

A linear mixed model is generally formulated as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, \ldots, m, \tag{1}$$

where i is the subject index; $\mathbf{Y}_i$ is the n_i-vector of observed responses on the ith subject; $\boldsymbol{\beta}$ is the p-vector of fixed effects with $\mathbf{X}_i$ as the corresponding $n_i \times p$ design matrix; $\mathbf{b}_i$ is the q-vector of random effects with $\mathbf{Z}_i$ as the corresponding $n_i \times q$ design matrix; and $\mathbf{e}_i$ is the n_i-vector of within-subject random errors. The matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ have full column ranks. All the vectors here and elsewhere in this article are column vectors unless specified otherwise.

The standard version of this model assumes that

$$\mathbf{b}_i \sim \text{ independent } \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi}), \quad \mathbf{e}_i \sim \text{ independent } \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i), \quad i = 1, \ldots, m, \tag{2}$$

and the two are mutually independent. Here the $q \times q$ matrix $\boldsymbol{\Psi}$ and $n_i \times n_i$ matrix $\boldsymbol{\Sigma}_i$ are non-singular covariance matrices. The matrix $\boldsymbol{\Psi}$ may be unstructured or structured, but $\boldsymbol{\Sigma}_i$ is generally parameterized in terms of a small number of parameters that do not change with i. We refer to model (1) together with the normality assumption (2) as the *normal mixed model* (NMM). Its parameters are generally estimated using a maximum likelihood (ML) method and the large sample theory is used for testing hypotheses and constructing confidence intervals. See Pinheiro and Bates (2000) and Jiang (2007, 2013) for accounts of theory, applications and computations involving these models.

The assumption of normality in (2) is frequently violated in practice. In particular, there may be skewness or heavy tailedness (meaning: tails heavier than those of a normal distribution) in the distribution of either the random effects or the errors, causing skewness or heavy tailedness in the observed data. The tails are often heavy due to outliers which may occur either in the random effects (**b**-outliers) or in the errors (**e**-outliers; Pinheiro, Liu and Wu, 2001). Although

the likelihood-based estimates of fixed effects may be robust to non-normality of random effects (Butler and Louis, 1992), the same is not true for estimates of random effects, which may behave badly if the normality does not hold (Zhang and Davidian, 2001). Moreover, if outliers are present in the data even in moderate amounts, the ML estimates of all parameters suffer from loss of efficiency (Pinheiro et al., 2001). This loss is especially severe for estimates of variance-covariance parameters, which are needed to compute standard errors of the fixed effects estimates. These parameters may also be of interest in their own right, e.g., as in method comparison studies (Choudhary and Yin, 2010) — an application of interest in this article. If the normality assumption is violated, it may be possible make it tenable by transforming the data. But a transformation is not always successful. Besides the transformed data may be difficult to interpret. This is especially a concern in method comparison studies where a transformation other than the log is generally not recommended (Bland and Altman, 1999).

In any case, a viable alternative to NMM is a *robust mixed model* whose development has received considerable attention in recent years. Four approaches for robust modeling have become especially popular. The first is a semiparametric approach where only the first two moments of the response vector are modeled using generalized estimating equations (Liang and Zeger, 1986). The second approach bounds the influence of outlying observations on parameter estimates by obtaining them as solutions of appropriately defined estimating equations (Richardson, 1997; Stahel and Welsh, 1997). The third is a fully parametric approach that replaces the assumption of normality for random effects and/or errors with a more general distribution that has normality as a special case — e.g., a mixture of normals, a t, a skew-normal (Azzalini, 1985) or a skew-t distribution (Azzalini and Capitanio, 2003). This approach is in the spirit of Box (1980) and Lange, Little and Taylor (1989) as it embeds the normal model in a larger model with additional parameters for accommodating non-normality. The fourth approach uses only weak distributional assumptions about random effects (Jiang, 1999; Wang, Tsai and Qu, 2012). The reader interested in a comparison of some of these approaches is referred to Heritier et al. (2009, ch. 4).

This article focuses on the third approach, i.e., robust mixed models of the form (1) that are flexible enough to incorporate skewness and heavy tailedness in the data. It has been considered

by several authors. Specifically, Verbeke and Lesaffre (1996) assume a finite mixture of normals as the distribution of random effects. Zhang and Davidian (2001) approximate the random effects density by a seminonparametric representation. Pinheiro et al. (2001) assume a joint multivariate t-distribution for random effects and errors. This model is called the *t mixed model* (TMM). Arellano-Valle, Bolfarine and Lachos (2005) assume independent skew-normal distributions for random effects and errors. Lachos, Ghosh and Arellano-Valle (2010) assume a joint multivariate skew-normal/independent distribution (Branco and Dey, 2001) for random effects and errors, with skewness only in random effects. Ho and Lin (2010) study a special case of this model called the *skew-t mixed model* (STMM), wherein a joint multivariate skew-t distribution is assumed for random effects and errors, and skewness is incorporated only in random effects. These models are generally fit using variants of the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977; McLachlan and Krishnan, 2007). An alternative algorithm is given by Song, Zhang and Qu (2007). See also Verdinelli and Wasserman (1991) and Rosa, Gianola and Padovani (2004) for Bayesian approaches to outlier problems and robust mixed models.

Among the aforementioned robust mixed models, the STMM is an especially attractive choice as it simultaneously captures the effects of skewness and heavy tailedness in data, while remaining computationally efficient. Moreover, the NMM and TMM are its special cases. However, despite the flexibility offered by the STMM, its applicability is limited by the drawback that the degrees of freedom of the assumed t-distributions for random effects and errors must be the *same*. Thus, this model cannot accommodate different levels of heaviness in the tails of random effects and error distributions. This phenomenon arises, e.g., when only the errors have heavy tails but not the random effects. It is observed in the crab claws data introduced later in this section, which motivated this work. In this article, we propose a mixed model that assumes a multivariate skew-t distribution for random effects and an *independent* multivariate t-distribution for errors. The independence of these two distributions allows their degrees of freedom to be different, thereby overcoming the foregoing limitation of STMM and broadening the scope of mixed models.

This article is organized as follows. Section 2 presents the proposed model and describes an EM-type algorithm for fitting it. Section 3 summarizes a simulation study to compare estimators

based on GSTMM and STMM. Section 4 illustrates an application by analyzing the crab data. Section 5 concludes with a discussion. The appendices contain technical details. We use the software `R` (R Development Core Team, 2012) for all statistical computations in this article.

The motivating example: Crab claws data

The claws of fiddler crabs are lost in fighting and their length is an important component of their size and strength (Lailvaux, Reaney and Backwell, 2009). As part of a study comparing the biochemistry and physical characteristics of original and regenerated claws, the laboratory of the last author measured the lengths (in millimeters) of 25 fiddler crab claws. Every claw was measured three times by each of three observers using two Mitutoyo vernier calipers (an older Dial set and a digital Digimatic set). Thus, each observer takes six measurements on every claw — 3 from caliper 1 and 3 from caliper 2. The measurements are taken in a random order. There is a total of $25 \times 3 \times 6 = 450$ observations in the data.

Although the main goal of this method comparison study was to compare the extent of agreement between the two calipers for each observer (Sengupta, 2012), here we restrict our attention to examining separately for each observer whether the means and error variances of the two calipers are same. Figure 1 displays a trellis plot of the data. The measurements from the two calipers largely overlap, indicating similar means and variances between them for all observers. We return this issue in Section 4 after finding an adequate model for these data.

Following a preliminary data analysis, we adopt a mixed model of the form

$$Y_{ijkl} = \beta_{jl} + b_{ij} + e_{ijkl},\ i = 1, \ldots, 25,\ j = 1, 2,\ k = 1, 2, 3,\ l = 1, 2, 3, \tag{3}$$

where Y_{ijkl} is the kth repeated measurement of the length of the ith claw, taken by the lth observer using the jth caliper; β_{jl} is the fixed intercept associated with the combination of jth caliper and lth observer; b_{ij} is the random effect of ith specimen on jth caliper; and e_{ijkl} is the random error term. This model can be written in the form (1) by taking

$$\mathbf{Y}_i = (Y_{i111}, Y_{i121}, Y_{i131}, Y_{i112}, Y_{i122}, Y_{i132}, \ldots, Y_{i213}, Y_{i223}, Y_{i233})$$

as the vector of $n_i = 18$ observations on the ith claw, $\boldsymbol{\beta}$ as the vector of $p = 6$ elements $(\beta_{11}, \beta_{12}, \ldots, \beta_{23})$, $\mathbf{b}_i$ as the vector of $q = 2$ elements (b_{i1}, b_{i2}), and defining $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{e}_i$ in a conformable manner. This model is initially fit assuming the usual normality (2) with $\boldsymbol{\Psi}$ as an unstructured matrix and $\boldsymbol{\Sigma}$ as a diagonal matrix,

$$\boldsymbol{\Sigma} = \text{diag}\left\{\sigma^2_{11}, \sigma^2_{11}, \sigma^2_{11}, \sigma^2_{12}, \sigma^2_{12}, \sigma^2_{12}, \ldots, \sigma^2_{23}, \sigma^2_{23}, \sigma^2_{23}\right\}. \tag{4}$$

The last assumption is equivalent to assuming that $e_{ijkl} \sim$ independent $\mathcal{N}(0, \sigma^2_{jl})$. The $\boldsymbol{\Sigma}$ matrix does not depend on i due to balanced design of the data. This model allows each observer $\times$ caliper combination to have its own population mean and variance, which is typical for a method comparison study (Choudhary and Yin, 2010).

This NMM is fit by the ML method using the `nlme` package (Pinheiro et al., 2012) in `R`. The parameter estimates are given in Table 5. Figure 2 shows the resulting normal quantile-quantile (QQ) plots of the predicted b_{i1} and b_{i2} values and the standardized residuals. Also shown is a histogram of the residuals. These graphs suggest that the normality assumption is reasonable for the random effects, whereas a heavy tailed distribution is needed for the errors. There are also four outliers in the data and the tails of the error distribution appear heavier than normal even upon ignoring them. These features of the crab data justify the need for a robust mixed model that allows the random effects and error distributions to differ in heaviness of tails. We model these data in Section 4 using a special case of the proposed model that assumes normality for random effects and a t-distribution for errors.

2 The general skew-t mixed model

2.1 Preliminaries

Let $f(\mathbf{y}|\boldsymbol{\theta})$ denote the probability density function of a random quantity $\mathbf{Y}$ with parameter $\boldsymbol{\theta}$. Also, let $\mathcal{G}(\alpha, \beta)$ denote a gamma distribution with parameters $\alpha, \beta > 0$, and density

$$f(y|\alpha, \beta) = \frac{\beta^\alpha}{g(\alpha)} y^{\alpha-1} \exp(-\beta y), \quad y > 0,$$

where $g(\cdot)$ is the gamma function. We use $\mathcal{TN}(\mu, \sigma^2; (a, b))$ to denote a $\mathcal{N}_1(\mu, \sigma^2)$ distribution truncated to lie in the interval (a, b). Next, let $\mathcal{SN}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, $t_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ and $\mathcal{S}t_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ respectively denote q-dimensional skew-normal, t and skew-t distributions. Here $\boldsymbol{\mu} \in \mathbb{R}^q$ is a location vector; $\boldsymbol{\Sigma}$ is a $q \times q$ positive definite scale matrix; $\boldsymbol{\lambda} \in \mathbb{R}^q$ is a vector of skewness parameters; and $\nu\,(> 0)$ is degrees of freedom. To define these distributions, let $\phi_q(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the density of a $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution; $\Phi(\cdot)$ denote the distribution function of a univariate standard normal distribution; and $\tau(\cdot \mid \nu)$ denote the distribution function of a univariate t-distribution with ν degrees of freedom. Also, let $\boldsymbol{\Sigma}^{1/2}$ denote a symmetric square root of a symmetric, positive-definite matrix $\boldsymbol{\Sigma}$ so that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$; and let $\boldsymbol{\Sigma}^{-1/2}$ denote the inverse of $\boldsymbol{\Sigma}^{1/2}$.

