Statistica Sinica **5**(1995), 19-39

ML ESTIMATION OF THE t DISTRIBUTION USING EM AND ITS EXTENSIONS, ECM AND ECME

Chuanhai Liu and Donald B. Rubin

Harvard University

Abstract. The multivariate **t** distribution has many potential applications in applied statistics. Current computational advances will make it routinely available in practice in the near future. Here we focus on maximum likelihood estimation of the parameters of the multivariate **t**, with known and unknown degrees of freedom, with and without missing data, and with and without covariates. We describe EM, ECM and ECME algorithms and indicate their relative computational efficiencies. All three algorithms are analytically quite simple, and all have stable monotone convergence to a local maximum likelihood estimate. ECME, however, can have a dramatically faster rate of convergence.

Key words and phrases: EM, ECM, ECME, incomplete data, missing data, multivariate **t**, robust estimation.

1. Introduction

The multivariate **t** distribution can be a useful theoretical tool for applied statistics. Of particular importance, it can be used for robust estimation of means, regression coefficients, and variance-covariance matrices in multivariate linear models, even in cases with missing data. A brief history of the theoretical development leading to such uses is as follows. Dempster, Laird and Rubin (1977) show that the EM algorithm can be used to find maximum likelihood (ML) estimates (MLEs) with complete univariate data and fixed degrees of freedom, and Dempster, Laird and Rubin (1980) extend these results to the regression case. Rubin (1983) shows how this result is easily extended to the multivariate **t**, and Little and Rubin (1987) and Little (1988) further extend the results to show how EM can deal with cases with missing data. Lange, Little and Taylor (1989) consider the more general situation with unknown degrees of freedom and find the joint MLEs of all parameters using EM; they also provide several applications of this general model. Related discussion appears in many places; a recent example is Lange and Sinsheimer (1993).

Here, using a generalization of the ECM algorithm (Meng and Rubin (1993)), called the ECME algorithm (Liu and Rubin (1994)), we find the joint MLE

much more efficiently than by EM or ECM. We include comparisons of ECME with both EM and multicycle ECM and provide some new theoretical results. Care must be used with these ML procedures, however, especially with small or unknown degrees of freedom, because the likelihood function can have many spikes with very high likelihood values but little associated posterior mass under any reasonable prior. The associated parameter estimates, can therefore, be of little practical interest by themselves even though they are formally local or even global maxima of the likelihood function. It is, nevertheless, important to locate such maxima because they can critically influence the behavior of iterative simulation algorithms designed to summarize the entire posterior distribution.

The notation and theory needed to present our results are presented sequentially. First, in Section 2, we summarize fundamental results concerning the multivariate **t** distribution, and in Section 3, derive the "complete-data" likelihood equations and associated ML estimates. In Section 4, we present the EM algorithm for ML estimation with known degrees of freedom, and in Section 5 the EM and multicycle ECM algorithms when the degrees of freedom are to be estimated. Section 6 derives the efficient ECME algorithm, and Section 7 extends ECME for the **t** to the case of linear models with fully observed predictor variables. Section 8 illustrates the extra efficiency of ECME over EM and ECM in two examples, and Section 9 provides a concluding discussion.

2. Multivariate t Distribution

When we say a p-dimensional random variable Y follows the multivariate **t** distribution $\mathbf{t}_p(\mu, \Psi, \nu)$ with center μ, positive definite inner product matrix Ψ, and degrees of freedom $\nu \in (0, \infty]$, we mean that, first, given the weight τ, Y has the multivariate normal distribution, and second, that $\nu\tau$ is χ^2_ν, that is, the weight τ is Gamma distributed:

$$Y|\mu, \Psi, \nu, \tau \ \sim N_p(\mu, \Psi/\tau),$$

and (1)

$$\tau|\mu, \Psi, \nu \ \sim \text{Gamma}(\nu/2, \nu/2),$$

where the Gamma(α, β) density function is

$$\beta^\alpha \tau^{\alpha-1} \exp\{-\beta\tau\}/\Gamma(\alpha), \quad \tau > 0, \alpha > 0, \beta > 0.$$

As $\nu \to \infty$, then $\tau \to 1$ with probability 1, and Y becomes marginally $N_p(\mu, \Psi)$. Standard algebraic operations integrating τ from the joint density of (Y, τ) lead to the density function of the marginal distribution of Y, namely, $\mathbf{t}_p(\mu, \Psi, \nu)$:

$$\frac{\Gamma(\frac{\nu+p}{2})|\Psi|^{-1/2}}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})[1+\delta_Y(\mu,\Psi)/\nu]^{(\nu+p)/2}}, \tag{2}$$

where

$$\delta_Y(\mu,\Psi) = (Y-\mu)'\Psi^{-1}(Y-\mu),$$

which is the Mahalanobis distance from Y to the center μ with respect to Ψ. If $\nu > 1$, μ is the mean of Y, and if $\nu > 2$, $\Psi\nu/(\nu-2)$ is its variance-covariance matrix. Because density (2) depends on Y only through $\delta_Y(\mu,\Psi)$, the density is the same for all Y that have the same Ψ distance from μ, and thus the distribution is ellipsoidally symmetric about μ.

