Bayesian inference in joint modelling of location and scale parameters of the t distribution for longitudinal data

Tsung-I Lin [a,b,c,⋆], Wan-Lun Wang [d]

[a] *Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan*

[b] *Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan*

[c] *Department of Public Health, China Medical University, Taichung 404, Taiwan*

[d] *Department of Statistics, Feng Chia University, Taichung 40724, Taiwan*

Abstract

This paper presents a fully Bayesian approach to multivariate t regression models whose mean vector and scale covariance matrix are modelled jointly for analyzing longitudinal data. The scale covariance structure is factorized in terms of unconstrained autoregressive and scale innovation parameters through a modified Cholesky decomposition. A computationally flexible *data augmentation* sampler coupled with the *Metropolis-within-Gibbs* scheme is developed for computing the posterior distributions of parameters. The Bayesian predictive inference for the future response vector is also investigated. The proposed methodologies are illustrated through a real example from a sleep dose-response study.

Key words: Cholesky decomposition; Data augmentation; Deviance information criterion; Maximum likelihood estimation; Outliers; Predictive distribution.

⋆ Corresponding author. Tel.: +886-4-22850420; fax: +886-4-22873028.

E-mail address: tilin@amath.nchu.edu.tw (T.I. Lin)

Revision invited by Journal of Statistical Planning and Inference (JSPI-S-10-00326-R1)

1. Introduction

During the last decade, methods of joint mean-covariance estimation for general linear models based on the modified Cholesky decomposition have received considerable attention in the literature (see, e.g., Pourahmadi, 1999) as a powerful tool for analyzing *nonstationary* longitudinal data. The key features with such an approach can be attributed to the facts that the covariance matrix is constrained to be nonnegative definite and has a statistically meaningful unconstrained parameterization. Pourahmadi (2000) provided a fast scoring method for obtaining the maximum likelihood (ML) estimates of unconstrained parameters and recommended using the Bayesian Information Criterion (BIC) to select the optimal model. Pan and MacKenzie (2003) later provided a data-driven technique for efficiently finding the global optimum of selected models. Recently, Cepeda and Gamerman (2004) presented a fully Bayesian approach to fitting this class of models via a generalization of the Metropolis-Hastings (M-H) algorithm (Hastings, 1970) of Cepeda and Gamerman (2000).

It is well known that statistical modelling built on the normality assumption tends to be vulnerable to the presence of outliers. In contrast, the multivariate t distribution has been recognized as a useful generalization of the normal distribution for robust inferences. For instance, Lange et al. (1989) introduced t-distributed regression model as a straightforward extension of the normal one. Pinheiro et al. (2001) considered a robust approach to linear mixed models (Larid and Ware, 1982) by assuming that both random effects and within-subject errors follow a multivariate t distribution. Further developments for analyzing multi-outcome longitudinal data in the context of multivariate t linear mixed models were considered by Wang and Fan (2010), among others. A Bayesian analysis of t linear mixed models of Pinheiro et al. (2001) was investigated by Lin and Lee (2007). Recently, Lin and Wang (2009) adopted a t-based joint modelling of mean and scale covariance for the analysis of

longitudinal data and demonstrated its robustness through several real examples.

A d-dimensional random vector $\boldsymbol{Y}$ is said to follow a multivariate t distribution with location vector $\boldsymbol{\mu}$, scale covariance matrix $\boldsymbol{\Sigma}$ and degrees of freedom ν, denoted by $\boldsymbol{Y} \sim T_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, if its density function is given by

$$f(\boldsymbol{Y}) = \frac{\Gamma\Big((\nu+d)/2\Big)}{\Gamma(\nu/2)(\pi\nu)^{d/2}}|\boldsymbol{\Sigma}|^{-1/2}\left[1+\frac{(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})}{\nu}\right]^{-(\nu+d)/2}.$$

The mean and covariance matrix of $\boldsymbol{Y}$ are $\boldsymbol{\mu}$ ($\nu > 1$) and $\nu\boldsymbol{\Sigma}/(\nu-2)$ ($\nu > 2$), respectively. If $\nu \to \infty$, then the distribution of $\boldsymbol{Y}$ reduces to $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Liu and Rubin (1995) offered some efficient EM-type algorithms for ML estimation of multivariate t distributions. For a comprehensive introduction to fundamental theories and characterizations of the multivariate t distribution along with its applications, the reader is referred to the monograph of Kotz and Nadarajah (2004).

With the fast development of Markov chain Monte Carlo (MCMC) methods, the Bayesian sampling-based approach has become an attractive alternative for drawing richer inferences. The advantages of adopting Bayesian methods involve the incorporation of prior information and the interpretation of parameters uncertainties. Bayesian inference often involves computing moments or quantiles of the posterior distribution, however, it requires a large amount of integration tasks. Typically, it is difficult to obtain the marginal posterior distributions by adopting the multidimensional integration. To address this issue, the MCMC method has been shown the most feasible tool from many important problems encountered in practice. In a Bayesian paradigm, the posterior inference can be accurately extracted from a large set of converged MCMC samples.

In this paper, we develop an efficient data augmentation (DA) sampler (Tanner and Wong, 1987) coupled with the idea of Metropolis-within-Gibbs scheme (Tierney, 1994) for implementing the Bayesian treatment of models proposed by Lin and Wang (2009). The priors are chosen to be weakly informative to avoid improper posterior distributions. Moreover, the selected prior distributions are conditionally

conjugate, which are very convenient when implementing the MCMC algorithm. Posterior predictive inferences for future values are also addressed.

The rest of the paper is organized as follows. In the next section, we formulate the model and review some of its relevant properties. In Sections 3, we present a full Bayesian inference for marginal posterior distributions as well as the predictive distribution of future observations. The proposed methodologies are illustrated with a real data set in Section 4 and a brief conclusion is given in the final section.

2. Model and estimation

Suppose there are N subjects in a longitudinal study. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}$ be the vector of n_i measurements on the ith subject collected at the time points $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})^{\mathrm{T}}$, and $\boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_{i1} & \cdots & \boldsymbol{x}_{in_i} \end{bmatrix}^{\mathrm{T}}$ be an $n_i \times p$ design matrix, where $\boldsymbol{x}_{ij}$ is a $p \times 1$ vector of covariate corresponding to Y_{ij}.