We say $\mathbf{Y} \sim \mathcal{SN}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ if its density function is

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_q(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}'\mathbf{y}^*), \quad \mathbf{y} \in \mathbb{R}^q,$$

where $\mathbf{y}^* = \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$. Next, $\mathbf{Y} \sim t_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ if its density function is

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = (\nu\pi)^{-q/2}\,\frac{g((\nu+q)/2)}{g(\nu/2)}|\det(\boldsymbol{\Sigma})|^{-1/2}\,(1 + \mathbf{y}^{*\prime}\mathbf{y}^*/\nu)^{-(\nu+q)/2}, \quad \mathbf{y} \in \mathbb{R}^q.$$

Furthermore, $\mathbf{Y} \sim \mathcal{S}t_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ if its density function is

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) = 2\, f_t(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)\, \tau\left(\boldsymbol{\lambda}'\mathbf{y}^*\{(\nu+q)/(\nu + \mathbf{y}^{*\prime}\mathbf{y}^*)\}^{1/2} \mid \nu + q\right), \quad \mathbf{y} \in \mathbb{R}^q,$$

where $f_t(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ is the density of a $t_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ distribution. Stochastic representations of these distributions are given in Appendix A. See Azzalini and Capitanio (2003) for additional properties.

Consider a random vector $\mathbf{Y}$ following the mixed model (1) along with the assumptions that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}, \quad \mathbf{b} \sim \mathcal{S}t_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \nu_b), \quad \mathbf{e} \sim t_n(\mathbf{0}, \boldsymbol{\Sigma}, \nu_e), \tag{5}$$

where $\mathbf{b}$ and $\mathbf{e}$ are mutually independent. This $\mathbf{Y}$ serves as a building block in the development of the proposed skew-t mixed model. Reparameterize $(\boldsymbol{\Psi}, \boldsymbol{\lambda})$ as $(\boldsymbol{\Gamma}, \boldsymbol{\gamma})$, where

$$\boldsymbol{\delta} = \boldsymbol{\lambda}/(1 + \boldsymbol{\lambda}'\boldsymbol{\lambda})^{1/2}, \quad \boldsymbol{\gamma} = \boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}, \quad \boldsymbol{\Gamma} = \boldsymbol{\Psi}^{1/2}(\mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}')\boldsymbol{\Psi}^{1/2} = \boldsymbol{\Psi} - \boldsymbol{\gamma}\boldsymbol{\gamma}', \tag{6}$$

with $\mathbf{I}_q$ denoting a $q \times q$ identity matrix.

Proposition 1. *Consider* $\mathbf{Y}$ *as defined in* (5).

(a) A hierarchical representation for $\mathbf{Y}$ *is as follows:*

$$\mathbf{Y} \,|\, \mathbf{b}, W_e \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma}/W_e),\ \mathbf{b} \,|\, T, W_b \sim \mathcal{N}_q(\boldsymbol{\gamma}\, T, \boldsymbol{\Gamma}/W_b),$$
$$T \,|\, W_b \sim \mathcal{TN}(0, 1/W_b; (0,\infty)),\ W_b \sim \mathcal{G}(\nu_b/2, \nu_b/2),\ W_e \sim \mathcal{G}(\nu_e/2, \nu_e/2). \tag{7}$$

This representation may also be written by transforming (W_b, W_e) *to* $(U = W_b, V = W_b/W_e)$.

(b) The mean vector and variance matrix of $\mathbf{Y}$ *are as follows:*

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} + \sqrt{\frac{\nu_b}{\pi}}\,\frac{g((\nu_b-1)/2)}{g(\nu_b/2)}\,\mathbf{Z}\boldsymbol{\gamma}, \quad \nu_b, \nu_e > 1,$$
$$var[\mathbf{Y}] = \frac{\nu_b}{\nu_b - 2}\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}' + \frac{\nu_e}{\nu_e - 2}\boldsymbol{\Sigma} - \frac{\nu_b}{\pi}\left(\frac{g((\nu_b-1)/2)}{g(\nu_b/2)}\right)^2 \mathbf{Z}\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Z}', \quad \nu_b, \nu_e > 2. \tag{8}$$

Additional properties of $\mathbf{Y}$, including its marginal density, are summarized in Proposition 2 in Appendix A. This density, given by (A.7), is not available in a closed-form. It must be computed via a one-dimensional numerical integration.

2.2 The proposed mixed model

The proposed model for the data $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, m$, is model (1) but instead of the normality assumption (2), it assumes

$$\mathbf{b}_i \sim \text{ independent } \mathcal{S}t_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \nu_b), \quad \mathbf{e}_i \sim \text{ independent } t_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i, \nu_e), \tag{9}$$

and $\mathbf{b}_i$ and $\mathbf{e}_i$ are mutually independent. We call it a *general skew-t mixed model* (GSTMM). Clearly, it allows ν_b and ν_e to differ. The response vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ are independent copies of $\mathbf{Y}$ defined in the previous section. Therefore, we can get a hierarchical representation for $\mathbf{Y}_i$, and $E[\mathbf{Y}_i]$ and var$[\mathbf{Y}_i]$ simply by adding a subscript i to the random quantities in (7) and (8). Further, using (A.7) the observed data log-likelihood function is

$$\log L(\boldsymbol{\theta}) = \log\{f(\mathbf{y}_1, \ldots, \mathbf{y}_m|\boldsymbol{\theta})\} = \sum_{i=1}^{m} \log\{f(\mathbf{y}_i|\boldsymbol{\theta})\} = \sum_{i=1}^{m} \log\left(\int_0^\infty f(\mathbf{y}_i, v_i|\boldsymbol{\theta})\, dv_i\right), \tag{10}$$

with $f(\mathbf{y}_i, v|\boldsymbol{\theta})$ as in (A.6) and $\boldsymbol{\theta}$ as the vector of unknown model parameters.

The proposed GSTMM reduces to the usual NMM when $\boldsymbol{\lambda} = \mathbf{0}$ and $\nu_b, \nu_e \to \infty$. To compare it with STMM (Lachos et al., 2010; Ho and Lin, 2010) and its special case TMM (Pinheiro et al., 2001), let us first define the STMM. This model is also of the form (1) but it assumes

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{e}_i \end{bmatrix} \sim \text{ independent } \mathcal{S}t_{q+n_i} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_i \end{bmatrix}, \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{0} \end{bmatrix}, \nu \right). \tag{11}$$

It reduces to TMM when $\boldsymbol{\lambda} = 0$, meaning there is no skewness. Notice that STMM *jointly* models $(\mathbf{b}_i, \mathbf{e}_i)$ as a multivariate skew-t with ν degrees of freedom. Marginally, $\mathbf{b}_i \sim \mathcal{S}t_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \nu)$, $\mathbf{e}_i \sim t_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i, \nu)$ and $\text{cov}[\mathbf{b}_i, \mathbf{e}_i] = \mathbf{0}$; but $\mathbf{b}_i$ and $\mathbf{e}_i$ are not mutually independent unless $\nu \to \infty$. In contrast, $\mathbf{b}_i$ and $\mathbf{e}_i$ appear independently in GSTMM (see (9)), allowing them to have different degrees of freedom. To gain further insight into the difference between the two models, consider the hierarchical representation of STMM (Ho and Lin, 2010) similar to (7) for GSTMM:

$$\begin{aligned} &\mathbf{Y} \,|\, \mathbf{b}, W \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma}/W), \ \mathbf{b} \,|\, T, W \sim \mathcal{N}_q(\boldsymbol{\gamma}\, T, \boldsymbol{\Gamma}/W), \\ &T \,|\, W \sim \mathcal{TN}(0, 1/W; (0, \infty)), \ W \sim \mathcal{G}(\nu/2, \nu/2). \end{aligned} \tag{12}$$

A comparison of (7) and (12) shows that to model the t-distributions of $\mathbf{b}|T$ and $\mathbf{e}$ as scale mixtures of normals, STMM uses a common gamma variable W, whereas GSTMM uses two independent gamma variables, W_b and W_e. Clearly, STMM is a special case of GSTMM when W_b and W_e are *identical*, justifying calling the new model a "generalization." By identical we mean W_b and W_e are equal with probability one; merely having the same distribution (i.e., $\nu_b = \nu_e = \nu$) is not enough. Nevertheless, in this case the two models have the same $E[\mathbf{Y}]$ and $\text{var}[\mathbf{Y}]$, provided they exist.

In general, GSTMM and STMM are not expected to produce similar results when $\nu_b = \nu_e = \nu$. But they do tend to be similar when all ν_b, ν_e and ν are large, without necessarily being equal. In this case, the two models are similar to the skew-normal mixed model (Arellano-Valle et al., 2005) with skewness only in random effects. It may be noted that STMM is not nested within GSTMM because it is not obtained by imposing a constraint in the parameter space of GSTMM. As a result, a likelihood ratio test cannot be used to distinguish between them. One has to rely on a model selection criterion, e.g., the Akaike Information Criterion (AIC) for this purpose.

Besides NMM, other special cases of GSTMM can be obtained by making certain a priori assumptions about the parameters in (9). For example, assuming $\boldsymbol{\lambda} = \mathbf{0}$ gives $\mathbf{b}_i \sim t_q(\mathbf{0}, \boldsymbol{\Psi}, \nu_b)$,

whereas assuming $\nu_b \to \infty$ gives $\mathbf{b}_i \sim \mathcal{SN}_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$. These two assumptions together give $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi})$. Further, assuming $\nu_e \to \infty$ gives $\mathbf{e}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$.

2.3 An ECM algorithm for ML estimation

In principle, the log-likelihood function in (10) can be maximized directly to get the ML estimator $\hat{\boldsymbol{\theta}}$. But it is generally not practical due to the dimension of $\hat{\boldsymbol{\theta}}$. So we consider an EM-type algorithm — the ECM algorithm (Meng and Rubin, 1993) — wherein the M-step of EM is replaced by a sequence of computationally simpler constrained maximization (CM) steps. Each iteration of ECM increases the likelihood function and the algorithm typically converges to a local or global maxima.

To develop the ECM algorithm, we take $\mathbf{Y}_{\text{miss},i} = (\mathbf{b}_i, T_i, U_i, V_i)$ as the *missing data* and $(\mathbf{Y}_i, \mathbf{Y}_{\text{miss},i})$ as the *complete data* on the ith subject. The quantities $(\mathbf{b}, T, U, V)$ are from (7) with (W_b, W_e) transformed to $(U = W_b, V = W_b/W_e)$. It is convenient to partition $\boldsymbol{\theta}$ into three blocks: $(\boldsymbol{\beta}, \boldsymbol{\theta}_{\boldsymbol{\Sigma}})$, $(\boldsymbol{\gamma}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ and (ν_b, ν_e). Here $\boldsymbol{\theta}_{\boldsymbol{\Sigma}}$ is a vector of parameters, not depending on the subject i, that parameterizes $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m$. Similarly, $\boldsymbol{\theta}_{\boldsymbol{\Gamma}}$ is a vector of parameters that parameterizes $\boldsymbol{\Gamma}$. If $\boldsymbol{\Psi}$ (and hence $\boldsymbol{\Gamma}$) is unstructured, $\boldsymbol{\theta}_{\boldsymbol{\Gamma}}$ represents the distinct elements of $\boldsymbol{\Gamma}$.