Further critical properties of the multivariate **t** concern its marginal and conditional distributions. Suppose Y is partitioned into $Y = (x, y)$, where the dimensions of x and y are p_x and p_y, respectively. Given τ, we have the well-known normal results:

$$x|\mu,\Psi,\nu,\tau \sim N_{p_x}(\mu_x, \Psi_x/\tau) \tag{3}$$

and

$$y|x,\mu,\Psi,\nu,\tau \sim N_{p_y}(\mu_{y|x}, \Psi^*_{y|x}/\tau), \tag{4}$$

where

$$\mu_{y|x} = \mu_y - \Psi_{y,x}\Psi_x^{-1}(x-\mu_x) \tag{5}$$

and

$$\Psi^*_{y|x} = \Psi_y - \Psi_{y,x}\Psi_x^{-1}\Psi_{x,y}, \tag{6}$$

with $(\mu_x, \mu_y) = \mu$, Ψ_x the inner product matrix corresponding to the components x of Y, and $\Psi_{y,x} = \Psi'_{x,y}$ the corresponding submatrix of Ψ corresponding to the x columns and y rows of Ψ; $\mu_{y|x}$ and $\Psi_{y|x}$ can be found either analytically as in (5) and (6) or numerically by the sweep operator (e.g., Goodnight (1979), Little and Rubin (1987)). Thus, for the marginal distribution of x we have from (1) and (3)

$$x \sim \mathbf{t}_{p_x}(\mu_x, \Psi_x, \nu).$$

From (3), given (μ,Ψ,ν,τ) the random variable $\tau\delta_x(\mu_x,\Psi_x)$ is $\chi^2_{p_x}$ distributed, that is, Gamma$(p_x/2, 1/2)$, so that treating x as data, the likelihood of τ given (μ,Ψ,ν,x) is

$$L(\tau|\mu,\Psi,\nu,x) \propto \text{Gamma}\left(\frac{p_x}{2}, \frac{\delta_x(\mu_x,\Psi_x)}{2}\right).$$

Since the Gamma distribution is the conjugate prior distribution for the parameter τ, from (1) and this likelihood, the conditional posterior distribution of τ, i.e., its distribution given (μ,Ψ,ν,x), is

$$\tau|x,\mu,\Psi,\nu = \tau|\delta_x(\mu_x,\Psi_x),\nu \sim \text{Gamma}\left(\frac{\nu+p_x}{2}, \frac{\nu+\delta_x(\mu_x,\Psi_x)}{2}\right), \tag{7}$$

whence

$$E(\tau|x,\mu,\Psi,\nu)=\frac{\nu+p_x}{\nu+\delta_x(\mu_x,\Psi_x)}. \tag{8}$$

From (4) and (7), the conditional distribution of y given x is $\mathbf{t}_{p_y}(\mu_{y|x},\Psi_{y|x},\nu+p_x)$, where

$$\Psi_{y|x}=\Psi^*_{y|x}\left[\frac{\nu+\delta_x(\mu_x,\Psi_x)}{\nu+p_x}\right].$$

3. ML Estimation of (μ,Ψ,ν) with Observed Y and τ

From the definition of the multivariate $\mathbf{t}$ distribution, n independent draws from $\mathbf{t}_p(\mu,\Psi,\nu)$ can be described as:

$$Y_i|\mu,\Psi,\tau \overset{\text{ind}}{\sim} N_p(\mu,\Psi/\tau_i) \quad \text{for } i=1,\ldots,n, \tag{9}$$

and

$$\tau_i|\nu \overset{\text{iid}}{\sim} \text{Gamma}\left(\frac{\nu}{2},\frac{\nu}{2}\right) \quad \text{for } i=1,\ldots,n. \tag{10}$$

When both $\boldsymbol{Y}=\{Y_1,\ldots,Y_n\}$ and $\boldsymbol{\tau}=\{\tau_1,\ldots,\tau_n\}$ are considered observed, $\{Y_1,\ldots,Y_n,\,\tau_1,\ldots,\tau_n\}$ comprise the complete data.