The model considered here is

$$\boldsymbol{Y}_i \sim T_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu) \quad (i = 1, \ldots, N), \tag{1}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^{\mathrm{T}}$ is the mean response vector for subject i. Following Pourahmadi (1999), we reparameterize $\boldsymbol{\Sigma}_i$ via the modified Cholesky decomposition as

$$\boldsymbol{L}_i \boldsymbol{\Sigma}_i \boldsymbol{L}_i^{\mathrm{T}} = \boldsymbol{D}_i, \tag{2}$$

where $\boldsymbol{D}_i = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_{n_i}^2\}$ and $\boldsymbol{L}_i = [\ell_{jk}]$ is a unit lower triangular matrix with the (j,k)th entry being $-\phi_{jk}$. It follows immediately that $\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{L}_i^{\mathrm{T}} \boldsymbol{D}_i^{-1} \boldsymbol{L}_i$. The parameters ϕ_{jk} and σ_j^2 in $\boldsymbol{L}_i$ and $\boldsymbol{D}_i$ are referred to as the *autoregressive parameters* and *scale innovation variances* of $\boldsymbol{\Sigma}_i$, respectively. Statistical interpretations for such reparameterization include (a) the below-diagonal entries of $\boldsymbol{L}_i$ are the negatives of

the autoregressive parameters, namely $-\phi_{jk}$, in

$$\hat{Y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{jk}(Y_{ik} - \mu_{ik}),$$

which is the linear least squares predictor of Y_{ij} based on its predecessors; (b) the diagonal entries of $\boldsymbol{D}_i$ are the scale innovation variances $\sigma_j^2 = c_\nu^{-1}\mathrm{var}(Y_{ij} - \hat{Y}_{ij})$, where $c_\nu = \nu/(\nu-2)$.

To make the dimension of unconstrained parameters μ_{ij}, ϕ_{jk} and $\log\sigma_j^2$ more parsimonious, we model them using covariates in the spirit of Pourahmadi (1999), namely, for $j = 1\ldots, n_i$ and $k = 1, \ldots, j-1$,

$$\mu_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, \quad \log\sigma_j^2 = \boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{\lambda}, \quad \phi_{jk} = \boldsymbol{z}_{jk}^{\mathrm{T}}\boldsymbol{\gamma}, \tag{3}$$

where $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are $p\times 1$, $q\times 1$ and $d\times 1$ vectors of unknown and unrestricted parameters, respectively. Note that $\boldsymbol{w}_j$ and $\boldsymbol{z}_{jk}$ are $q\times 1$ and $d\times 1$ parsimonious covariate vectors, which can usually be determined in terms of polynomial of measurement times t_{ij}'s with degrees of $q-1$ and $d-1$, respectively. Note that $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are assumed to be common for all $\boldsymbol{\Sigma}_i$'s for exhibiting the same covariance structure. Simply speaking, $\boldsymbol{\Sigma}_i$'s can be shrinkable or extendible, depending on i only through its dimension $n_i \times n_i$. Accordingly, the parameters of interest are $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \nu)^{\mathrm{T}}$. The log-likelihood function for the observed data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}|\boldsymbol{Y}) = &\sum_{i=1}^{N}\log\Gamma\Big(\frac{\nu+n_i}{2}\Big) - N\log\Gamma\Big(\frac{\nu}{2}\Big) - \frac{n}{2}\log(\pi\nu) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{\lambda}\\ &-\frac{1}{2}\sum_{i=1}^{N}(\nu+n_i)\log\Big(1+\frac{\Delta_i(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda})}{\nu}\Big),\end{aligned}$$

where $\Delta_i(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}) = (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$ and $n = \sum_{i=1}^{N} n_i$ is the total number of observations. The ML estimates $\hat{\boldsymbol{\theta}}$ are obtained as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\,\ell(\boldsymbol{\theta}|\boldsymbol{y}),$$

which generally have no analytical solution. Lin and Wang (2009) provided a fast scoring algorithm for carrying out ML estimation with the standard errors as a by-product.

It follows from the essential property of the multivariate t distribution that $\boldsymbol{Y}_i \sim T_{n_i}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ can be hierarchically expressed as $\boldsymbol{Y}_i \mid \tau_i \sim N_{n_i}\left(\boldsymbol{\mu}, \tau_i^{-1}\boldsymbol{\Sigma}\right)$ and $\tau_i \sim \Gamma(\nu/2, \nu/2)$, where $\Gamma(\alpha, \beta)$ stands for a gamma distribution with mean $\alpha\beta^{-1}$. Writing $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$, the complete data likelihood function is expressed as

$$L_c(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{\tau}) \propto \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{N}\Big(\sum_{j=1}^{n_i} \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\lambda} + \tau_i\Big(\nu + \Delta_i(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma})\Big)\Big)\Big\} \times \left[\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\right]^N \prod_{i=1}^{N} \tau_i^{(\nu+n_i-2)/2}, \tag{4}$$

where $|\boldsymbol{D}_i| = |\boldsymbol{L}_i||\boldsymbol{\Sigma}_i||\boldsymbol{L}_i^{\mathrm{T}}| = |\boldsymbol{\Sigma}_i|$, $\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} = \{r_{ij}\}_{j=1}^{n_i}$ and $\Delta_i(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = (\boldsymbol{r}_i - \boldsymbol{Z}_i\boldsymbol{\gamma})^{\mathrm{T}}\boldsymbol{D}_i^{-1}(\boldsymbol{r}_i - \boldsymbol{Z}_i\boldsymbol{\gamma})$ with $\boldsymbol{Z}_i = [\boldsymbol{z}(i,1) \ \cdots \ \boldsymbol{z}(i,n_i)]^{\mathrm{T}}$ and $\boldsymbol{z}(i,j) = \sum_{k=1}^{j-1} r_{ik}\boldsymbol{z}_{jk}$.

The ECME algorithm (Liu and Rubin, 1994), a generalization of EM (Dempster et al., 1977), extends the ECM algorithm of Meng and Rubin (1993) with the CM-steps by maximizing *either* the expected complete data log-likelihood function or the correspondingly constrained actual log-likelihood function, called the 'CML-step'. The main advantage of the ECME algorithm is that it not only preserves the nice features of EM and ECM, but also converges substantially faster than EM and ECM, as demonstrated by Liu and Rubin (1995) for ML estimation of multivariate t distribution with unknown degrees of freedom. Instead of the scoring method employed in Lin and Wang (2009), a computationally efficient ECME algorithm for obtaining the ML estimates $\hat{\boldsymbol{\theta}}$ in this context is given in a longer version of this paper which is available upon request from the authors.

3. Bayesian Methodology

3.1. *Prior settings, full conditionals and the DA sampler*

To accomplish a Bayesian setup of the model, one must specify a prior distribution for the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \nu)^{\mathrm{T}}$. Assuming that the prior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$ and ν are independent *a priori*, the joint prior density is

$$\pi(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\gamma})\pi(\nu).$$

When there is no prior information about $\boldsymbol{\theta}$, a convenient strategy of avoiding improper posterior distribution is to use diffuse proper priors. In this case, the posterior inference will be very close to those obtained by the ML method for large samples. In many circumstances, the Bayesian approach is more useful for situations in which the large sample properties cannot be applied for ML inference, especially when the collected samples are not sufficient enough.

The prior specifications adopted here are as follows:

$$\begin{aligned} &\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \quad \boldsymbol{\lambda} \sim N_q(\boldsymbol{\mu}_{\boldsymbol{\lambda}}, \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}), \\ &\boldsymbol{\gamma} \sim N_d(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}), \quad \log(1/\nu) \sim U(-10, 10), \end{aligned} \tag{5}$$

where the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, $\boldsymbol{\mu}_{\boldsymbol{\lambda}}$, $\boldsymbol{\mu}_{\boldsymbol{\gamma}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ are assumed to be known quantities. Typically, $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, $\boldsymbol{\mu}_{\boldsymbol{\lambda}}$ and $\boldsymbol{\mu}_{\boldsymbol{\gamma}}$ are zero vectors. To reflect the prior independence among $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$, the structures of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ can be taken as diagonal matrices with relatively large numbers on the main diagonal. In the later analysis, we take $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^4 \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}} = 10^4 \boldsymbol{I}_q$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = 10^4 \boldsymbol{I}_d$ for ensuring a reasonably flat prior distribution for $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$. Note that the prior for ν in (5) follows that employed in Liu and Rubin (1998) on the basis of vagueness.