Ignoring the terms that are free of $\boldsymbol{\theta}$, the expected complete-data log-likelihood in the rth ECM iteration, i.e., $E\left[\log\left\{f(\mathbf{Y}_1, \mathbf{Y}_{\text{miss},1}, \ldots, \mathbf{Y}_m, \mathbf{Y}_{\text{miss},m}|\boldsymbol{\theta})\right\} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_m, \boldsymbol{\theta}^{(r)}\right]$, can be written as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \sum_{i=1}^{m} \left\{ Q_{i1}\big(\boldsymbol{\beta}, \boldsymbol{\theta}_{\boldsymbol{\Sigma}}|\boldsymbol{\theta}^{(r)}\big) + Q_{i2}\big(\boldsymbol{\gamma}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}}|\boldsymbol{\theta}^{(r)}\big) + Q_{i3}\big(\nu_b, \nu_e|\boldsymbol{\theta}^{(r)}\big) \right\}, \tag{13}$$

where upon using E_r to denote the expectation over the conditional distribution of $\mathbf{Y}_{\text{miss},i}|\mathbf{Y}_i$ evaluated at $\boldsymbol{\theta}^{(r)}$, we have

$$\begin{aligned}
Q_{i1}(\boldsymbol{\beta}, \boldsymbol{\theta}_{\boldsymbol{\Sigma}}|\boldsymbol{\theta}^{(r)}) &= -(1/2)\log(\det \boldsymbol{\Sigma}_i) - (1/2)(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\, E_r[U_i/V_i] \\
&\quad + (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i\, E_r[\mathbf{b}_i U_i/V_i] - (1/2)\ \text{trace}\left(\mathbf{Z}_i'\,\boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i\, E_r[\mathbf{b}_i\mathbf{b}_i' U_i/V_i]\right), \\
Q_{i2}(\boldsymbol{\gamma}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}}|\boldsymbol{\theta}^{(r)}) &= -(1/2)\log(\det \boldsymbol{\Gamma}) - (1/2)\ \text{trace}\big(\boldsymbol{\Gamma}^{-1}\big\{E_r[\mathbf{b}_i\mathbf{b}_i' U_i] - E_r[\mathbf{b}_i T_i U_i]\boldsymbol{\gamma}' \\
&\quad - \boldsymbol{\gamma}(E_r[\mathbf{b}_i T_i U_i])' + E_r[T_i^2 U_i]\boldsymbol{\gamma}\boldsymbol{\gamma}'\big\}\big), \\
Q_{i3}(\nu_b, \nu_e|\boldsymbol{\theta}^{(r)}) &= (\nu_e/2)E_r[\log(U_i/V_i) - (U_i/V_i)] + (\nu_b/2)E_r[\log U_i - U_i] + (\nu_b/2)\log(\nu_b/2) \\
&\quad - \log(g(\nu_b/2)) + (\nu_e/2)\log(\nu_e/2) - \log(g(\nu_e/2)).
\end{aligned}$$

The E and CM steps in the rth iteration of the proposed ECM algorithm are as follows:

E-step: Compute the conditional expectations in (13) as described in Appendix B.

CM-step 1: Fix $\boldsymbol{\theta}_{\boldsymbol{\Sigma}} = \boldsymbol{\theta}_{\boldsymbol{\Sigma}}^{(r)}$ and update $\boldsymbol{\beta}$ by maximizing $\sum_{i=1}^{m} Q_{i1}\big(\boldsymbol{\beta}, \boldsymbol{\theta}_{\boldsymbol{\Sigma}}^{(r)} | \boldsymbol{\theta}^{(r)}\big)$ over $\boldsymbol{\beta}$, yielding

$$\boldsymbol{\beta}^{(r+1)} = \left(\sum_{i=1}^{m} \mathbf{X}_i' \big(\boldsymbol{\Sigma}_i^{(r)}\big)^{-1} \mathbf{X}_i E_r[U_i/V_i]\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i' \big(\boldsymbol{\Sigma}_i^{(r)}\big)^{-1} \big(\mathbf{Y}_i E_r[U_i/V_i] - \mathbf{Z}_i E_r[\mathbf{b}_i U_i/V_i]\big).$$

CM-step 2: Fix $\boldsymbol{\beta} = \boldsymbol{\beta}^{(r+1)}$ and update $\boldsymbol{\theta}_{\boldsymbol{\Sigma}}$ by numerically maximizing $\sum_{i=1}^{m} Q_{i1}\big(\boldsymbol{\beta}^{(r+1)}, \boldsymbol{\theta}_{\boldsymbol{\Sigma}} | \boldsymbol{\theta}^{(r)}\big)$ over $\boldsymbol{\theta}_{\boldsymbol{\Sigma}}$ to get $\boldsymbol{\theta}_{\boldsymbol{\Sigma}}^{(r+1)}$.

CM-step 3: Fix $\boldsymbol{\theta}_{\boldsymbol{\Gamma}} = \boldsymbol{\theta}_{\boldsymbol{\Gamma}}^{(r)}$ and update $\boldsymbol{\gamma}$ by maximizing $\sum_{i=1}^{m} Q_{i2}\big(\boldsymbol{\gamma}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}}^{(r)} | \boldsymbol{\theta}^{(r)}\big)$ over $\boldsymbol{\gamma}$, yielding

$$\boldsymbol{\gamma}^{(r+1)} = \sum_{i=1}^{m} E_r[\mathbf{b}_i T_i U_i] / \sum_{i=1}^{m} E_r[T_i^2 U_i].$$

CM-step 4: Fix $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(r+1)}$ and update $\boldsymbol{\theta}_{\boldsymbol{\Gamma}}$ by maximizing $\sum_{i=1}^{m} Q_{i2}\big(\boldsymbol{\gamma}^{(r+1)}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}} | \boldsymbol{\theta}^{(r)}\big)$ over $\boldsymbol{\theta}_{\boldsymbol{\Gamma}}$ to get $\boldsymbol{\theta}_{\boldsymbol{\Gamma}}^{(r+1)}$. If $\boldsymbol{\Gamma}$ is unstructured then $\boldsymbol{\theta}_{\boldsymbol{\Gamma}}^{(r+1)}$ consists of distinct elements of

$$\boldsymbol{\Gamma}^{(r+1)} = \frac{1}{m} \sum_{i=1}^{m} \Big(E_r[\mathbf{b}_i \mathbf{b}_i' U_i] - E_r[\mathbf{b}_i T_i U_i] \big(\boldsymbol{\gamma}^{(r+1)}\big)' - \boldsymbol{\gamma}^{(r+1)} \big(E_r[\mathbf{b}_i T_i U_i]\big)' + E_r[T_i^2 U_i] \boldsymbol{\gamma}^{(r+1)} \big(\boldsymbol{\gamma}^{(r+1)}\big)' \Big),$$

otherwise the maximization is done numerically.

CM-step 5: Update (ν_b, ν_e) by numerically maximizing $\sum_{i=1}^{m} Q_{i3}\big(\nu_b, \nu_e | \boldsymbol{\theta}^{(r)}\big)$ over (ν_b, ν_e) to get $\big(\nu_b^{(r+1)}, \nu_e^{(r+1)}\big)$.

The aforementioned expressions for $\boldsymbol{\beta}^{(r+1)}$, $\boldsymbol{\gamma}^{(r+1)}$ and $\boldsymbol{\Gamma}^{(r+1)}$ are verified in Appendix C. This ECM algorithm can be suitably modified to fit some special cases of GSTMM (see Appendix D). When a numerical maximization is needed, it is a good idea to transform the parameters to make the parameter space unconstrained (Pinheiro and Bates, 2000, ch. 2). Moreover, as is true for any EM-type algorithm, one needs to run the ECM algorithm with several starting points to have some assurance that the algorithm converges to a global maxima, $\hat{\boldsymbol{\theta}}$.

Next, let $\mathcal{I} = -\partial^2 \log L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^2 |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ denote the *observed information matrix* for $\boldsymbol{\theta}$. It is obtained by numerically differentiating $\log L(\boldsymbol{\theta})$ given by (10). From the large sample theory of ML estimators (Lehmann, 1998, ch. 7), $\hat{\boldsymbol{\theta}}$ is approximately normal with mean $\boldsymbol{\theta}$ and covariance matrix $\mathcal{I}^{-1}$ when the number of subjects m is large. This result can be used to compute standard errors for ML estimators, test hypotheses and construct confidence intervals.

Let the n_i-vector $\hat{\mathbf{Y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i$ denote the fitted response for the ith subject, $i = 1, \ldots, m$. Here $\hat{\mathbf{b}} = \widehat{E}[\mathbf{b}|\mathbf{Y}]$ estimates the best predictor of $\mathbf{b}$, namely $E[\mathbf{b}|\mathbf{Y}]$, which minimizes the mean squared prediction error in the class of all predictors of $\mathbf{b}$ based on $\mathbf{Y}$. It is computed by substituting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ in $E[\mathbf{b}|\mathbf{Y}]$ given by Proposition 5 in Appendix E. Next, let $\hat{\mathbf{e}}_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i$ denote the n_i-vector of residuals for the ith subject. These residuals can be used for model checking.

3 A Monte Carlo simulation study

In this section, we use simulation to compare ML estimators based on GSTMM and STMM. We do not include NMM and TMM in the comparison as they are special cases of the models being compared and have already been compared in Pinheiro et al. (2001) and Ho and Lin (2010). Pinheiro et al. conclude that the TMM-based estimators are more efficient than those based on NMM when outliers are present in the data, even in moderate amounts. Moreover, the gains in efficiency are bigger for variance-covariance parameters than fixed effects. Ho and Lin conclude that ML estimators based on TMM and STMM are more precise than those based on NMM.

We saw in Section 2 that $E[\mathbf{Y}]$, the mean response of a subject, under STMM or GSTMM does not equal the usual $\mathbf{X}\boldsymbol{\beta}$ unless the skewness parameter $\boldsymbol{\lambda} = \mathbf{0}$. For these models, the practitioner may be more interested in inference on $E[\mathbf{Y}]$ rather than $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ separately. Therefore, we compare estimators of $E[\mathbf{Y}]$ instead of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. We also compare estimators of the scale matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$. We take $m = 100$ subjects and focus on a simplified version of model (3) considered for the crab data, with no observer effect and only two repeated measurements. This model is

$$Y_{ijk} = \beta_j + b_{ij} + e_{ijk},\ i = 1, \ldots, 100,\ j = 1, 2,\ k = 1, 2. \tag{14}$$

It can be expressed in the familiar form (1) by taking $\mathbf{Y}_i = (Y_{i11}, Y_{i12}, Y_{i21}, Y_{i22})$, $\boldsymbol{\beta} = (\beta_1, \beta_2)$, $\mathbf{b}_i = (b_{i1}, b_{i2})$ and appropriately defining $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{e}_i$. In this case, $\boldsymbol{\Psi}$ has (ψ_1^2, ψ_2^2) as diagonal elements and ψ_{12} as the off-diagonal element. Moreover, $\boldsymbol{\Sigma} = \text{diag}\{\sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2\}$.