Because of the conditional structure of the complete-data model given by distributions (9) and (10), the complete-data likelihood function can be factored into the product of two distinct functions: the likelihood of (μ,Ψ) corresponding to the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{\tau}$, and the likelihood function of ν corresponding to the marginal distribution of $\boldsymbol{\tau}$. More precisely, given the complete data (Y,τ), the log-likelihood function of the parameters μ,Ψ and ν, ignoring constants, is

$$L(\mu,\Psi,\nu|Y,\boldsymbol{\tau})=L_N(\mu,\Psi|Y,\boldsymbol{\tau})+L_G(\nu|\boldsymbol{\tau}), \tag{11}$$

where

$$\begin{aligned}L_N(\mu,\Psi|Y,\boldsymbol{\tau})=&-\frac{n}{2}\ln|\Psi|-\frac{1}{2}\text{trace}\Psi^{-1}\boxed{\textstyle\sum_{i=1}^n\tau_iY_iY_i'}\\&+\mu'\Psi^{-1}\boxed{\textstyle\sum_{i=1}^n\tau_iY_i}-\frac{1}{2}\mu'\Psi^{-1}\mu\boxed{\textstyle\sum_{i=1}^n\tau_i},\end{aligned} \tag{12}$$

and

$$L_G(\nu|\boldsymbol{\tau})=-n\ln\left(\Gamma\left(\frac{\nu}{2}\right)\right)+\frac{n\nu}{2}\ln\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\boxed{\textstyle\sum_{i=1}^n(\ln(\tau_i)-\tau_i)}. \tag{13}$$

The complete-data sufficient statistics for μ, Ψ and ν, boxed in (12) and in (13), are

$$S_{\tau Y}=\sum_{i=1}^n\tau_iY_i,\;\; S_{\tau YY}=\sum_{i=1}^n\tau_iY_iY_i',\;\; S_\tau=\sum_{i=1}^n\tau_i,\;\text{and}\;S_{\tau*\tau}=\sum_{i=1}^n(\ln(\tau_i)-\tau_i). \tag{14}$$

Given the complete data $(\boldsymbol{Y}, \boldsymbol{\tau})$, $\{S_\tau, S_{\tau Y}, S_{\tau YY}\}$ is the set of complete-data sufficient statistics for (μ, Ψ), and $S_{\tau * \tau}$ is the sufficient statistic for ν.

Given the sufficient statistics S_τ, $S_{\tau Y}$, $S_{\tau YY}$, and $S_{\tau * \tau}$, the MLE of (μ, Ψ) and the MLE of ν can be obtained from $L_N(\mu, \Psi | \boldsymbol{Y}, \boldsymbol{\tau})$ and $L_G(\nu | \boldsymbol{\tau})$, respectively. Specifically, the maximum likelihood estimates of μ and Ψ from $L_N(\mu, \Psi | \boldsymbol{Y}, \boldsymbol{\tau})$ are

$$\hat{\mu} = \frac{S_{\tau Y}}{S_\tau} = \frac{\sum_{i=1}^n \tau_i Y_i}{\sum_{i=1}^n \tau_i}, \tag{15}$$

and

$$\hat{\Psi} = \frac{1}{n}\left(S_{\tau YY} - \frac{1}{S_\tau} S_{\tau Y} S'_{\tau Y}\right) = \frac{1}{n}\sum_{i=1}^n \tau_i (Y_i - \hat{\mu})(Y_i - \hat{\mu})'. \tag{16}$$

Therefore, the MLE of the center μ, namely $\hat{\mu}$, is the weighted mean of the observations $Y_1, \ldots, Y_n$ with weights $\tau_1, \ldots, \tau_n$; the MLE of the inner product matrix Ψ, namely $\hat{\Psi}$, is the average weighted sum of squares of the observations $Y_1, \ldots, Y_n$ about $\hat{\mu}$ with weights $\tau_1, \ldots, \tau_n$. The estimators of (μ, Ψ) given by (15) and (16) are known as weighted least squares estimators. With vector $\boldsymbol{\tau}$ observed, the MLE of ν can be obtained by maximizing $L_G(\nu | \boldsymbol{\tau})$ given by (13), that is, by solving

$$-\phi(\nu/2) + \ln(\nu/2) + \frac{1}{n} S_{\tau * \tau} + 1 = 0 \tag{17}$$

for ν, where $\phi(x) = d\ln(\Gamma(x))/dx$ is the digamma function; Equation (17) is discussed in the Appendix.