Multiplying the complete data likelihood function (4) with the prior distributions in (5), the joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \nu, \boldsymbol{\tau})$ is given by

$$
\begin{aligned}
& p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \nu, \boldsymbol{\tau} \mid \boldsymbol{Y}) \\
& \propto \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{N} \Big(\sum_{j=1}^{n_i} \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{\lambda} + \tau_i \big(\nu + \Delta_i(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \big) \Big) \Big\} \prod_{i=1}^{N} \tau_i^{(\nu + n_i - 2)/2} \\
& \quad \times \left[\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \right]^N \exp\Big\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}}) \Big\} \\
& \quad \times \exp\Big\{ -\frac{1}{2} (\boldsymbol{\lambda} - \boldsymbol{\mu}_{\boldsymbol{\lambda}})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}^{-1} (\boldsymbol{\lambda} - \boldsymbol{\mu}_{\boldsymbol{\lambda}}) - \frac{1}{2} (\boldsymbol{\gamma} - \boldsymbol{\mu}_{\boldsymbol{\gamma}})^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\mu}_{\boldsymbol{\gamma}}) \Big\} J_\nu, \qquad (6)
\end{aligned}
$$

where $J_\nu = 1/\nu (0 < \nu < \infty)$ is the Jacobian of transforming $\log(1/\nu)$ to ν.

The DA algorithm introduced by Tanner and Wong (1987) has been shown to be a powerful tool for the computation of the entire posterior distribution. The key idea behind the algorithm is to augment the observed data $\boldsymbol{Y}$ to the latent vector $\boldsymbol{\tau}$. The DA algorithm consists of the imputation step (I-step) and the posterior step (P-step). At the kth iteration, the I-step imputes the missing data by drawing $\tau_i^{(k)}$ from the predictive distributions $p(\tau_i \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(k)})$. The P-step refers to generating $\boldsymbol{\theta}^{(k+1)}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \tau_i^{(k+1)})$. Under suitable regularity conditions, the simulated samples of $\tau_i^{(k)}$'s and $\boldsymbol{\theta}^{(k)}$ converge to their associated target distributions after a sufficiently long burn-in period.

The following proposition provides the required conditional posterior distributions used in the DA algorithm.

Proposition 1. *From (6), the full conditionals are given as follows (the notation "$\mid \cdots$" denotes conditioned on $\boldsymbol{Y}$ and all other variables):*

$$
\begin{aligned}
& \tau_i \mid \cdots \sim \Gamma\left(\frac{n_i + \nu}{2}, \frac{\nu + \Delta_i(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma})}{2} \right), && (7) \\
& \boldsymbol{\beta} \mid \cdots \sim N_p(\boldsymbol{b}^*, \mathbf{B}^*), && (8) \\
& \boldsymbol{\gamma} \mid \cdots \sim N_d(\boldsymbol{g}^*, \boldsymbol{G}^*), && (9)
\end{aligned}
$$

where

$$\boldsymbol{b}^* = \mathbf{B}^*\Big(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{i=1}^{N}\tau_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Y}_i\Big),$$
$$\mathbf{B}^* = \Big(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \sum_{i=1}^{N}\tau_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\Big)^{-1},$$
$$\boldsymbol{g}^* = \boldsymbol{G}^*\Big(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \sum_{i=1}^{N}\tau_i \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{D}_i^{-1}\boldsymbol{r}_i\Big),$$
$$\boldsymbol{G}^* = \Big(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \sum_{i=1}^{N}\tau_i \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{D}_i^{-1}\boldsymbol{Z}_i\Big)^{-1}.$$

The full conditional distributions of $\boldsymbol{\lambda}$ and ν given below are not of standard forms.

$$p(\boldsymbol{\lambda} \mid \cdots)$$
$$\propto \exp\Big\{-\frac{1}{2}\sum_{i=1}^{N}\Big(\sum_{j=1}^{n_i}\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\lambda} + \tau_i\Delta_i(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma})\Big) - \frac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{\mu}_{\boldsymbol{\lambda}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\lambda}-\boldsymbol{\mu}_{\boldsymbol{\lambda}})\Big\} \quad (10)$$

and

$$p(\nu \mid \cdots) \propto \left[\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\right]^N \Big(\prod_{i=1}^{N}\tau_i^{\nu/2}\Big)\exp\Big\{-\frac{\nu}{2}\sum_{i=1}^{N}\tau_i\Big\}J_\nu. \quad (11)$$

Proof: The proof follows from standard calculations and hence is omitted.

In the simulation process, samples for τ_i's and $\boldsymbol{\theta}$ are alternately generated. In summary, the implementation of DA algorithm proceeds as follows:

I-Step: Impute τ_i from its conditional posterior in (7) for $i = 1, \ldots, N$.

P-Step:

1. Draw $\boldsymbol{\beta}$ from its full conditional posterior in (8).

2. Draw $\boldsymbol{\gamma}$ from its full conditional posterior in (9).

3. Update $\boldsymbol{\lambda}$ from (10) via the M-H algorithm.

4. Update ν from (11) via the M-H algorithm.

To elaborate on P-Step 3 of the above algorithm, a multivariate normal distribution $N_q(\boldsymbol{\lambda}^{(k)}, c^2\hat{\boldsymbol{I}}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1})$ is chosen as the *random walk* jumping distribution $J(\tilde{\boldsymbol{\lambda}}|\boldsymbol{\lambda}^{(k)})$, where the scale c can be taken around $2.4/\sqrt{q}$ (Gelman et al., 1996). Moreover, $\hat{\boldsymbol{I}}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1}$ is the inverse information matrix evaluated at the ML estimate of $\boldsymbol{\lambda}$. Note that the

information matrix of $\boldsymbol{\lambda}$ is

$$\boldsymbol{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} = \frac{1}{2}\sum_{i=1}^{N}\frac{\nu+n_i}{\nu+n_i+2}\left\{\sum_{j=1}^{n_i}\boldsymbol{w}_j\boldsymbol{w}_j^{\mathrm{T}} - \frac{\left(\sum_{j=1}^{n_i}\boldsymbol{w}_j\right)\left(\sum_{j=1}^{n_i}\boldsymbol{w}_j^{\mathrm{T}}\right)}{\nu+n_i}\right\}.$$