We consider two scenarios for simulating $\mathbf{b}_i$ and $\mathbf{e}_i$. In the first scenario, we follow Pinheiro et al. (2001) and simulate them from mixtures of normals that represent *contaminated* normal

distributions. Specifically,

$$\begin{aligned}
\mathbf{b}_i &\sim \text{ independent } (1-p_b)\,\mathcal{N}_2(\mathbf{0},\boldsymbol{\Psi}) + p_b c\,\mathcal{N}_2(\mathbf{0},\boldsymbol{\Psi}),\\
\mathbf{e}_i &\sim \text{ independent } (1-p_e)\,\mathcal{N}_4(\mathbf{0},\boldsymbol{\Sigma}) + p_e c\,\mathcal{N}_4(\mathbf{0},\boldsymbol{\Psi}),\ i=1,\ldots,100,
\end{aligned} \tag{15}$$

where p_b and p_e denote the respective expected proportions of **b**- and **e**-outliers in the data and c denotes the contamination factor. There are no **b**-outliers when $p_b = 0$ and no **e**-outliers when $p_e = 0$. We take $p_b, p_e = 0, 0.05, 0.10, 0.25$, and $c = 2, 4$, resulting in a total of 32 outlier patterns. Pinheiro et al. call the contamination pattern "close" when $c = 2$, and "distant" when $c = 4$. Thus, the simulation model in the first scenario is given by (14) and (15). The following values, motivated by the ML estimates from the crab data, are assigned to the parameters of this model:

$$(\beta_1,\beta_2) = (32,35),\ (\sigma_1^2,\sigma_2^2) = (1,1.25^2),\ (\psi_1^2,\psi_2^2,\psi_{12}) = (36,49,40.74),\ (\lambda_1,\lambda_2) = (0,0). \tag{16}$$

As in Pinheiro et al., the values in (16) are taken as the *target* values for the parameters. It follows that the target value for $E[\mathbf{Y}_1]$, the common mean response of the subjects, is $(32, 32, 35, 35)$.

We simulate 500 datasets from the foregoing simulation model for each combination of (p_b, p_e, c). Both GSTMM and STMM are fit to each dataset and the ML estimates of $E[\mathbf{Y}_1]$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are computed. The ECM algorithm of Section 2 is employed for fitting GSTMM and it is suitably modified for fitting STMM. The resulting estimates are used to estimate biases and mean squared errors of the estimators. The efficiencies of GSTMM-based estimators relative to STMM are computed by dividing the mean squared errors under STMM by those under GSTMM. The biases and relative efficiencies for $c = 4$ are presented in Tables 1 and 2, respectively. We do not present results for $c = 2$ as the conclusions remain qualitatively similar to the $c = 4$ case.

The results in Table 1 suggest that the estimators of $E[\mathbf{Y}_1]$ have comparable biases under both models. The biases in estimation of (σ_1^2, σ_2^2) can also be considered comparable unless $p_e = 0.25$ in which case STMM estimators have more bias than their GSTMM counterparts. The estimators of $\boldsymbol{\Psi}$ also generally have more bias under STMM than GSTMM, especially so when p_b is large.

We next consider the relative efficiencies presented in Table 2. For ease in interpretation of these results, we take the view that a gain or loss of up to 20% is too modest to be important for a practitioner, especially if the change is not in the same direction for all parameters. We see

that the two models are equally efficient when there are no outliers. When outliers are present, most efficiencies are one or more, a few are 0.9 and 0.8, but none is less than 0.8. This suggests that although GSTMM may lose some efficiency over STMM, the loss is never significant from a practical viewpoint. In contrast, GSTMM may offer substantial gains in efficiency over STMM, especially for $\mathbf{\Sigma}$ and $\mathbf{\Psi}$, depending upon the proportion of outliers. Specifically, the GSTMM's gain for $E[\mathbf{Y}_1]$ is notable when $p_e = 0.25$ and $p_b < p_e$, with maximum gain of 40%. The gain for $\mathbf{\Sigma}$ ranges between 40% to 500% when $p_e = 0, 0.10, 0.25$, with largest gains in case of $p_e = 0.25$, and efficiencies practically one in case of $p_e = 0.05$. The gain in efficiency for $\mathbf{\Psi}$ is quite substantial (between 400% to 1300%) when $p_b = 0.25$. On the whole, these findings suggest that there may not be much practical difference between the two models for estimation of the mean response, but GSTMM may be considerably more efficient than STMM for estimating $\mathbf{\Psi}$ and $\mathbf{\Sigma}$, especially when there is a sizable proportion of either **b**- or **e**-outliers.

On a reviewer's suggestion, we also compare the average estimated marginal densities of $\mathbf{b}_i$ under the two models. The predicted values of $\mathbf{b}_i$ are computed using Proposition 5 in Appendix E in case of GSTMM and using Ho and Lin (2010) in case of STMM. The two average densities are virtually indistinguishable and both capture the corresponding true density quite well.

We would like to note that in case of close contamination pattern ($c = 2$), average $\hat{\nu}_b$ equals 30, 27, 25 and 19 respectively when $p_b = 0, 0.05, 0.10$ and 0.25; whereas the average $\hat{\nu}_e$ equals 31, 25, 21 and 14 respectively when $p_e = 0, 0.05, 0.10$ and 0.25. In case of distant contamination pattern ($c = 4$), these averages are 30, 10, 4 and 2 for $\hat{\nu}_b$, and 34, 7, 4 and 2 for $\hat{\nu}_e$. Obviously, we expect $\hat{\nu}_b$ and $\hat{\nu}_e$ to get calibrated by the proportion and the size of outliers — these results give an idea of how this calibration works. Note, in particular, that the average fitted t-distribution in case of 25% distant outliers does not have a finite variance.

In the second scenario, we simulate $\mathbf{b}_i$ and $\mathbf{e}_i$ directly from t-distributions:

$$\mathbf{b}_i \sim \text{ independent } t_2(\mathbf{0}, \mathbf{\Psi}, \nu_b), \quad \mathbf{e}_i \sim \text{ independent } t_4(\mathbf{0}, \mathbf{\Sigma}, \nu_e), \ i = 1, \ldots, 100. \tag{17}$$

The focus is on 16 combinations obtained by taking $\nu_b, \nu_e = 4, 10, 30, \infty$, with values for other parameters given by (16). A t-distribution with small degrees of freedom has heavy tails and it converges to a normal distribution as its degrees of freedom approaches ∞. The simulation model

in the second scenario is given by (14) and (17). We proceed exactly as in the first scenario to estimate biases and efficiencies ML estimators based on GSTMM and STMM. These results are presented in Tables 3 and 4.

The results in Table 3 show that estimators of $E[\mathbf{Y}_1]$ have comparable biases under both models. However, the estimators of (σ_1^2, σ_2^2) and $\mathbf{\Psi}$ tend to have more bias under STMM than GSTMM unless both ν_b and ν_e are large. This finding is consistent with what we saw in the first scenario.

As for efficiency, we see that no entry in Table 4 is less than one, implying that the GSTMM-based estimators are at least as efficient as those based on STMM. However, like the first scenario, the gain in efficiency for GSTMM depends not only on the parameter being estimated but also on the extent of heavy tailedness. In particular, the relative efficiencies for $E[\mathbf{Y}_1]$ are practically one in all cases. The efficiencies for other parameters are also one when both ν_b and ν_e are large. On the other hand, in case of $\nu_e = 4$, the efficiency for $\mathbf{\Sigma}$ range between 4.8 to 8.5, with values increasing as ν_b increases. Similarly, when $\nu_e = 4$, the efficiency for $\mathbf{\Psi}$ range between 4.3 to 10.2, with values increasing as ν_e increases. Overall, these results show that there is no practical difference in the two models for estimation of $E[\mathbf{Y}_1]$. However, if either ν_b or ν_e is small, GSTMM may be substantially more efficient than STMM for estimation of $\mathbf{\Psi}$ and $\mathbf{\Sigma}$. The gain in efficiency is largest when one degree of freedom parameter is small and the other is large.

Taken together, the findings for scenarios 1 and 2 can be summarized as follows: There is not much practical difference in the two models when both random effects and error distributions are either normal or nearly so or when the estimand is mean response. However, when at least one of the distributions has heavy tails, possibly due to a sizable proportion of either **b**- or **e**- outliers, the GSTMM estimators of scale matrices $\mathbf{\Psi}$ and $\mathbf{\Sigma}$ may be substantially more efficient and less biased than the STMM estimators. These conclusions remain unchanged when the simulation study is repeated with non-zero skewness in random effects, three repeated measurements per subject and other combinations for location and scale parameters (results not shown).

4 Analysis of crab claws data

We now return to the crab data introduced in Section 1 where we found that normality was reasonable for random effects but a heavy tailed distribution was needed for errors. So we model these data using a special case of GSTMM that assumes $(\boldsymbol{\lambda}, \nu_b) = (\mathbf{0}, \infty)$, i.e., $\mathbf{b}_i \sim$ independent $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Psi})$ and $\mathbf{e}_i \sim$ independent $t_{18}(\mathbf{0}, \boldsymbol{\Sigma}, \nu_e)$. Here $\boldsymbol{\Psi}$ is unstructured with (ψ_1^2, ψ_2^2) as diagonal elements and ψ_{12} as the off-diagonal element, and $\boldsymbol{\Sigma}$ is as defined in (4). This model has a total of 16 parameters. Table 5 summarizes the ML estimates and their standard errors when this model is fit as described in Appendix D. The numerical derivatives and quadratures needed for this computation are obtained using `numDeriv` package (Gilbert, 2011) and `statmod` (Smyth et al., 2011) packages in `R`. For the fitted model, $\log L(\hat{\boldsymbol{\theta}}) = 284.95$ and AIC $= -537.90$.

The assumed GSTMM is preferred by AIC over four other competing models — full GSTMM (without any constraint on $\boldsymbol{\lambda}$ and ν_b; 19 parameters, AIC $= -532.18$), STMM (18 parameters, AIC $= -524.48$), TMM (16 parameters, AIC $= -528.39$), NMM (15 parameters, AIC $= -429.62$). The fitting of the various models takes the following amounts of time (in seconds) on a Linux machine with a 2.8 GHz processor and 1.5 GB memory: 2 sec (NMM), 39 sec (TMM), 49 sec (STMM), 116 sec (assumed GSTMM) and 196 sec (full GSTMM).

Figure 3 presents a QQ plot of standardized residuals $(\hat{e}_{ijkl}/\hat{\sigma}_{jl})$, computed as described in Section 2. The plot is generated using the `car` package (Fox and Weisberg, 2011) in `R` by taking t-distribution with degrees of freedom $\hat{\nu}_e = \exp(1.28) = 3.6$ as the reference distribution. The plot corroborates that, except for the four outlying points, the t-distribution assumption for the errors is quite reasonable. The adequacy of the assumed model is confirmed by graphical checks recommended by Pinheiro and Bates (2000). To assess the impact of the four outliers on the fitted model, we refit the model twice — first by removing them from the data and next by replacing them with their presumably correct values. We see that $\hat{\nu}_e$ increases to about 8 in both cases, but there is little to no change in estimates of other parameters and their standard errors. This demonstrates that the estimates of the fixed effects and variance-covariance parameters in a GSTMM are robust to outliers but, nor surprisingly, the estimate of degrees of freedom gets calibrated by them.

The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ have the same interpretation under the assumed GSTMM and NMM.

Therefore, we can compare their estimates and standard errors reported in Table 5. In case of $\boldsymbol{\beta}$, the estimates are roughly the same but their standard errors are smaller under GSTMM than NMM. In case of $(\log(\psi_1^2), \log(\psi_2^2))$, both estimates and standard errors are similar. In case of $z(\rho)$, the Fisher's z-transformation of the random effects correlation ρ, both estimate and standard error are smaller under NMM than GSTMM. Nevertheless, they lead to practically the same confidence interval on the original scale of ρ. These results show that the assumed GSTMM not only fits better than NMM, but it also produces estimates of $\boldsymbol{\beta}$ that are more precise than NMM. In case of $\boldsymbol{\Sigma}$, we see that the NMM-based log-scale estimates are a bit larger and have slightly smaller standard errors compared to GSTMM. However, these estimates are not strictly comparable as $\boldsymbol{\Sigma}$ does not have the same interpretation under the two models.