4. MLE of μ and Ψ with Known ν Using EM

In statistical practice with the multivariate t distribution, the vector $\boldsymbol{\tau}$ is missing and also commonly some of the values in Y are missing. In some cases, however, ν can be assumed to be known, as when statistical analyses with different specified degrees of freedom ν are used for judging the robustness of the analyses. The EM algorithm for the MLE of (μ, Ψ) is given in this section with missing $\boldsymbol{\tau}$ and known ν when the data matrix $\boldsymbol{Y}$ can have missing values that arise from an ignorable mechanism (Rubin (1976), Little and Rubin (1987)). Although the results have appeared in the literature referenced in Section 1, we derive them here because the notation and results are needed to present our new results. The E-step of the EM algorithm involves finding the conditional expectation of the complete-data sufficient statistics $\{S_\tau, S_{\tau Y}, S_{\tau YY}\}$ given the observed values of Y, the known value of ν, and the current estimate of (μ, Ψ). The M-step involves weighted least squares estimation of μ and Ψ as in (15) and (16).

Let $Y_{i,\text{obs}}$ denote the observed components of Y_i and $Y_{i,\text{mis}}$ denote the missing components of Y_i; $\boldsymbol{Y}_{\text{obs}} = \{Y_{i,\text{obs}} : i = 1, \ldots, n\}$ and $\boldsymbol{Y}_{\text{mis}} = \{Y_{i,\text{mis}} : i = 1, \ldots, n\}$. Let p_i, $\mu_{i,\text{obs}}$ and $\Psi_{i,\text{obs}}$ be the corresponding dimension, center and inner product

of $Y_{i,\text{obs}}$, respectively. Also let $\delta_{i,\text{obs}}(\mu,\Psi)=(Y_{i,\text{obs}}-\mu_{i,\text{obs}})'\Psi_{i,\text{obs}}^{-1}(Y_{i,\text{obs}}-\mu_{i,\text{obs}})$ and $\Omega^{(t)}=\{Y_{\text{obs}},\mu^{(t)},\Psi^{(t)},\nu\}$. From (8), we have

$$w_{i,\text{obs}}^{(t+1)}=E(\tau_i|\Omega^{(t)})=\frac{\nu+p_i}{\nu+\delta_{i,\text{obs}}^{(t)}}, \tag{18}$$

where $\delta_{i,\text{obs}}^{(t)}=\delta_{i,\text{obs}}(\mu^{(t)},\Psi^{(t)})$, which is the $\Psi_{i,\text{obs}}^{(t)}$ distance of $Y_{i,\text{obs}}$ from $\mu_{i,\text{obs}}$. Thus for the expectations of S_τ, $S_{\tau Y}$ and $S_{\tau YY}$ at the $(t+1)$th E-step, we have

$$S_\tau^{(t+1)}\equiv E(S_\tau|\Omega^{(t)})=\sum_{i=1}^n w_{i,\text{obs}}^{(t+1)}, \tag{19}$$

where $w_i^{(t+1)}$ is given in (18). Also from (14), (4) and (5)

$$\begin{aligned}S_{\tau Y}^{(t+1)}&\equiv E(S_{\tau Y}|\Omega^{(t)})=\sum_{i=1}^n E\left(\tau_i Y_i|\Omega^{(t)}\right)=\sum_{i=1}^n E\left[\tau_i E\left(Y_i|\Omega^{(t)},\tau_i\right)|\Omega^{(t)}\right]\\&=\sum_{i=1}^n E\left(\tau_i|\Omega^{(t)}\right)E\left(Y_i|\Omega^{(t)}\right)=\sum_{i=1}^n w_{i,\text{obs}}^{(t+1)}\hat{Y}_i^{(t)},\end{aligned} \tag{20}$$

where $w_i^{(t+1)}$ is given in (18) and $\hat{Y}_i^{(t)}\equiv E\left(Y_i|\Omega^{(t)}\right)=E\left(Y_i|\Omega^{(t)},\tau_i\right)$. The jth component of $\hat{Y}_i^{(t)}$, namely $\hat{Y}_{ij}^{(t)}$, is the conditional mean of Y_{ij} given $\Omega^{(t)}$; when Y_{ij} is observed, $\hat{Y}_{ij}^{(t)}=Y_{ij}$; otherwise, $\hat{Y}_{ij}^{(t)}$ can be found using the Sweep operator applied to $(\mu^{(t)},\Psi^{(t)})$ to predict $Y_{i,\text{mis}}$ by its linear regression from $Y_{i,\text{obs}}$. Finally, using an argument analogous to that used for (20),