Because the proposal distribution is symmetric, it follows from the Metropolis jumping rule (Metropolis et al., 1953) that

$$\boldsymbol{\lambda}^{(k+1)} = \begin{cases} \tilde{\boldsymbol{\lambda}} & \text{with probability} \quad \min\{1,\alpha\}, \\ \boldsymbol{\lambda}^{(k)} & \text{otherwise,} \end{cases}$$

where

$$\alpha = \frac{p(\tilde{\boldsymbol{\lambda}} \mid \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}, \nu^{(k)}, \boldsymbol{\tau}^{(k+1)}, \boldsymbol{Y})}{p(\boldsymbol{\lambda}^{(k)} \mid \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}, \nu^{(k)}, \boldsymbol{\tau}^{(k+1)}, \boldsymbol{Y})}.$$

For sampling ν in P-Step 4, we first transform ν to $\nu^* = \log(1/\nu)$ and then apply the M-H algorithm to the function $g(\nu^* \mid \cdots) = p(\nu(\nu^*) \mid \cdots)e^{-\nu^*}$. The jumping distribution $J(\nu^{*(k+1)}|\nu^{*(k)})$ can be chosen as a truncated normal distribution with mean $\nu^{*(k)}$ and variance $2.4^2\hat{\sigma}^2_{\nu^*}$ before truncation, and truncated region $\mathbb{A} = (-10, 10)$, where $\hat{\sigma}^2_{\nu^*} = \hat{\nu}^{-2}\hat{I}_{\nu}^{-1}$ and $\hat{I}_{\nu}^{-1}$ is the inverse information matrix of ν evaluated at $\hat{\nu}$. Note that I_ν is given as

$$I_\nu = \frac{1}{4}\sum_{i=1}^{N}\left\{\psi\left(\frac{\nu}{2}\right) - \psi\left(\frac{\nu+n_i}{2}\right) - \frac{2n_i(\nu+n_i+4)}{\nu(\nu+n_i)(\nu+n_i+2)}\right\},$$

where $\psi(x) = d^2\log\Gamma(x)/dx^2$ is the trigamma function. Notably, at the $(k+1)$st iteration, one can generate a truncated normal variate (Gelfand et al., 1992) by

$$\nu^{*(k+1)} = \nu^{*(k)} + \sigma^{(k)}_{\nu^*}\Phi^{-1}\left\{\Phi\left(\frac{-10-\nu^{*(k)}}{\sigma^{(k)}_{\nu^*}}\right) + U\left[\Phi\left(\frac{10-\nu^{*(k)}}{\sigma^{(k)}_{\nu^*}}\right) - \Phi\left(\frac{-10-\nu^{*(k)}}{\sigma^{(k)}_{\nu^*}}\right)\right]\right\},$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0,1)$ and U denotes a random uniform $(0,1)$ variate. Note that $\nu^{*(k+1)}$ is accepted according to the probability $\min\left\{1, \frac{g(\nu^{*(k+1)}|\cdots)J(\nu^{*(k)}|\nu^{*(k+1)})}{g(\nu^{*(k)}|\cdots)J(\nu^{*(k+1)}|\nu^{*(k)})}\right\}$. If accepted, invert it back to $\nu^{(k+1)} = \exp\{-\nu^{*(k+1)}\}$. Otherwise, set $\nu^{(k+1)} = \nu^{(k)}$.

The above procedure is processing until the convergence is achieved. Notably, the

model uncertainty or any posterior inference of interest can be effectively taken into account by converged Monte Carlo samples, say $\{\boldsymbol{\theta}^{(k)}\}_{k=1}^{K}$. For example, an approximate posterior mean of $\boldsymbol{\theta}$ can be estimated by $\bar{\boldsymbol{\theta}} = K^{-1}\sum_{k=1}^{K}\boldsymbol{\theta}^{(k)}$ and its posterior variance-covariance can be approximated by $\sum_{k=1}^{K}(\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}})^{\mathrm{T}}/(K-1)$.

3.2. *Posterior predictive inference for future values*

We consider the extended prediction of $\tilde{\boldsymbol{y}}_i$, a $g \times 1$ future observations of $\boldsymbol{Y}_i$. The posterior predictive density of $\tilde{\boldsymbol{y}}_i$ is given by

$$p(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{Y}_i) = \int p(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{Y}_i, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{Y}_i) d\boldsymbol{\theta}. \tag{12}$$

The integration in (12) is usually intractable, but can be easily approximated by constructing a set of stationary MCMC samples.

Let $\tilde{\boldsymbol{x}}_i$ be a $g \times p$ matrix of prediction regressors corresponding to $\tilde{\boldsymbol{y}}_i$ such that $E(\tilde{\boldsymbol{y}}_i) = \tilde{\boldsymbol{x}}_i\boldsymbol{\beta}$. Let $\tilde{\boldsymbol{w}}_j$ and $\tilde{\boldsymbol{z}}_{jk}$, respectively, denote a $q\times 1$ and a $d\times 1$ covariate vector associated with $\tilde{\boldsymbol{y}}_i$ such that $\log\sigma_j^2 = \tilde{\boldsymbol{w}}_j^{\mathrm{T}}\boldsymbol{\lambda}$ and $\phi_{jk} = \tilde{\boldsymbol{z}}_{jk}^{\mathrm{T}}\boldsymbol{\gamma}$ for $j = n_i+1, \ldots, n_i+g$ and $k = 1, \ldots, j-1$. Further, we assume that the joint distribution for $(\boldsymbol{Y}_i^{\mathrm{T}}, \tilde{\boldsymbol{y}}_i^{\mathrm{T}})^{\mathrm{T}}$ follows an (n_i+g)-variate t distribution, given by

$$\begin{bmatrix} \boldsymbol{Y}_i \\ \\ \tilde{\boldsymbol{y}}_i \end{bmatrix} \sim T_{n_i+g}(\boldsymbol{X}_i^*\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^*, \nu)$$

with $\boldsymbol{X}_i^* = (\boldsymbol{X}_i^{\mathrm{T}}, \tilde{\boldsymbol{x}}_i^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\Sigma}_i^{*-1} = \boldsymbol{L}_i^{*\mathrm{T}}\boldsymbol{D}_i^{*-1}\boldsymbol{L}_i^*$, where $\boldsymbol{L}_i^*$ is a $(n_i+g)\times(n_i+g)$ lower triangular matrix having $-\phi_{jk}$ at the (j,k)th position, $k<j$, $j = 1, \ldots, n_i+g$, and $\boldsymbol{D}_i^* = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_{n_i}^2, \sigma_{n_i+1}^2, \ldots, \sigma_{n_i+g}^2\}$. More specifically,

$$\boldsymbol{L}_i^* = \begin{bmatrix} & \boldsymbol{L}_i & & \mathbf{0} \\ & & & \\ -\tilde{\boldsymbol{z}}_{n_i+1,1}^{\mathrm{T}}\boldsymbol{\gamma} & \ddots & -\tilde{\boldsymbol{z}}_{n_i+1,n_i}^{\mathrm{T}}\boldsymbol{\gamma} & 1 \\ \vdots & \ddots & \cdots & \ddots \quad \ddots \\ -\tilde{\boldsymbol{z}}_{n_i+g,1}^{\mathrm{T}}\boldsymbol{\gamma} & \vdots & \cdots & -\tilde{\boldsymbol{z}}_{n_i+g,n_i+g-1}^{\mathrm{T}}\boldsymbol{\gamma} \quad 1 \end{bmatrix}$$

and

$$\boldsymbol{D}_i^* = \mathrm{diag}\Big\{ \exp(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{\lambda}), \ldots, \exp(\boldsymbol{w}_{n_i}^{\mathrm{T}}\boldsymbol{\lambda}), \exp(\tilde{\boldsymbol{w}}_{n_i+1}^{\mathrm{T}}\boldsymbol{\lambda}), \ldots, \exp(\tilde{\boldsymbol{w}}_{n_i+g}^{\mathrm{T}}\boldsymbol{\lambda}) \Big\}.$$