Our next task is to use the fitted GSTMM to separately examine for each observer whether the means and error variances of the two calipers are the same. To this end, we first test the null hypothesis $\beta_{1l} = \beta_{2l}$, $l = 1, 2, 3$, against its complement. The p-value for the asymptotic likelihood ratio test of this hypothesis is less than 0.001. Next, we test the null hypothesis $\sigma_{1l}^2 = \sigma_{2l}^2$, $l = 1, 2, 3$, against its complement. The p-value for this test is also less than 0.001. Thus, there is at least one observer for whom the means as well as variances of the two methods differ significantly. Nevertheless, the estimates in Table 5 show that the differences, especially in the means, are minor relative to the scale of measurement, and are not practically significant.

5 Discussion

The GSTMM proposed in this article offers more flexibility in modeling of data with outliers, heavy tailedness and skewness than the STMM. Although the GSTMM does not gain substantial efficiency over the STMM for estimating mean response, its efficiency gain for estimating scale matrices of random effects and error distributions can be quite considerable. Accurate estimation of these matrices is important as they determine standard errors of the estimated fixed effects and mean response besides often being of independent interest. These advantages of GSTMM come at the cost of additional complexity in fitting of the model. For either model, however, the

computations need to be programmed using a software package such as R.

We saw both in simulation study and analysis of crab data that the estimates of the degrees of freedom parameters in the GSTMM are sensitive to outliers. If a large number of extreme outliers are present then the fitted t-distribution may not have finite mean or variance, which may pose a problem in applications where inference on mean or variance of response is desired. In such a case, alternative robust mixed modeling approaches that either constrain the influence of outlying observations on the fitted model (Richardson, 1997) or produce estimators with high breakdown points (Heritier et al., 2009) may be more attractive than the GSTMM.

We fit the GSTMM using an ECM algorithm wherein even the degrees of parameters are updated in a CM step. However, the literature on fitting t-distributions generally suggests replacing the CM step for updating the degrees of freedom by a step that maximizes the actual observed data likelihood over this parameter while keeping all other parameters at their current values. This results in the so-called ECME algorithm (Liu and Rubin, 1994), which tends to converge faster than its ECM counterpart. Although ECME can be used in place of ECM for GSTMM as well, the computation of the likelihood (10) needed for ECME is quite time consuming, potentially outweighing its benefit. This is indeed the case in crab data where ECME takes a much longer time to converge than ECM.

We invert the numerically computed observed information matrix to estimate the covariance matrix of the ML estimators. Alternatives to this approach include the method of Louis (1982) and others reviewed in McLachlan and Krishnan (2007, section 4.7). Further work is needed to develop and compare these methods for covariance matrix estimation.

Acknowledgments

The authors gratefully acknowledge the assistance of Golo Maurer, Rebecca Boulton and Leanne Reaney of the last author's laboratory in collection of the crab claws data used in this article. They also thank two reviewers and the Associate Editor for providing comments that led to substantial improvements in this article.

Appendix A Properties of Y (Section 2.1)

Consider $\mathbf{b}$ and $\mathbf{e}$ as defined in (5), and W_b and W_e as defined in (7). Let $\mathbf{b}^* \sim \mathcal{SN}_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$, $\mathbf{G}_1^* \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$, $\mathbf{G}_2^* \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}_q)$ and $T^* \sim \mathcal{TN}(0, 1; (0, \infty))$. Assume that $\mathbf{G}_1^*$, $\mathbf{G}_2^*$, T^*, W_b and W_e are mutually independent. Then, with $\boldsymbol{\delta}$ given by (6), the following stochastic representations hold for $\mathbf{e}$, $\mathbf{b}^*$ and $\mathbf{b}$ (Ho and Lin, 2010):

$$\mathbf{e} \stackrel{d}{=} \boldsymbol{\Sigma}^{1/2}\mathbf{G}_1^*/W_e^{1/2}, \quad \mathbf{b}^* \stackrel{d}{=} \boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}T^* + \boldsymbol{\Psi}^{1/2}(\mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}')^{1/2}\mathbf{G}_2^*, \quad \mathbf{b} \stackrel{d}{=} \mathbf{b}^*/W_b^{1/2}. \tag{A.1}$$

Proof of Proposition 1. The result in (a) follows from (A.1) upon using (6). The result in (b) follows from (a) upon applying the law of iterated expectations and using well-known results about moments involving normal and gamma variates. □

Next, we present two results from literature that will help us derive the marginal density of $\mathbf{Y}$ in Proposition 2. Define for $v > 0$,

$$\begin{aligned} &\boldsymbol{\Pi}_v = (\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}' + v\boldsymbol{\Sigma}), \quad \boldsymbol{\lambda}_v = \frac{\boldsymbol{\Pi}_v^{-1/2}\mathbf{Z}\boldsymbol{\Psi}^{1/2}\boldsymbol{\lambda}}{\left(1 + \boldsymbol{\lambda}'\boldsymbol{\Psi}^{-1/2}(\boldsymbol{\Psi}^{-1} + \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}/v)^{-1}\boldsymbol{\Psi}^{-1/2}\boldsymbol{\lambda}\right)^{1/2}}, \\ &\xi_v = \boldsymbol{\lambda}_v'\boldsymbol{\Pi}_v^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad \eta_v = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Pi}_v^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \nu_b + \nu_e/v, \end{aligned} \tag{A.2}$$

and let $h_v(\cdot)$ denote the density a $\mathcal{G}(n^*/2, \eta_v/2)$ distribution, where

$$n^* = n + \nu_b + \nu_e. \tag{A.3}$$

Lemma 1 (Arellano-Valle et al., 2005)**.** *Suppose* $\mathbf{Y}|\mathbf{b} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma})$ *and* $\mathbf{b} \sim \mathcal{SN}_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$. *Then marginally,* $\mathbf{Y} \sim \mathcal{SN}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Pi}_{v=1}, \boldsymbol{\lambda}_{v=1})$, *with* $\boldsymbol{\Pi}_v$ *and* $\boldsymbol{\lambda}_v$ *as in* (A.2).

Lemma 2 (Azzalini and Capitanio, 2003)**.** *Suppose* $W \sim \mathcal{G}(\alpha, \beta)$. *Then for any* $c \in \mathbb{R}$ *we have,* $E\big(\Phi(cW^{1/2})\big) = \tau\big(c\,(\alpha/\beta)^{1/2} \mid 2\alpha\big)$.

Proposition 2. *Consider* $\mathbf{Y}$ *as defined in* (5) *along with the quantities given by* (A.2) *and* (A.3).

(a) A hierarchical representation for $\mathbf{Y}$ *is as follows:*

$$\mathbf{Y} \,|\, U, V \sim \mathcal{SN}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Pi}_V/U, \boldsymbol{\lambda}_V), \; U \sim \mathcal{G}(\nu_b/2, \nu_b/2), \; U/V \sim \mathcal{G}(\nu_e/2, \nu_e/2). \tag{A.4}$$

(b) For $\mathbf{y} \in \mathbb{R}^n$, $u, v > 0$, the joint density of $(\mathbf{Y}, U, V)$ is

$$f(\mathbf{y}, u, v|\boldsymbol{\theta}) = 2\pi^{-n/2}\,\nu_b^{\nu_b/2}\nu_e^{\nu_e/2}\frac{g(n^*/2)}{g(\nu_b/2)g(\nu_e/2)}\frac{(det\ \boldsymbol{\Pi}_v)^{-1/2}}{v^{1+\nu_e/2}\,\eta_v^{n^*/2}}\,\Phi\big(\xi_v u^{1/2}\big)\,h_v(u). \tag{A.5}$$

(c) For $\mathbf{y} \in \mathbb{R}^n$, $v > 0$, the joint density of $(\mathbf{Y}, V)$ is

$$f(\mathbf{y}, v|\boldsymbol{\theta}) = 2\pi^{-n/2}\,\nu_b^{\nu_b/2}\nu_e^{\nu_e/2}\frac{g(n^*/2)}{g(\nu_b/2)g(\nu_e/2)}\frac{(det\ \boldsymbol{\Pi}_v)^{-1/2}}{v^{1+\nu_e/2}\,\eta_v^{n^*/2}}\,\tau\big(\xi_v(n^*/\eta_v)^{1/2} \mid n^*\big). \tag{A.6}$$

(d) For $\mathbf{y} \in \mathbb{R}^n$, the marginal density of $\mathbf{Y}$ can be computed as

$$f(\mathbf{y}|\boldsymbol{\theta}) = \int_0^\infty f(\mathbf{y}, v|\boldsymbol{\theta})\,dv. \tag{A.7}$$

Proof. For (a), we use (A.1) to represent $\mathbf{Y}$ as

$$\mathbf{Y}|\mathbf{b}, W_e \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma}/W_e), \quad \mathbf{b}|W_b \sim \mathcal{SN}_q(\mathbf{0}, \boldsymbol{\Psi}/W_b, \boldsymbol{\lambda}),$$

where $W_b \sim \mathcal{G}(\nu_b/2, \nu_b/2)$ and $W_e \sim \mathcal{G}(\nu_e/2, \nu_e/2)$. The result now holds from Lemma 1 upon transforming (W_b, W_e) to $(U = W_b, V = W_b/W_e)$. For (b), we use (a) to write the joint density of $(\mathbf{Y}, U, V)$ and simplify. For (c), we integrate out u from $f(\mathbf{y}, u, v|\boldsymbol{\theta})$ given in (b) using Lemma 2. For (d), we integrate out v from $f(\mathbf{y}, v|\boldsymbol{\theta})$ given in (c). □

Appendix B Expectations for E-step (Section 2.3)

In this section, we omit the subject index i as subscript for random variates to simplify the notation. Our strategy is to first obtain the joint distribution of the missing data $(\mathbf{b}, T, U, V)$ conditional on the observed $\mathbf{Y}$ and then use it to get the desired expectations. Define for $v > 0$,

$$\boldsymbol{\Omega}_v = \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}' + v\boldsymbol{\Sigma}, \quad \boldsymbol{\Lambda}_v = (\boldsymbol{\Gamma}^{-1} + \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}/v)^{-1}, \quad \zeta_v^2 = (1 + \boldsymbol{\gamma}'\mathbf{Z}'\boldsymbol{\Omega}_v^{-1}\mathbf{Z}\boldsymbol{\gamma})^{-1}. \tag{A.8}$$

Using well-known results for patterned matrices (Seber and Lee, 2003, page 467), one can see that $\boldsymbol{\Omega}_v$ and $\boldsymbol{\Lambda}_v$ are related as $\boldsymbol{\Lambda}_v = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{Z}'\boldsymbol{\Omega}_v^{-1}\mathbf{Z}\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}_v\mathbf{Z}'\boldsymbol{\Sigma}^{-1}/v = \boldsymbol{\Gamma}\mathbf{Z}'\boldsymbol{\Omega}_v^{-1}$. Moreover, ξ_v defined in (A.2) can also be written as $\xi_v = \zeta_v\boldsymbol{\gamma}'\mathbf{Z}'\boldsymbol{\Omega}_v^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. We also need the following result that gives first two moments of a normal random variable truncated to be positive.