$$\begin{aligned}S_{\tau YY}^{(t+1)}&\equiv E(S_{\tau YY}|\Omega^{(t)})=\sum_{i=1}^n E\left[\tau_i\left(E(Y_i|\Omega^{(t)})(E(Y_i|\Omega^{(t)})'+\text{Cov}(Y_i|\Omega^{(t)},\tau_i)\right)|\Omega^{(t)}\right]\\&=\sum_{i=1}^n w_{i,\text{obs}}^{(t+1)}\hat{Y}_i^{(t)}\left(\hat{Y}_i^{(t)}\right)'+\sum_{i=1}^n\Psi_i^{(t)},\end{aligned} \tag{21}$$

where $\Psi_i^{(t)}=E\left[\tau_i\,\text{Cov}(Y_i|\Omega^{(t)},\tau_i)|\Omega^{(t)}\right]$. The (j,k)th element of $\Psi_i^{(t)}$ is zero if either Y_{ij} or Y_{ik} is observed, and, if both Y_{ij} and Y_{ik} are missing, it is the corresponding element of

$$\Psi_{i,\text{mis}|\text{obs}}^{(t)}=\Psi_{i,\text{mis}}^{(t)}-\Psi_{i,\text{mis,obs}}^{(t)}(\Psi_{i,\text{obs}}^{(t)})^{-1}\Psi_{i,\text{obs,mis}}^{(t)}, \tag{22}$$

where $\Psi_{i,\text{mis}|\text{obs}}^{(t)}$ is found using the Sweep operator on $\Psi^{(t)}$ to predict $Y_{i,\text{mis}}$ from $Y_{i,\text{obs}}$.

From (15) and (16) we have the EM algorithm for finding the MLE of (μ,Ψ) with known ν and incomplete $\boldsymbol{Y}$ as follows.

At iteration $t+1$ with input $(\mu^{(t)},\Psi^{(t)})$,

E-step: Calculate $w_{i,\text{obs}}^{(t+1)}$ for $i=1,\ldots,n$ in (18) and the expected sufficient statistics, $S_\tau^{(t+1)}$ in (19), $S_{\tau Y}^{(t+1)}$ in (20), and $S_{\tau YY}^{(t+1)}$ in (21).

M-step: Calculate

$$\mu^{(t+1)} = \frac{S_{\tau Y}^{(t+1)}}{S_{\tau}^{(t+1)}} = \frac{\sum_{i=1}^{n} w_{i,\text{obs}}^{(t+1)} \hat{Y}_i^{(t)}}{\sum_{i=1}^{n} w_{i,\text{obs}}^{(t+1)}} \tag{23}$$

and

$$\begin{aligned} \Psi^{(t+1)} &= \frac{1}{n}\left(S_{\tau YY}^{(t+1)} - \frac{1}{S_{\tau}^{(t+1)}} S_{\tau Y}^{(t+1)} (S_{\tau Y}^{(t+1)})'\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} w_{i,\text{obs}}^{(t+1)} \left[\hat{Y}_i^{(t)} - \mu^{(t)}\right]\left[\hat{Y}_i^{(t)} - \mu^{(t)}\right]' + \frac{1}{n}\sum_{i=1}^{n} \Psi_i^{(t)}. \end{aligned} \tag{24}$$

From (23) and (24), we see that in the case where ν is fixed and $\boldsymbol{Y}$ is incomplete, the EM algorithm involves both iterative imputation and iteratively reweighted least squares estimation. The E-step of the EM algorithm effectively calculates the weights $w_i^{(t+1)}$ in (18) and imputes the missing values $\boldsymbol{Y}_{\text{mis}}$ and their cross products with their conditional expectations given $\boldsymbol{Y}_{\text{obs}}$, ν, and the current estimate of (μ, Ψ), $(\mu^{(t)}, \Psi^{(t)})$. Then the M-step of the EM algorithm does weighted least squares estimation on the imputed sufficient statistics.