Let $\boldsymbol{\Sigma}_i^*$ be partitioned conformably with the dimension of $(\boldsymbol{Y}_i^{\mathrm{T}},\ \tilde{\boldsymbol{y}}_i^{\mathrm{T}})^{\mathrm{T}}$. Therefore, we have

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{13}$$

Here we suppress the subscript i of the partitioned matrices in (13) for notational convenience. Note that $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^{\mathrm{T}}$. Making use of the conditional property concerning the multivariate t distribution, we have

$$\tilde{\boldsymbol{y}}_i \mid (\boldsymbol{Y}_i, \boldsymbol{\theta}) \ \sim \ T_g\left(\boldsymbol{\mu}_{2\cdot 1}, \frac{\nu + \Delta_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})}{\nu + n_i} \boldsymbol{\Sigma}_{22\cdot 1}, \nu + n_i \right),$$

where $\boldsymbol{\mu}_{2\cdot 1} = \tilde{\boldsymbol{x}}_i\boldsymbol{\beta} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$ and $\boldsymbol{\Sigma}_{22\cdot 1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

Let $\boldsymbol{\theta}^{(k)}$ be the generated sample at the kth iteration of the DA sampler when the convergence is achieved. We can obtain the approximate predictive distribution of $\tilde{\boldsymbol{y}}_i$ using the Rao-Blackwellization Theorem (Gelfand and Smith, 1990). That is,

$$p(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{Y}) \approx \frac{1}{K} \sum_{k=1}^{K} t_g\Big(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{\mu}_{2\cdot 1}^{(k)}, \frac{\nu^{(k)} + \Delta_i(\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\gamma}^{(k)})}{\nu^{(k)} + n_i} \boldsymbol{\Sigma}_{22\cdot 1}^{(k)}, \nu^{(k)} + n_i \Big),$$

where $\boldsymbol{\mu}_{2\cdot 1}^{(k)} = \tilde{\boldsymbol{x}}_i\boldsymbol{\beta}^{(k)} + \boldsymbol{\Sigma}_{21}^{(k)}\boldsymbol{\Sigma}_{11}^{-1(k)}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(k)})$ and $t_g(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denotes the density of $T_g(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ evaluated at $\tilde{\boldsymbol{y}}_i$.

4. A Numerical Illustration

Sleep is a fundamental and physiological demand of mankind. Recent investigations indicate that sleep deprivation and chronic sleep restriction would impact health, safety and productivity. In general, it would result in reduction of the neural behavior, irregular influence on the immune function, increased fatigue, decreased alertness and impaired performance in a variety of cognitive psychomotor tests.

As an illustration, we apply the proposed methods to the 3-hour group of a sleep dose-response study for 18 volunteer participants. They spent a total of 14 days in the Johns Hopkins Bayview General Clinical Research Center (GCRC, Baltimore, MD, USA) to complete this study. The first 2 days were adaptation and training (T1 and T2) and the third served as baseline (B). During the three days, all participants were required to be in bed from 23:00 to 07:00, i.e., 8-hour time in bed (TIB). Beginning on the fourth day, they were restricted to 3-hour TIB from 04:00 to 07:00 for the next 7 days (E1-E7). On the 11th day and lasting to the 13th day (R1-R3), they were requested 8 hours daily TIB from 23:00 to 07:00 for recovery. They are also placed a final night of 8-hour TIB to release from the study, but no testing occurred on the final recovery day (R4). The experimental design and TIB schedule are shown in Figure 1 of Belenky et al. (2003).

Of these 18 subjects, they received a series of cognitive and alterness tests, e.g., psychomotor vigilance task (PVT), polysomnography (PSG) measures and sleep latency test (SLT). The PVT is a generally acknowledged test to measure reaction time in ms (millisecond) to a visual stimulus, a thumb-operated and hand-held device (Dinges and Powell, 1985). Subjects heeded the LED timer display on the device and pressed the response button as soon as possible after the emergence of the visual stimulus. The group performed the PVT test four times per day (09:00, 12:00, 15:00 and 21:00). In this example, the response variable Y_{ij} is the average reaction time measured from the PVT test over the first 10 days for the 18 subjects.

This data set can be extracted from the database of `lme4` R-package (Bates, 2007). A detailed description of participants, design and procedures of the study can be found in Balkin et al. (2000), Belenky et al. (2003) and Balkin et al. (2004).

Table 1 about here

Table 1 lists the sample variances and correlations among the repeated measurements of the data. Apparently, the variances tend to increase across time after the baseline. Furthermore, the correlations exhibit an irregular pattern within the same lagged times, suggesting that the data have a non-stationary covariance structure.

Figure 1 about here

Figure 1(a) displays the trajectories of average reaction times evolved over an equally spaced 10-day period along with its mean profile and ± 1 standard deviations across days. A few of subjects appear to have sudden jumps and drops in experimental days and hence are suspected to be discordant outliers. This indicates that normal-based models could be inappropriate for this data set.

For the specification of design matrices, because the trend of population mean profile evolves linearly over time, it is reasonable to use a 1st-degree polynomial in time to model the mean responses. The covariates are thus taken as $\boldsymbol{x}_{ij} = (1, t_{ij})^{\mathrm{T}}$, where $t_{ij} = j$ for $i = 1, \ldots, 18$ and $j = 1, \ldots, 10$. In addition, Figures 1(b) and 1(c) suggest that ϕ_{jk}'s and $\log \sigma_j^2$'s could be well suited to a cubic or a high-order polynomial functions in lags. Following Pourahmadi (2000), we use the Poly(q, d) as a shorthand for imposing two distinct polynomials of lagged times j and $j-k$ with degrees q and d for $\log \sigma_j^2$ and ϕ_{jk}, respectively. Specifically, the covariates $\boldsymbol{z}_{jk}$ and $\boldsymbol{w}_j$ are chosen as

$$\begin{aligned}
&\boldsymbol{w}_j = (1, j, j^2, \ldots, j^q)^{\mathrm{T}}, \qquad j = 1, \ldots, 10;\ k = 1, \ldots, j-1, \\
&\boldsymbol{z}_{jk} = (1, (j-k), (j-k)^2, \ldots, (j-k)^d)^{\mathrm{T}}.
\end{aligned}$$

For parsimony, the largest degrees q and d are limited to 5. The best pair of degrees, say (q^*, d^*), will satisfy $(q^*, d^*) = \arg\min_{(q,d)}\{\text{DIC}(q, d)\}$.