Lemma 3 (Johnson et al, 1994, pages 156-163)**.** *Let* $T \sim \mathcal{TN}(\eta, \omega^2; (0, \infty))$. *Then*

$$E[T] = \eta + \frac{\phi(\eta/\omega)}{\Phi(\eta/\omega)}\,\omega, \quad E[T^2] = \eta^2 + \omega^2 + \frac{\phi(\eta/\omega)}{\Phi(\eta/\omega)}\,\eta\omega.$$

Proposition 3. *Consider* $\mathbf{Y}$ *and other quantities as in Proposition 2 together with* (A.8).

(a) The conditional density of $V|\mathbf{Y}$ *is* $f(v|\mathbf{y}, \boldsymbol{\theta}) = f(\mathbf{y}, v|\boldsymbol{\theta})/f(\mathbf{y}|\boldsymbol{\theta})$, $\mathbf{y} \in \mathbb{R}^n, v > 0$.

(b) The conditional density of $U|\mathbf{Y}, V$ *is* $f(u|\mathbf{y}, v, \boldsymbol{\theta}) = f(\mathbf{y}, u, v|\boldsymbol{\theta})/f(\mathbf{y}, v|\boldsymbol{\theta})$, $\mathbf{y} \in \mathbb{R}^n, u, v > 0$.

(c) $T|\mathbf{Y}, U, V \sim \mathcal{TN}\big(\xi_V\zeta_V, \zeta_V^2/U; (0, \infty)\big)$.

(d) $\mathbf{b}|\mathbf{Y}, T, U, V \sim \mathcal{N}_q\big(\boldsymbol{\Gamma}\mathbf{Z}'\boldsymbol{\Omega}_V^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\Lambda}_V\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}\,T, \boldsymbol{\Lambda}_V/U\big)$.

Proof. The results in (a) and (b) hold from the definition of a conditional density. For (c), we can see from (7) that $\mathbf{Y}|T, U, V \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}\,T, \boldsymbol{\Omega}_V/U)$, $T|U, V \sim \mathcal{TN}(0, 1/U; (0, \infty))$. Moreover as a function of t, $f(t|\mathbf{y}, u, v, \boldsymbol{\theta}) \propto f(\mathbf{y}|t, u, v, \boldsymbol{\theta})f(t|u, v, \boldsymbol{\theta})$. Now an application of Arellano-Valle et al. (2005, Lemma 3) shows that the product on the right is further proportional to the density of a $\mathcal{TN}(\xi_v\zeta_v, \zeta_v^2/u; (0, \infty))$ distribution evaluated at t. This establishes the result.

For (d), we proceed in a similar manner to write $f(\mathbf{b}|\mathbf{y}, t, u, v, \boldsymbol{\theta}) \propto f(\mathbf{y}|\mathbf{b}, u, v, \boldsymbol{\theta})f(\mathbf{b}|t, u, \boldsymbol{\theta})$ as a function of $\mathbf{b}$. Next an application of Arellano-Valle et al. (2005, Lemma 2) shows that the product on the right is proportional to the density of a $\mathcal{N}_q(\boldsymbol{\gamma}t + \boldsymbol{\Lambda}_v\mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}t)/v, \boldsymbol{\Lambda}_v/u)$ distribution evaluated at $\mathbf{b}$. The result follows from further simplifying the mean using (A.8). □

Proposition 4. *Consider* $\mathbf{Y}$ *and other quantities as in Proposition 3. Then we have the following expectations conditional on* $\mathbf{Y}$ *and* V.

(a) For an integer r *such that* $n^* + 2r > 0$,

$$E[U^r|\mathbf{Y}, V] = 2^r\,\frac{g((n^* + 2r)/2)}{g(n^*/2)}\,\frac{\tau\big(\xi_V\{(n^* + 2r)/\eta_V\}^{1/2} \mid n^* + 2r\big)}{\eta_V^r\,\tau\big(\xi_V\{n^*/\eta_V\}^{1/2} \mid n^*\big)}.$$

(b) For an integer r *such that* $n^* + r > 0$,

$$E\left[U^{r/2}\,\frac{\phi(\xi_V U^{1/2})}{\Phi(\xi_V U^{1/2})}\,|\mathbf{Y}, V\right] = \frac{2^{(r-1)/2}}{\pi^{1/2}}\,\frac{g((n^* + r)/2)}{g(n^*/2)}\,\frac{\eta_V^{n^*/2}}{(\eta_V + \xi_V^2)^{(n^*+r)/2}\,\tau\big(\xi_V\{n^*/\eta_V\}^{1/2} \mid n^*\big)}.$$

(c) For an integer r such that $n^* + 2r > 1$,

$$E[TU^r \,|\mathbf{Y}, V] = \zeta_V \left(\xi_V E[U^r|\mathbf{Y}, V] + E\left[U^{(2r-1)/2} \frac{\phi(\xi_V U^{1/2})}{\Phi(\xi_V U^{1/2})} \,|\mathbf{Y}, V \right] \right).$$

(d) For an integer r such that $n^* + 2r > 2$,

$$E[T^2U^r \,|\mathbf{Y}, V] = \zeta_V^2 E[U^{r-1}|\mathbf{Y}, V] + \xi_V \zeta_V E[TU^r \,|\mathbf{Y}, V].$$

(e) $E[\mathbf{b}TU \,|\mathbf{Y}, V] = \mathbf{\Gamma Z}'\mathbf{\Omega}_V^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})E[TU \,|\mathbf{Y}, V] + \mathbf{\Lambda}_V \mathbf{\Gamma}^{-1}\boldsymbol{\gamma} E[T^2U \,|\mathbf{Y}, V]$.

(f) $E[\mathbf{b}U \,|\mathbf{Y}, V] = \mathbf{\Gamma Z}'\mathbf{\Omega}_V^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})E[U \,|\mathbf{Y}, V] + \mathbf{\Lambda}_V \mathbf{\Gamma}^{-1}\boldsymbol{\gamma} E[TU \,|\mathbf{Y}, V]$.

(g) $E[\mathbf{b}\mathbf{b}'U \,|\mathbf{Y}, V] = \left(\mathbf{I}_q + E[\mathbf{b}U \,|\mathbf{Y}, V](\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{\Sigma}^{-1}\mathbf{Z}/V + E[\mathbf{b}TU \,|\mathbf{Y}, V]\boldsymbol{\gamma}'\mathbf{\Gamma}^{-1} \right) \mathbf{\Lambda}_V$.

Proof. For (a), we write $E[U^r|V, \mathbf{Y}] = \int_0^\infty u^r f(u|\mathbf{y}, v, \boldsymbol{\theta})\, du$, and simplify using Proposition 2 and Lemma 2. A similar argument is used for (b). For the rest, we use Proposition 3, Lemma 3 and the law of iterated expectations. □

The aforementioned results suggest that the expectations involved in the E-step can be computed in the following manner. The two expectations, $E[\log(V)|\mathbf{Y}] = \int_0^\infty \log(v) f(v|\mathbf{y}, \boldsymbol{\theta})\, dv$ and $E[\log(U) \,|\mathbf{Y}] = \int_0^\infty \left(\int_0^\infty \log(u) f(u|\mathbf{y}, v, \boldsymbol{\theta})\, du \right) f(v|\mathbf{y}, \boldsymbol{\theta})\, dv$, need to be computed numerically with the densities given by Proposition 3. To compute expectation of any other random quantity, first compute its expectation conditional on $(\mathbf{Y}, V)$ using Proposition 4 and then average it over the conditional distribution of $V|\mathbf{Y}$. The one-dimensional integral involved in this averaging must be computed numerically. All the expectations are evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(r)}$.

Appendix C Maximizers in CM-steps (Section 2.3)

The expressions for $\boldsymbol{\beta}^{(r+1)}$ in CM-step 1 and $\boldsymbol{\gamma}^{(r+1)}$ in CM-step 3 are obtained by respectively solving $(\partial/\partial\boldsymbol{\beta}) \sum_{i=1}^m Q_{i1}\left(\boldsymbol{\beta}, \boldsymbol{\theta}_{\mathbf{\Sigma}}^{(r)} | \boldsymbol{\theta}^{(r)}\right) = 0$ for $\boldsymbol{\beta}$ and $(\partial/\partial\boldsymbol{\gamma}) \sum_{i=1}^m Q_{i2}\left(\boldsymbol{\gamma}, \boldsymbol{\theta}_{\mathbf{\Gamma}}^{(r)} | \boldsymbol{\theta}^{(r)}\right) = 0$ for $\boldsymbol{\gamma}$. The expression for $\mathbf{\Gamma}^{(r+1)}$ in CM-step 4 follows from a standard result that the quantity $-(m/2)\log(\det \mathbf{\Gamma}) - (1/2)\,\mathrm{trace}(\mathbf{\Gamma}^{-1} \sum_{i=1}^m \mathbf{A}_i)$ is maximized with respect to $\mathbf{\Gamma}$ when $\mathbf{\Gamma} = \sum_{i=1}^m \mathbf{A}_i/m$ (Johnson and Wichern, 2002, Result 4.10).

Appendix D Fitting some special cases of GSTMM

We now briefly explain how the ECM algorithm of Section 2.3 can be modified to fit some special cases of GSTMM. In case 1, we take $\boldsymbol{\lambda} = \mathbf{0}$, i.e., $\mathbf{b}_i \sim t_q(\mathbf{0}, \boldsymbol{\Psi}, \nu_b)$, implying $(\boldsymbol{\gamma}, \boldsymbol{\Gamma}) = (\mathbf{0}, \boldsymbol{\Psi})$. The algorithm works by setting $\boldsymbol{\gamma} = \mathbf{0}$ in the expected log-likelihood (13) and omitting CM-step 3.

In case 2, we take $\nu_b \to \infty$, i.e., $\mathbf{b}_i \sim \mathcal{SN}_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$. In this case, we do not need the gamma distribution for W_b as $U = W_b \equiv 1$. The ECM algorithm can be applied after suitably modifying the likelihood (13). Essentially this means setting $U \equiv 1$ and removing the terms involving ν_b.

In case 3, we take $\boldsymbol{\lambda} = \mathbf{0}$ and $\nu_b \to \infty$, i.e., $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi})$. The ECM algorithm can be applied after modifying the likelihood (13) as suggested for cases 1 and 2.

In case 4, we take $\nu_e \to \infty$, i.e., $\mathbf{e}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. As in case 2, the gamma distribution for W_e is not needed as $W_e = U/V \equiv 1$. We can apply the ECM algorithm essentially after setting $U/V \equiv 1$ and dropping the terms involving ν_e.

In case 5, $\boldsymbol{\lambda} = \mathbf{0}$ and $\nu_b, \nu_e \to \infty$, i.e., $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi})$ and $\mathbf{e}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. This is the usual NMM. Although the ECM algorithm can be applied after modifying (13) as suggested for cases 3 and 4, an alternative approach is generally used in practice (Pinheiro and Bates, 2000, ch. 2).

Appendix E Computing $E[\mathbf{b}|\mathbf{Y}]$ (Section 2.3)

Proposition 5. *Consider* $\mathbf{Y}$ *as defined in* (5). *Then* $E[\mathbf{b}|\mathbf{Y}] = \int_0^\infty E[\mathbf{b}|\mathbf{y}, v]\, f(v|\mathbf{y}, \boldsymbol{\theta})\, dv$, *where* $f(v|\mathbf{y}, \boldsymbol{\theta})$ *is given by Proposition 3 and*

$$E[\mathbf{b}|\mathbf{Y}, V] = \boldsymbol{\Gamma}\mathbf{Z}'\boldsymbol{\Omega}_V^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\Lambda}_V\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}\,\zeta_V\left(\xi_V + E\left[U^{-1/2}\,\frac{\phi(\xi_V U^{1/2})}{\Phi(\xi_V U^{1/2})}\,|\mathbf{Y}, V\right]\right),$$

with the last expectation given by part (b) of Proposition 4.