5. MLE of μ and Ψ with Unknown ν Using EM or ECM

For the case where ν is unknown, Lange, Little and Taylor (1989) showed how to use EM to find the joint MLEs of all parameters μ, Ψ and ν. This extension is straightforward. First, with ν replaced by its current estimate $\nu^{(t)}$, the E-step of EM in this case is the same as that in the case of Section 4 where ν is fixed, except for the additional calculation of the conditional expectation of the sufficient statistic $S_{\tau * \tau}$ in (14) given the observed data and the current estimates, $\Omega^{(t)} = \{\mathbf{Y}_{\text{obs}}, \mu^{(t)}, \Psi^{(t)}, \nu^{(t)}\}$:

$$E\left(S_{\tau * \tau} | \Omega^{(t)}\right) = \sum_{i=1}^{n}\left[\phi\left(\frac{p_i + \nu^{(t)}}{2}\right) - \ln\left(\frac{p_i + \nu^{(t)}}{2}\right)\right] + \sum_{i=1}^{n}\left[\ln(w_{i,\text{obs}}^{(t+1)}) - w_{i,\text{obs}}^{(t+1)}\right], \tag{25}$$

where, from (8),

$$w_{i,\text{obs}}^{(t+1)} = E(\tau_i | \Omega^{(t)}) = \frac{\nu^{(t)} + p_i}{\nu^{(t)} + \delta_{i,\text{obs}}^{(t)}}, \tag{26}$$

and (25) follows from (7); as in (17), $\phi(x)$ in (25) is the digamma function.

Second, the M-step of EM in this case separately maximizes the expected L_N in (12) over (μ, Ψ) and the expected L_G in (13) over ν. Therefore, the M-step for (μ, Ψ) with unknown degrees of freedom ν is the same as that with known ν, which is given in (23) and (24). The M-step for ν is more difficult, however,

because $\nu^{(t+1)}$ must be obtained by finding the solution to the equation:

$$-\phi(\nu/2) + \ln(\nu/2) + \frac{1}{n}\sum_{i=1}^{n}\left[\ln(w_{i,\text{obs}}^{(t+1)}) - w_{i,\text{obs}}^{(t+1)}\right] + 1 + \frac{1}{n}\sum_{i=1}^{n}\left[\phi\left(\frac{\nu^{(t)}+p_i}{2}\right) - \ln\left(\frac{\nu^{(t)}+p_i}{2}\right)\right] = 0. \tag{27}$$

The left hand side of Equation (27), excluding the last term, equals the left hand side of the corresponding Equation (17) with the missing value of τ_i replaced by its conditional expectation for $i = 1, \ldots, n$ given $\Omega^{(t)}$; thus, the last term on the left hand side of Equation (27) can be interpreted as a correction for the mean value imputation of the missing weights $\tau_1, \ldots, \tau_n$. As a matter of fact, the solution to Equation (27) is always greater than or equal to the solution to Equation (17) with τ_i replaced by its conditional expectation $w_i^{(t)}$ for $i = 1, \ldots, n$ because the last term on the left side of Equation (27) is non-positive and $-\phi(\nu/2)+\ln(\nu/2)$ is strictly decreasing in $(0, +\infty)$. A one-dimensional search, such as the half-interval method (Carnahan, Luther and Wilkes (1969)), can be used to solve (27) for ν.

When ν is unknown, the convergence of the EM algorithm is very slow (see, for example, Liu and Rubin (1994)), and the one dimensional search for updating ν is time consuming as discussed and illustrated in Lange, Little and Taylor (1989). Consequently, we consider extensions of EM that can be more efficient, for example, the multicycle ECM algorithm of Meng and Rubin (1993).

The ECM algorithm (Meng and Rubin (1993)) generalizes the EM algorithm by replacing the M-step with a sequence of simple constrained (or conditional) maximization steps, abbreviated CM-steps, indexed by $s = 1, \ldots, S$, each of which fixes some function of the parameters θ to be maximized. Meng and Rubin (1993) show that the ECM algorithm is a GEM algorithm (Generalized EM (Dempster, Laird and Rubin (1977))) and shares the nice convergence properties of EM. The rate of convergence of the ECM algorithm is given in Meng (1994), and it is typically slower than EM, at least in terms of number of iterations. A multi-cycle version of ECM (MCECM (Meng and Rubin (1993))) is obtained by inserting an E-step before *each* CM-step rather than just before the set of CM-steps.

For the multivariate **t**, let the parameters $\theta = (\mu, \Psi, \nu)$ be partitioned into $\theta_1 = (\mu, \Psi)$ and $\theta_2 = \nu$. In this case, ECM is EM because the complete-data likelihood function of $\theta = (\mu, \Psi, \nu)$ factorizes into a factor for $\theta_1 = (\mu, \Psi)$ and a factor for $\theta_2 = \nu$. The MCECM algorithm that performs an E-step before each CM-step is as follows.

E-step of MCECM at iteration $t + 1$: The same as the E-step of EM, just conditioning on the current parameter estimates, $\theta^{(t)} = (\mu^{(t)}, \Psi^{(t)}, \nu^{(t)})$.