Figure 2 about here

For the models fitted by MCMC sampling, we ran five parallel chains with the starting values of each chain drawn independently from their prior distributions. Based on these parallel chains, the multivariate potential scale reduction factor (MPSRF) suggested by Brooks and Gelman (1998) was used to assess the validity of convergence. Figure 2 displays the MPSRF plots for some of the selected models. Inspection of these MPSRFs suggests that the convergence occurred after 2,500 iterations. Therefore, we discard the first 2500 iterations for each chain as the "burn-in" samples. Furthermore, the converged MCMC realizations of each chain are collected for every 20 iterations, in order to obtain approximately independent samples. We also conduct the analysis with different values of the hyperparameters, resulting in very similar results.

Table 2 about here

Bayesian model comparison is based on the deviance information criterion (DIC) advocated by Spiegelhalter et al. (2002). The DIC is defined as

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D = 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) = D(\bar{\boldsymbol{\theta}}) + 2p_D,$$

where $\overline{D(\boldsymbol{\theta})} = E_{\theta|\boldsymbol{Y}}[-2\ell(\boldsymbol{\theta} \mid \boldsymbol{Y})]$ is the posterior expectation of the deviance and $D(\bar{\boldsymbol{\theta}})$ is the deviance evaluated at the posterior means of the parameters. The penalty term, $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$, is regarded as the *effective number of parameters*. Note that the DIC can be interpreted as a Bayesian measure of fit penalized by twice the effective number of parameters p_D. By definition, large value of DIC corresponds to a poor fit. The DIC values associated with the competitive models are included in Table 2. Judging from this table, all DIC values of t models are smaller than their corresponding normal counterparts, confirming the appropriateness of the use of t

distributions. The fitted normal models are obtained by setting $\nu = 1,000$ when performing the MCMC algorithm. Results show that the Poly(3,4) is the preferred choice because it has the smallest DIC values for both the normal and t models.

Table 3 about here

Posterior inferences for Poly(3,4) summarized by MCMC samples, including the means, standard deviations together with 2.5% and 97.5% quantiles, are shown in Table 3. As can be seen, the 95% posterior interval of ν is (2.8, 15.1), signifying the presence of longer-than-normal errors.

Figure 3 about here

Figure 3 displays the logarithm of fitted innovation variances for the normal- and t- Poly (3,4) models along with their 95% credible intervals. In light of the graphical visualization, it appears that the t-fitted $\log \sigma_j^2$, which is a cubic function of time, adapts the pattern of sample estimates more closely than does the normal one.

Figure 4 about here

To detect possible outlying observations, Figure 4 shows the boxplots of MCMC samples of τ_i for $i = 1, \ldots, 18$. As pointed out by Wakefield *et al.* (1994), the sampled τ_i's can be used as concise indicators for detecting outliers with prior expectation of one. When the value of τ_i is substantially lower than one, it gives a strong indication that the ith participant should be regarded as an outlier in the population. Figure 4 reveals that subjects 1 and 6 are suspected outliers since none of 95% upper posterior limits exceeds 1. We marked the two subjects in colored lines in Figure 1(a). It looks fairly clear that their growth curve patterns are quite different from the others. For instance, subject 6 exhibits a sudden jump at day 6 then a sudden drop at day 7.

5. Conclusion

This paper presents a fully Bayesian approach to jointly modelling the mean and scale covariance structures for longitudinal data in the framework of multivariate t regression models. We have developed a workable DA sampler that enables practitioners to simulate the posterior distributions of model parameters. We also demonstrated how the model provides flexibility in analyzing heavy-tailed longitudinal data from a Bayesian perspective. The proposed approach allows the user to analyze real-world data in a wide variety of considerations. Numerical results illustrated in Section 4 indicate that the t model for the sleep data is evidently more adequate than the normal one. In particular, the graphical Bayesian outputs provide both easily understood inferential summaries and informative diagnostic aids for detecting outliers. Future researches along this line include a joint mean-covariance approach to multivariate skew normal models (Azzalini and Capitanio, 1999; Lin and Lee, 2008) or multivariate skew t models (Azzalini and Capitaino, 2003; Ho and Lin, 2010) to cover more possible classes or patterns for longitudinal data.

Acknowledgements

The authors would like to express their deepest gratitude to the Chief Editor, the Associate Editor and one anonymous referee for carefully reading the paper and for their great help in improving the paper. We are also grateful to Miss Wen-Chi Lin for her kind and skilful assistance in the initial simulation study. This research was supported by the National Science Council of Taiwan (Grant Nos. NSC97-2118-M-005-001-MY2 and NSC99-2118-M-035-004).

References

Azzalini, A., Capitaino, A., 1999. Statistical applications of the multivariate skew-normal distribution. J. Roy. Statist. Soc. B 61, 579–602.

Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J. Roy. Stat. Soc. B 65, 367–389.

Bates, D.M., 2007. lme4 R package. `http://cran.r-project.org.`

Belenky, G., Wesensten, N.J., Thorne, D.R., Thomas, M.L., Sing, H.C., Redmond, D.P., Russo, M.B., Balkin, T.J., 2003. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. J. Sleep Res. 12, 1–12.

Balkin, T.J., Thorne, D., Sing, H., Thomas, M., Redmond, D., Wesensten, N., Russo, M., Williams, J., Hall, S., Belenky, G., 2000. Effects of sleep schedules on commercial motor vehicle driver performance. Report MC-00-133, National Technical Information Service, U.S. Dept. of Transportation, Springfield, VA.

Balkin, T.J., Bliese, P.D., Belenky, G., Sing, H., Thorne, D.R., Thomas, M., Redmond, D.P., Russo, M., Wesensten, N.J., 2004. Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. J. Sleep Res. 13, 219–227.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Statist. 7, 434–455.

Cepeda, E.C., Gamerman, D., 2000. Bayesian modeling of variance heterogeneity in normal regression models. Braz. J. Probab. Stat. 14, 207–221.

Cepeda, E.C., Gamerman, D., 2004. Bayesian modeling of joint regressions for the mean and covariance matrix. Biom. J. 46, 430–440.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc. Ser. B 39, 1–38.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling based approaches to calculate marginal densities. J. Am. Stat. Assoc. 85, 398–409.

Gelfand, A.E., Smith, A.F.M., Lee, T.M., 1992. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. J. Am. Stat. Assoc. 85, 523–532.

Gelman, A., Robert, G., Gilks. W., 1996. Efficient Metropolis jumping rules. In Bayesian Statistics 5 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford University Press, New York

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov Chains and their applications. Biometrika 57, 97–109.

Ho, H.J., Lin, T.I., 2010. Robust linear mixed models using the skew t distribution with application to schizophrenia data. Biom. J. 52, 449–469.

Kotz, S, Nadarajah, S., 2004. Multivariate t Distributions and Their Applications. third ed. Cambridge University Press, New York

Lange, K.L., Little, R.J.A., Taylor, J.M.G., 1989. Robust statistical modeling using the t distribution. J. Am. Stat. Assoc. 84, 881–896.

Laird, N.M., Ware, J.H., (1982). Random effects models for longitudinal data. Biometrics 38:963–974

Lin, T.I., Lee, J.C., 2007. Bayesian analysis of hierarchical linear mixed modeling using the multivariate t distribution. J. Statist. Plann. Inference 137, 484–495.