Proof. It suffices to verify the expression for $E[\mathbf{b}|\mathbf{Y}, V]$ to establish the result. For this, we write $E[\mathbf{b}|\mathbf{Y}, V] = E[E[\mathbf{b}|\mathbf{Y}, T, U, V]]$, where the inner expectation is $\boldsymbol{\Gamma}\mathbf{Z}'\boldsymbol{\Omega}_V^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\Lambda}_V\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}\,T$ from part (d) of Proposition 3 and the outer expectation is with respect to the distribution of $(T, U)|\mathbf{Y}, V]$. The result follows from computing this expectation by again applying the law of iterated expectations — by first computing it with respect to $T|\mathbf{Y}, U, V$ using part (c) of Propo-

sition 3 and Lemma 3, and then averaging the result over the conditional distribution of $U|\mathbf{Y}, V$ using part (b) of Proposition 4. □

References

Arellano-Valle, R. B., Bolfarine, H. and Lachos, V. H. (2005). Skew-normal linear mixed models. *Journal of Data Science* **3**, 415–438.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171–178.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B* **65**, 367–389.

Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8**, 135–160.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A* **143**, 383–430.

Branco, M. D. and Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis* **79**, 99–113.

Butler, S. M. and Louis, T. A. (1992). Random effects models with nonparametric priors. *Statistics in Medicine* **11**, 1981–2000.

Choudhary, P. K. and Yin, K. (2010). Bayesian and frequentist methodologies for analyzing method comparison studies with multiple methods. *Statistics in Biopharmaceutical Research* **2**, 122–132.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*, 2nd edn. Sage, Thousand Oaks, CA.

Gilbert, P. (2011). *numDeriv: Accurate numerical derivatives*. `http://CRAN.R-project.org/package=numDeriv`.

Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. John Wiley, New York.

Ho, H. J. and Lin, T. I. (2010). Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Biometrical Journal* **52**, 449–469.

Jiang, J. (1999). Conditional inference about generalized linear mixed models. *Annals of Statistics* **27**, 1974–2007.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York.

Jiang, J. (2013). The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond. *Annals of Statistics* **41**, 177–195.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd edn. John Wiley, New York.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*, 5th edn. Prentice Hall, Upper Saddle River, NJ.

Lachos, V. H., Ghosh, P. and Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica* **20**, 303–322.

Lailvaux, S. P., Reaney, L. T. and Backwell, P. R. Y. (2009). Dishonest signalling of fighting ability and multiple performance traits in the fiddler crab Uca mjoebergi. *Functional Ecology* **23**, 359–366.

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

Lehmann, E. L. (1998). *Elements of Large-Sample Theory*. Springer, New York.

Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**(4), 633–648.

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.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and Extensions*, second edn. Wiley, New York.

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

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Pinheiro, J. C., Bates, D., DebRoy, S., Sarkar, D. and R Development Core Team (2012). *nlme: Linear and nonlinear mixed effects models*. `http://CRAN.R-project.org/package=nlme`.

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.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. `http://www.R-project.org`.

Richardson, A. M. (1997). Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association* **92**(437), 154–161.

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.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, 2nd edn. John Wiley, New York.

Sengupta, D. (2012). *A Robust Linear Mixed Effects Model With Application to Method Comparison Studies*. Ph.D. Dissertation, The University of Texas at Dallas.

Smyth, G., Hu, Y., Dunn, P. and Phipson, B. (2011). *statmod: Statistical Modeling*. `http://CRAN.R-project.org/package=statmod`.

Song, P. X.-K., Zhang, P. and Qu, A. (2007). Maximum likelihood inference in robust linear mixed-effects models using multivariate t distribution. *Statistica Sinica* **17**, 929–943.

Stahel, W. A. and Welsh, A. H. (1997). Approaches to robust estimation in the simplest variance components model. *Journal of Statistical Planning and Inference* **57**, 295–319.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.

Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing* **1**, 105–117.

Wang, P., Tsai, G.-F. and Qu, A. (2012). Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association* **107**, 725–736.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57**(3), 795–802.

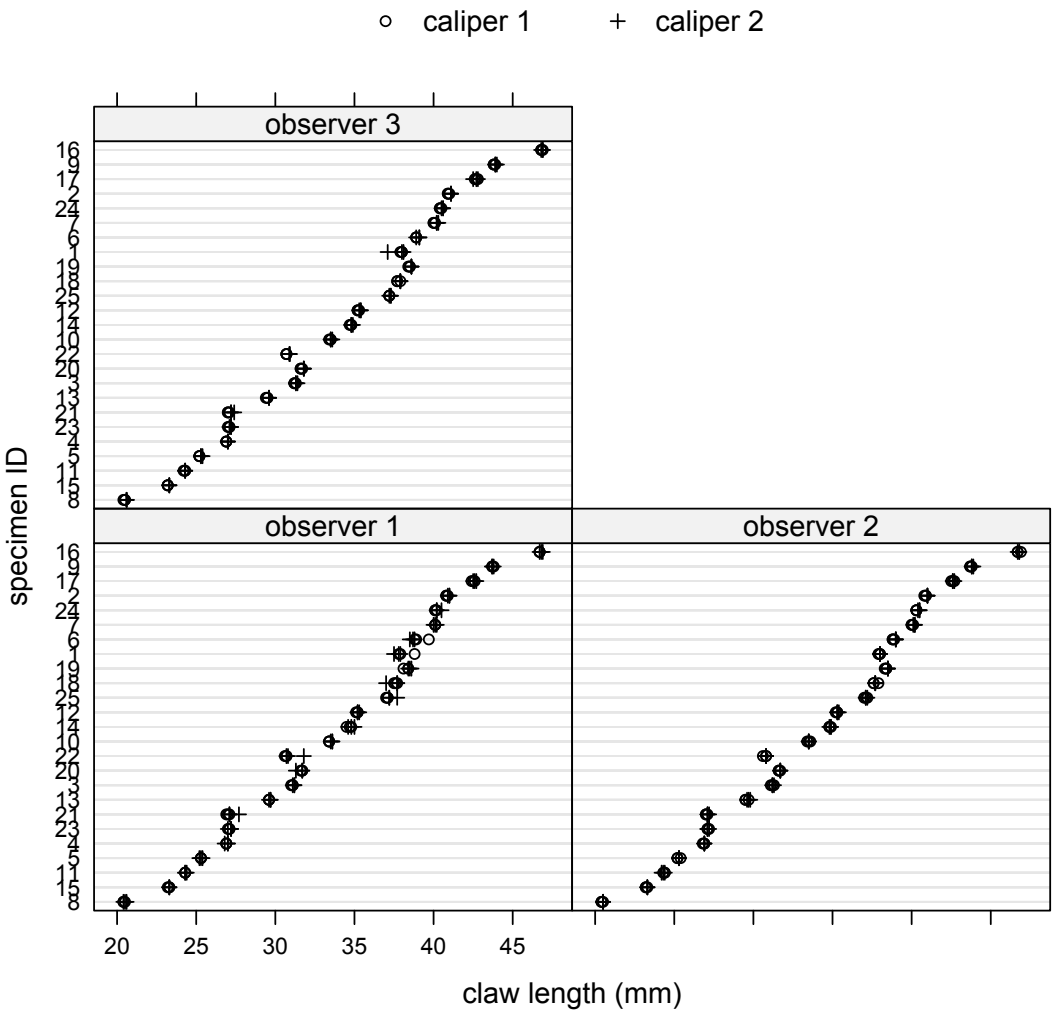

Figure 1: Trellis plot of crab data.

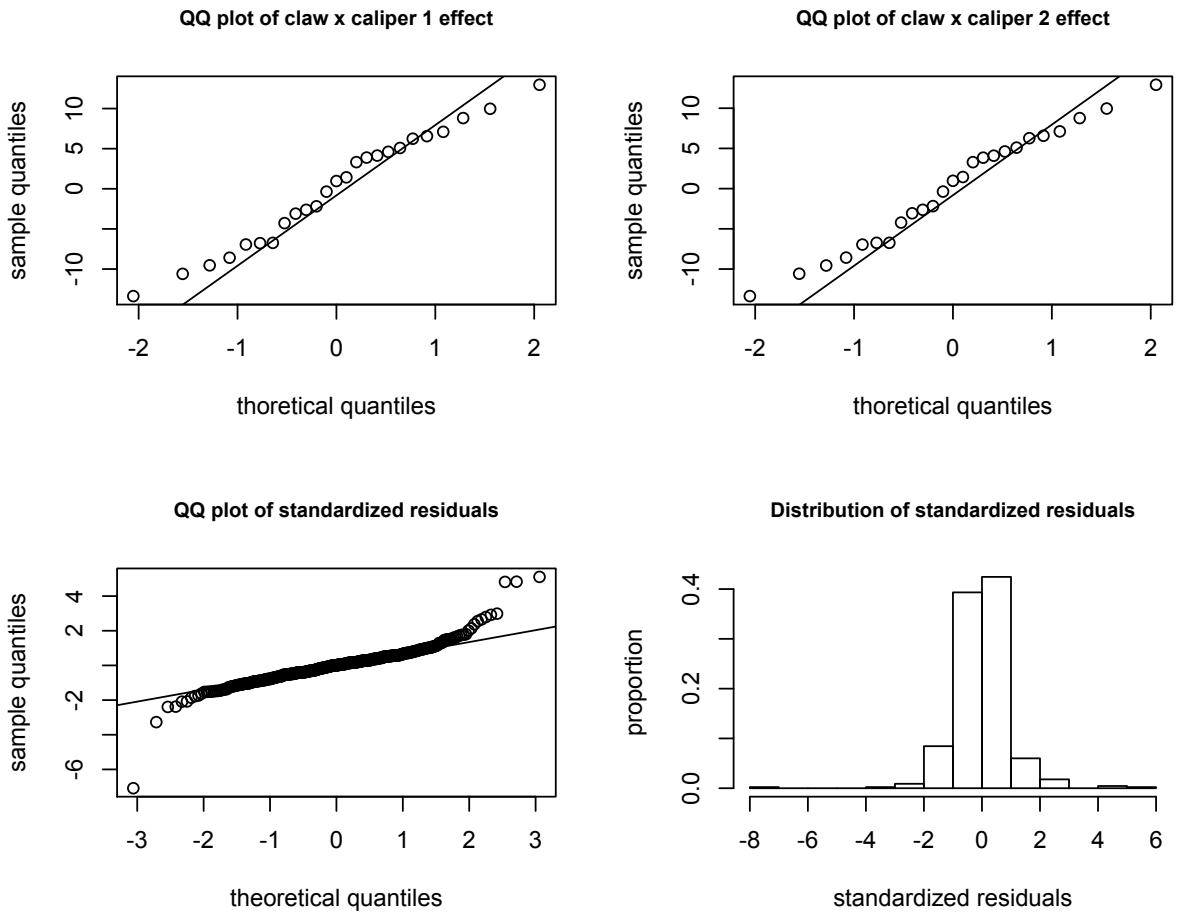

Figure 2: Normal QQ plots of predicted claw × caliper random effects and standardized residuals, and a histogram of standardized residuals. A line passing through the first and third quartiles is added in each QQ plot.