CM-step 1 of MCECM at iteration $t+1$ **:** Fix $\nu = \nu^{(t)}$, and calculate $\mu^{(t+1)}$ and $\Psi^{(t+1)}$ using (23) and (24) with ν replaced by $\nu^{(t)}$.

E-step of MCECM at iteration $t+1$**:** The same as the E-step of EM, just conditioning on the current parameter estimates, $\theta^{(t+1/2)} = (\mu^{(t+1)}, \Psi^{(t+1)}, \nu^{(t)})$.

CM-step 2 of MCECM at iteration $t+1$ **:** Fix $\mu = \mu^{(t+1)}$ and $\Psi = \Psi^{(t+1)}$, and calculate $\nu^{(t+1)}$, which is the solution to Equation (27) with $w_{i,\text{obs}}^{(t+1)}$ defined as the conditional expectation of τ_i given $\theta^{(t+1/2)}$ and Y_{obs} rather than given $\theta^{(t)}$ and Y_{obs} as in (26); that is, in contrast to (26) for EM=ECM, we have for MCECM

$$w_{i,\text{obs}}^{(t+1)} = \frac{\nu^{(t)} + p_i}{\nu^{(t)} + \delta_{i,\text{obs}}^{(t+1)}}, \tag{28}$$

for $i = 1, \ldots, n$, where $\delta_{i,\text{obs}}^{(t+1)} = \delta_{i,\text{obs}}(\mu^{(t+1)}, \Psi^{(t+1)})$.

6. MLE of the t Distribution via an Efficient Algorithm: ECME

The ECME algorithm (Liu and Rubin (1994)) extends the ECM algorithm by allowing CM-steps to maximize *either* the constrained expected log-likelihood, as with ECM, or the correspondingly constrained actual log-likelihood function $L(\theta)$. Although the ECME algorithm is not a GEM algorithm, it shares the nice convergence properties of EM (Wu (1983)) and ECM (Meng (1994)) under mild conditions. Moreover, it typically converges more quickly than either EM or ECM (Liu and Rubin (1994)).

Our application of ECME maximizes the expected log-likelihood over $\theta_1 = (\mu, \Psi)$ given ν but maximizes the actual log-likelihood over $\theta_2 = \nu$ given $\theta_1 = (\mu, \Psi)$, and yields a much faster converging algorithm because of the lost-memory of the $\boldsymbol{\tau}$ in the second CM step. From Section 2, the loglikelihood function of $\theta = (\mu, \Psi, \nu)$ given the observed data Y_{obs}, ignoring constants, is

$$\begin{aligned} L(\mu, \Psi, \nu | Y_{\text{obs}}) = & \sum_{i=1}^{n} \ln\left(\Gamma\left(\frac{\nu + p_i}{2}\right)\right) - n \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{1}{2}\sum_{i=1}^{n} \ln|\Psi_{i,\text{obs}}| \\ & + \frac{n\nu}{2}\ln(\nu) - \sum_{i=1}^{n} \frac{\nu + p_i}{2} \ln\left(\nu + \delta_{i,\text{obs}}\right). \end{aligned} \tag{29}$$

Given $\theta_1 = (\mu, \Psi)$, we can maximize (29) over $\theta_2 = \nu$ by finding the solution to the following equation:

$$\begin{aligned} -\phi(\nu/2) + \ln(\nu/2) + \frac{1}{n}\sum_{i=1}^{n} [\ln(w_{i,\text{obs}}) - w_{i,\text{obs}}] + 1 \\ + \frac{1}{n}\sum_{i=1}^{n} \left[\phi\left(\frac{\nu + p_i}{2}\right) - \ln\left(\frac{\nu + p_i}{2}\right)\right] = 0, \end{aligned} \tag{30}$$

where $w_i = (\nu + p_i)/(\nu + \delta_{i,\text{obs}})$.

Thus, the E-step of ECME is the same as the E-step of EM and ECM; CM-step 1 of ECME is the same as CM-step 1 of ECM; but CM-step 2 of ECME maximizes the actual likelihood (29) over ν given $\theta_1 = (\mu^{(t+1)}, \Psi^{(t+1)})$. Code for this maximization involves a one-dimensional search as with EM.

To sharpen the comparison among ECME, EM and ECM for finding the MLE of the t distribution, assume that (μ, Ψ) is known. ECME finds the MLE of ν directly using a one dimensional search over the actual likelihood. In contrast, EM, ECM and MCECM find the MLE of ν iteratively based on (27), where each iteration itself requires a one dimensional search over the expected log-likelihood of ν. Therefore, the ECME algorithm should converge substantially faster than EM, ECM or MCECM. A numerical example comparing the convergence rates of EM, ECM, MCECM and ECME is presented in Section 8.