Lin, T.I., Lee, J.C., 2008. Estimation and prediction in linear mixed models with skew normal random effects for longitudinal data. Stat. Med. 27, 1490–1507.

Lin, T.I., Wang, Y.J., 2009. A robust approach to joint modeling of mean and scale covariance for longitudinal data. J. Statist. Plann. Inference 139, 3013–3026.

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

Liu, C., Rubin, D.B., 1995. ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statist. Sinica. 5, 19–39.

Liu, C.H., Rubin, D.B., 1998. Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. Biometrika 85, 673–688.

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

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculation by fast computing machines. J. Chem. Phy. 21, 1087–1092.

Pan, J., MacKenzie, G., 2003. On modelling mean-covariance structures in longitudinal studies. Biometrika 90, 239–244

Pinheiro, J.C., Liu, C.H., Wu, Y.N., 2001. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. J. Comput. Graph. Statist. 10, 249–276.

Pourahmadi, M., 1999. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. Biometrika 86, 677–690.

Pourahmadi, M., 2000. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. Biometrika 87, 425–435.

Spiegelhalter, D.J., Best N.G., Carlin, B.P., Linde, A.V.D., 2002. Bayesian measures of model complexity and fit. J. Roy. Stat. Soc. B 64, 583–639.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation (with discussion). J. Am. Stat. Assoc. 82, 528–550.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Stat. 22, 1701–1728.

Wakefield, J.C., Smith, A.F.M., Racine-Pooh, A., Gelfand, A.E., 1994. Bayesian analysis of linear and non-linear model by using Gibbs sampler. App. Stat. 43, 201–221.

Wang, W.L., Fan, T.H., 2010. Estimation in multivariate t linear mixed models for multiple longitudinal data. Statist. Sinica, to appear.

Table 1
Sample variances (along the main diagonal) and correlations (below the main diagonal).

t	1	2	3	4	5	6	7	8	9	10
1	1032.3									
2	0.737	1117.6								
3	0.470	0.770	868.7							
4	0.464	0.741	0.875	1509.9						
5	0.449	0.651	0.694	0.914	1809.5					
6	0.372	0.529	0.492	0.722	0.854	2680.1				
7	0.222	0.315	0.454	0.675	0.749	0.743	3990.9			
8	0.493	0.476	0.590	0.600	0.695	0.690	0.703	2510.4		
9	0.329	0.395	0.406	0.596	0.745	0.901	0.729	0.762	3624.0	
10	0.516	0.546	0.423	0.566	0.716	0.838	0.460	0.659	0.881	4487.2

Table 2

$\overline{D(\boldsymbol{\theta})}$, p_D and DIC values for various Poly(q, d) models.

(q, d)	$\overline{D(\boldsymbol{\theta})}$		p_D		DIC	
	normal	t	normal	t	normal	t
(2,2)	1734.648	1712.066	7.878	8.757	1742.526	1720.823
(2,3)	1734.430	1711.692	8.771	9.681	1743.202	1721.373
(2,4)	1732.630	1710.692	9.649	10.711	1742.279	1721.403
(2,5)	1731.979	1711.734	9.917	12.365	1741.895	1724.099
(3,2)	1727.604	1706.965	8.615	9.726	1736.219	1716.691
(3,3)	1726.363	1706.388	9.707	10.931	1736.070	1717.319
(3,4)*	1720.909	1702.673	10.665	11.733	1731.575	1714.406
(3,5)	1720.285	1702.851	11.867	12.993	1732.152	1715.844
(4,2)	1728.582	1708.131	9.717	10.880	1738.299	1719.011
(4,3)	1727.568	1707.088	10.812	11.691	1738.381	1718.779
(4,4)	1721.941	1703.486	11.953	12.861	1733.894	1716.347
(4,5)	1721.193	1703.385	12.745	13.718	1733.937	1717.103

* the best chosen model

Table 3

Summarized posterior results for the normal- and t- Poly$(3, 4)$ models.

Parameter	normal				t			
	Mean	S.D.	$Q_{0.025}$	$Q_{0.975}$	Mean	S.D.	$Q_{0.025}$	$Q_{0.975}$
$\hat{\beta}_0$	242.278	6.999	227.728	256.031	240.163	7.423	225.288	254.503
$\hat{\beta}_1$	9.907	1.493	7.117	12.988	8.983	1.526	5.832	11.859
$\hat{\lambda}_0$	8.393	0.687	7.109	9.838	8.439	0.758	7.046	9.987
$\hat{\lambda}_1$	-1.732	0.515	-2.786	-0.737	-1.745	0.557	-2.821	-0.677
$\hat{\lambda}_2$	0.375	0.105	0.169	0.595	0.352	0.115	0.125	0.574
$\hat{\lambda}_3$	-0.022	0.006	-0.035	-0.009	-0.020	0.007	-0.033	-0.006
$\hat{\gamma}_0$	2.292	0.467	1.362	3.213	2.058	0.447	1.152	2.943
$\hat{\gamma}_1$	-2.183	0.623	-3.418	-0.905	-1.817	0.593	-3.008	-0.644
$\hat{\gamma}_2$	0.737	0.253	0.231	1.245	0.583	0.238	0.126	1.066
$\hat{\gamma}_3$	-0.103	0.040	-0.183	-0.025	-0.080	0.037	-0.153	-0.009
$\hat{\gamma}_4$	0.005	0.002	0.001	0.009	0.004	0.002	0.000	0.008
$\hat{\nu}$	—	—	—	—	7.288	3.190	2.849	15.117