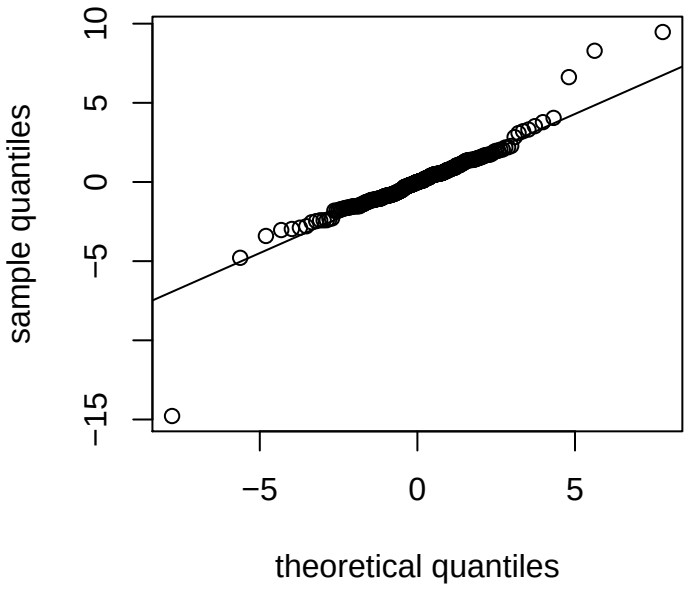

Figure 3: t QQ plot of standardized residuals for the crab data.

	$p_b = 0$				$p_b = 0.05$				$p_b = 0.10$				$p_b = 0.25$			
	p_e				p_e				p_e				p_e			
	0	.05	.10	.25	0	.05	.10	.25	0	.05	.10	.25	0	.05	.10	.25
	GSTMM															
$E[Y_{111}]$	0.0	0.0	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$E[Y_{121}]$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
σ_1^2	-0.1	-0.1	0.0	0.3	-0.1	-0.1	-0.1	0.2	-0.1	-0.1	-0.1	0.2	-0.1	-0.1	-0.1	0.2
σ_2^2	-0.1	-0.1	0.0	0.4	-0.1	-0.2	-0.1	0.3	-0.1	-0.2	-0.1	0.2	-0.1	-0.2	-0.1	0.3
ψ_1^2	-1.8	-1.9	-1.9	-2.0	-1.9	-2.3	-1.6	-1.1	-1.5	-1.9	-1.4	-0.7	9.0	8.3	9.0	10.3
ψ_2^2	-2.5	-2.6	-2.5	-2.7	-2.7	-3.2	-2.3	-1.5	-2.0	-2.6	-1.9	-0.9	12.2	11.2	12.2	14.0
ψ_{12}	-2.1	-2.1	-2.1	-2.2	-2.2	-2.6	-1.9	-1.3	-1.7	-2.2	-1.6	-0.9	10.1	9.3	10.0	11.5
	STMM															
$E[Y_{111}]$	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$E[Y_{121}]$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.2
σ_1^2	-0.1	0.1	0.2	0.7	-0.1	0.0	0.1	0.6	-0.1	0.0	0.1	0.6	-0.2	-0.1	0.0	0.6
σ_2^2	-0.1	0.1	0.2	1.1	-0.2	0.0	0.2	1.0	-0.2	-0.1	0.1	0.9	-0.3	-0.1	0.1	0.9
ψ_1^2	-1.6	-4.5	-6.0	-7.8	5.8	0.9	-0.7	-3.1	12.6	7.2	5.1	2.5	47.6	37.4	33.0	27.5
ψ_2^2	-2.2	-6.1	-8.1	-10.6	7.7	1.2	-1.1	-4.2	17.2	9.8	7.0	3.4	64.8	50.8	44.8	37.4
ψ_{12}	-1.8	-5.1	-6.8	-9.0	6.5	1.0	-1.0	-3.7	14.2	8.0	5.7	2.5	53.9	42.2	37.2	30.8

Table 1: Estimated biases of ML estimators when the simulation model, given by (14) and (15), is based on contaminated normal distributions with distant contamination pattern.

	$p_b = 0$				$p_b = 0.05$				$p_b = 0.10$				$p_b = 0.25$			
	p_e				p_e				p_e				p_e			
	0	.05	.10	.25	0	.05	.10	.25	0	.05	.10	.25	0	.05	.10	.25
$E[Y_{111}]$	1.0	1.1	1.2	1.4	1.0	1.1	1.2	1.4	1.0	1.0	1.1	1.3	0.9	0.9	1.0	1.2
$E[Y_{121}]$	1.0	1.1	1.2	1.4	1.0	1.1	1.1	1.4	1.0	1.0	1.1	1.3	0.9	0.9	1.0	1.2
σ_1^2	1.0	1.1	2.3	4.5	1.4	0.9	1.6	5.3	1.7	0.8	1.4	5.6	2.1	0.9	1.2	5.0
σ_2^2	1.0	1.1	2.2	4.5	1.3	0.8	1.5	5.5	1.7	0.8	1.3	5.7	2.1	1.0	1.2	5.2
ψ_1^2	1.0	1.7	2.2	3.0	1.8	1.0	0.8	0.9	4.7	2.3	1.6	1.1	13.8	9.5	6.8	4.1
ψ_2^2	1.0	1.7	2.2	2.9	1.7	1.0	0.8	0.9	4.6	2.3	1.6	1.0	13.8	9.5	6.7	4.1
ψ_{12}	1.0	1.7	2.2	3.0	1.8	1.0	0.8	0.9	4.6	2.2	1.6	1.0	13.8	9.5	6.8	4.1

Table 2: Efficiencies of GSTMM-based ML estimators relative to STMM when the simulation model, given by (14) and (15), is based on contaminated normal distributions with distant contamination pattern.

	$\nu_b = 4$				$\nu_b = 10$				$\nu_b = 30$				$\nu_b = \infty$			
	ν_e				ν_e				ν_e				ν_e			
	4	10	30	∞	4	10	30	∞	4	10	30	∞	4	10	30	∞
	GSTMM															
$E[Y_{111}]*$	-0.2	-0.2	-0.1	-0.1	0.2	-0.2	0.2	-0.4	-0.1	-0.3	-0.2	-0.3	-0.2	0.0	0.2	-0.4
$E[Y_{121}]*$	-0.3	-0.2	0.1	-0.2	0.4	-0.2	0.3	-0.4	0.0	-0.3	-0.1	-0.4	-0.1	-0.1	0.1	-0.4
σ_1^2*	0.2	0.0	0.0	-0.5	0.1	0.0	0.0	-0.5	0.1	0.0	0.0	-0.5	0.1	0.0	0.0	-0.5
σ_2^2*	0.2	-0.2	-0.1	-0.8	0.1	0.0	-0.1	-0.8	0.0	0.0	0.0	-0.9	0.0	-0.1	0.0	-1.0
ψ_1^2	0.9	0.7	0.6	0.9	0.0	0.1	-0.1	-0.1	0.0	-0.1	-0.1	0.3	-1.7	-1.8	-2.3	-2.2
ψ_2^2	1.4	1.4	1.0	1.1	0.1	0.4	0.2	-0.1	0.1	0.2	0.1	0.4	-2.1	-2.6	-3.2	-3.2
ψ_{12}	1.2	1.0	0.7	1.0	0.1	0.2	0.0	-0.1	0.0	0.1	0.0	0.4	-1.8	-2.1	-2.6	-2.6
	STMM															
$E[Y_{111}]*$	-0.3	-0.2	-0.1	0.0	0.2	-0.2	0.2	-0.4	-0.2	-0.3	-0.2	-0.3	-0.2	0.0	0.2	-0.4
$E[Y_{121}]*$	-0.4	-0.2	0.1	-0.1	0.4	-0.2	0.3	-0.4	0.0	-0.3	-0.1	-0.4	-0.1	-0.2	0.1	-0.4
σ_1^2*	2.6	0.6	-0.4	-0.8	3.3	1.1	0.0	-0.5	3.5	1.2	0.0	-0.5	3.6	1.2	0.0	-0.4
σ_2^2*	3.9	0.8	-0.7	-1.2	4.9	1.8	0.0	-0.8	5.2	1.9	0.0	-0.8	5.4	1.7	0.0	-1.0
ψ_1^2	11.2	15.8	17.4	18.0	1.0	4.6	5.0	5.0	-2.7	-0.2	0.1	0.6	-4.0	-1.9	-2.1	-2.0
ψ_2^2	15.4	22.0	24.0	24.3	1.5	6.7	7.0	6.8	-3.5	0.0	0.3	0.7	-5.2	-2.7	-2.9	-2.8
ψ_{12}	12.7	18.1	19.8	20.3	1.1	5.4	5.8	5.7	-3.0	-0.1	0.2	0.7	-4.4	-2.2	-2.4	-2.3

Table 3: Estimated biases of ML estimators when the simulation model, given by (14) and (17), is based on t-distributions. The entries for starred parameters have been multiplied by 10.

	$\nu_b = 4$				$\nu_b = 10$				$\nu_b = 30$				$\nu_b = \infty$			
	ν_e				ν_e				ν_e				ν_e			
	4	10	30	∞	4	10	30	∞	4	10	30	∞	4	10	30	∞
$E[Y_{111}]$	1.1	1.0	1.0	1.1	1.1	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.1	1.0	1.0	1.0
$E[Y_{121}]$	1.1	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.1	1.0	1.0	1.0
σ_1^2	4.6	1.7	1.2	1.3	6.9	1.9	1.0	1.0	8.4	2.1	1.0	1.0	8.8	2.2	1.0	1.0
σ_2^2	4.8	1.6	1.2	1.4	7.1	2.1	1.0	1.0	7.7	2.1	1.0	1.0	7.1	2.1	1.0	1.0
ψ_1^2	4.5	7.6	9.9	10	1.2	2.0	2.1	2.3	1.2	1.0	1.0	1.0	1.4	1.0	1.0	1.0
ψ_2^2	4.7	7.6	10	10	1.2	2.1	2.1	2.3	1.2	1.0	1.0	1.0	1.4	1.0	1.0	1.0
ψ_{12}	4.5	7.5	9.8	9.9	1.2	2.0	2.1	2.3	1.2	1.0	1.0	1.0	1.4	1.0	1.0	1.0

Table 4: Efficiencies of GSTMM-based ML estimators relative to STMM when the simulation model, given by (14) and (17), is based on t-distributions.

	NMM		GSTMM			NMM		GSTMM	
θ	$\hat{\theta}$	s.e.	$\hat{\theta}$	s.e.	θ	$\hat{\theta}$	s.e.	$\hat{\theta}$	s.e.
β_{11}	33.79	1.39	32.94	0.89	$\log(\sigma^2_{13})$	-5.56	0.21	-5.64	0.26
β_{12}	33.81	1.39	32.96	0.90	$\log(\sigma^2_{21})$	-3.10	0.17	-3.76	0.26
β_{13}	33.83	1.39	32.98	0.90	$\log(\sigma^2_{22})$	-6.06	0.20	-6.31	0.27
β_{21}	33.85	1.39	33.01	0.90	$\log(\sigma^2_{23})$	-4.07	0.17	-5.55	0.26
β_{22}	33.88	1.39	33.02	0.90	$\log(\psi^2_1)$	3.88	0.28	3.89	0.26
β_{23}	33.93	1.39	33.09	0.90	$\log(\psi^2_2)$	3.87	0.28	3.89	0.26
$\log(\sigma^2_{11})$	-3.54	0.17	-4.64	0.26	$z(\rho)$	5.95	0.33	6.30	0.43
$\log(\sigma^2_{12})$	-4.99	0.18	-5.19	0.24	$\log(\nu_e)$	n/a	n/a	1.28	0.32

Table 5: ML estimates of parameters and their standard errors (s.e.'s) that result from fitting the NMM and the assumed GSTMM to the crab data. Here $\rho = \psi_{12}/(\psi_1\psi_2)$ is the correlation between (b_{i1}, b_{i2}) and $z(\rho) = \tanh^{-1}(\rho)$ is the Fisher's z-transformation of ρ.