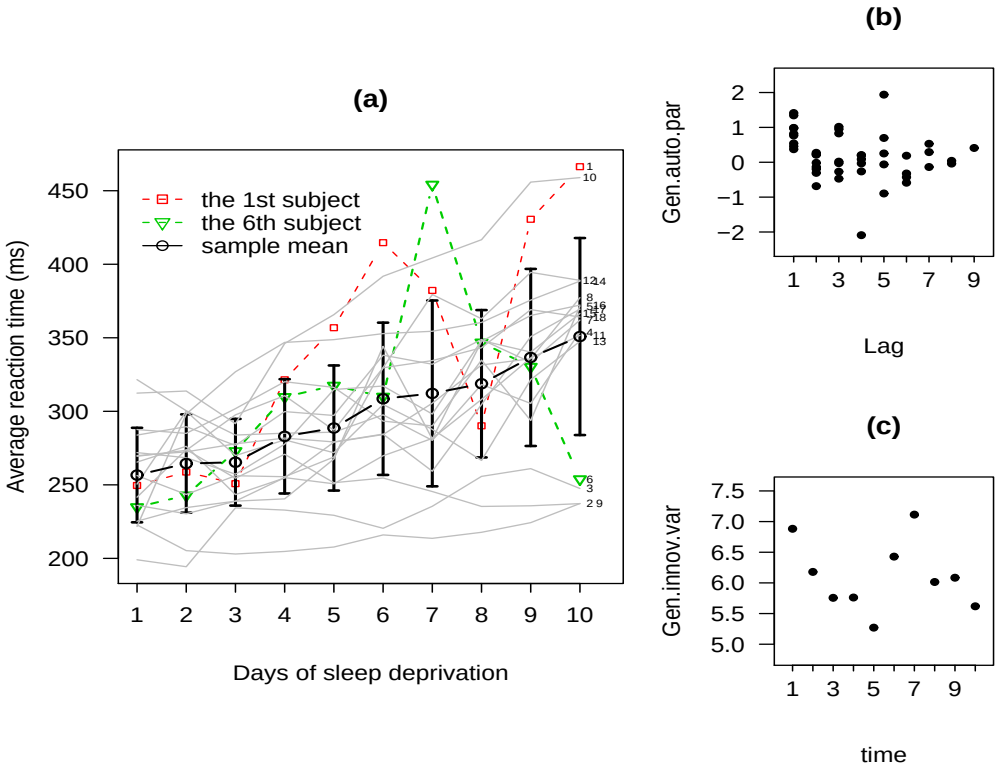

Fig. 1. (a) Trajectories of average reaction time for the 18 subjects. (b) Sample generalized autoregressive parameters. (c) Sample log-innovation variances.

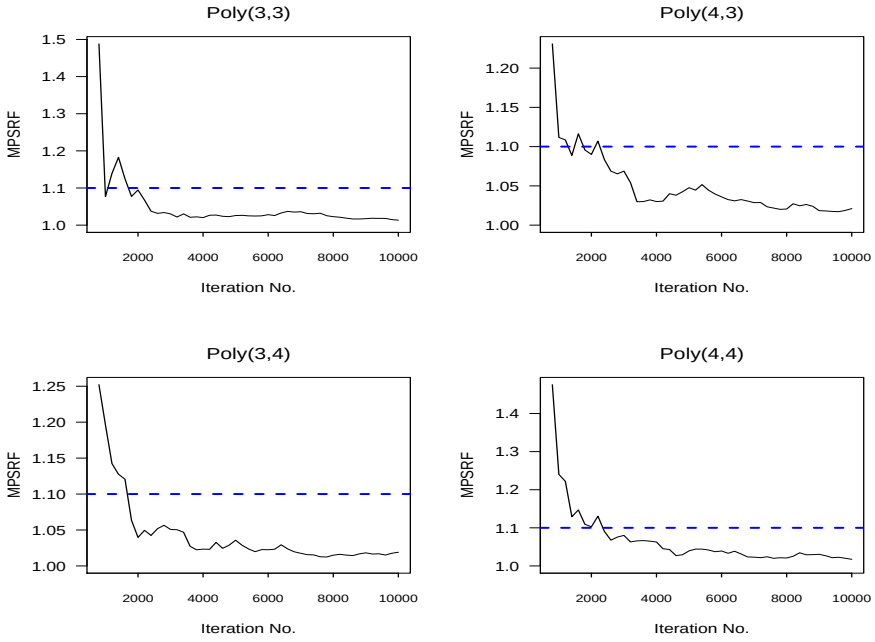

Fig. 2. MPSRF plots for four selected Poly(q, d) models.

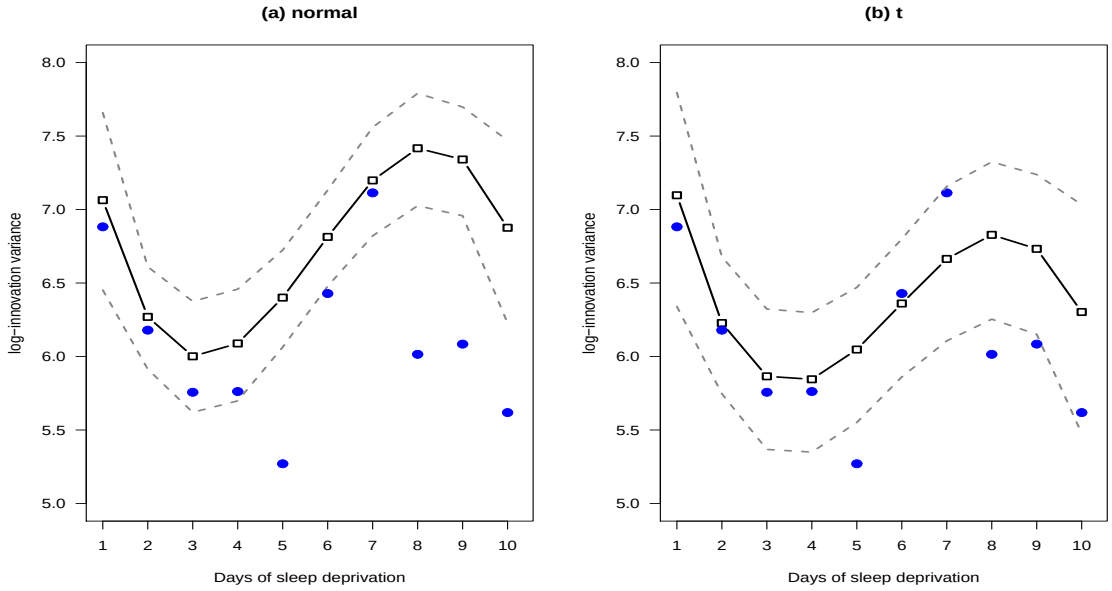

Fig. 3. Fitted log-innovation variances for the (a) normal-Poly(3,4)and (b) t-Poly(3,4) models. The solid circles represent sample log-innovation variances. The dashed and dotted lines represent the 95% credible intervals (equal-tailed probability).

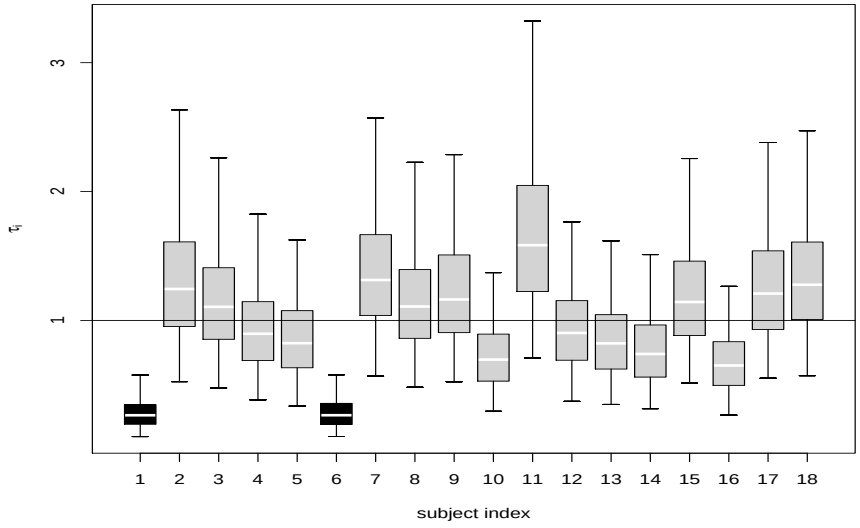

Fig. 4. Marginal posterior distributions of τ_i's for the 18 subjects. The boxplots are drawn containing 2.5%, 25%, 50%, 75%, 97.5% quantiles of the MCMC samples.