7. Extension to Linear Models with Fully Observed Predictor Variables

The previous results can be easily extended to linear models with q fully observed predictor variables, $\mathbf{X}$, where the residuals of Y_i given X_i are independently and identically multivariate $\mathbf{t}$ distributed. The required work is a straightforward modification of the previous development where now, instead of a common mean μ for all $i = 1, \ldots, n$, we have $\mu_i = \beta' X_i$ for $i = 1, \ldots, n$, where β is $(q \times p)$ matrix of regression coefficients. MLEs of the regression coefficients, the inner product matrix of the errors, and the degrees of freedom are now described as simple extensions of the previous results.

A multivariate linear model can be represented as follows:

$$\mathbf{Y}_{(n\times p)} = \mathbf{X}_{(n\times q)}\beta_{(q\times p)} + \mathbf{e}_{(n\times p)}, \tag{31}$$

where the predictor matrix $\mathbf{X}$ consists of n fully observed q-dimensional vectors, the response data matrix $\mathbf{Y}$ contains the corresponding n observations of p-dimensional outcome variables, the coefficient matrix β is $(q \times p)$ dimensional, and the error term $\mathbf{e}$ consists of the corresponding n p-dimensional errors. Let Y_i', X_i' and e_i' be the ith rows of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{e}$, respectively, so that model (31) can be written as

$$Y_i = \mu_i + e_i = \beta' X_i + e_i \tag{32}$$

for $i = 1, \ldots, n$, where, as in previous sections, the error terms $e_1, \ldots, e_n$ are independently and identically $\mathbf{t}_p(0, \Psi, \nu)$ distributed. In analogy with Section 3, we have the following complete-data model. The outcome vector observations Y_i are conditionally independent and normal given the parameters and the weight vector $\boldsymbol{\tau}$:

$$Y_i|(\mathbf{X}, \beta, \Psi, \boldsymbol{\tau}) \overset{\text{ind}}{\sim} N_p(\beta' X_i, \Psi/\tau_i) \quad \text{for} \quad i = 1, \ldots, n, \tag{33}$$

and the marginal distribution of weights is Gamma,

$$\tau_i|\nu \overset{\text{iid}}{\sim} \text{Gamma}(\nu/2, \nu/2) \quad \text{for} \quad i = 1, \ldots, n. \tag{34}$$

Following the development in Section 3, we now have a set of complete-data sufficient statistics for (β, Ψ, ν) that generalize S_τ to $S_{\tau XX}$ and $S_{\tau Y}$ to $S_{\tau XY}$. The set of complete-data sufficient statistics for (β, Ψ, ν) is:

$$S_{\tau YY} = \mathbf{Y}'W\mathbf{Y} = \sum_{i=1}^{n} \tau_i Y_i Y_i', \quad S_{\tau XY} = \mathbf{X}'W\mathbf{Y} = \sum_{i=1}^{n} \tau_i X_i Y_i',$$

$$S_{\tau XX} = \mathbf{X}'W\mathbf{X} = \sum_{i=1}^{n} \tau_i X_i X_i', \tag{35}$$

and $S_{\tau^*\tau}$ is given by (14), where $W = \text{diag}\{\tau_1, \ldots, \tau_n\}$. The MLE of ν given the complete-data $\{\mathbf{Y}, \mathbf{X}, \boldsymbol{\tau}\}$ is just as given in Section 3. The complete-data MLE of (β, Ψ) is easy to derive and is given by

$$\hat{\beta} = S_{\tau XX}^{-1} S_{\tau XY} = (\mathbf{X}'W\mathbf{X})^{-1}\mathbf{X}'W\mathbf{Y} \tag{36}$$

and

$$\hat{\Psi} = \frac{1}{n}\left[S_{\tau YY} - S_{\tau XY}' S_{\tau XX}^{-1} S_{\tau XY}\right] = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\beta})'W(\mathbf{Y} - \mathbf{X}\hat{\beta}); \tag{37}$$

$\hat{\beta}$ in (36) is the weighted least squares estimator of β, and $\hat{\Psi}$ in (37) is the associated estimator of residual variance.

When the weights $\tau_1, \ldots, \tau_n$ are unknown, and $\mathbf{Y}$ is either complete or incomplete, we can use EM, ECM, or ECME to find the MLEs of the parameters. The conditional expectations of the sufficient statistics needed for the E-step can be obtained case by case, i.e., for $i = 1, \ldots, n$, where, with the linear model, $\mu_i = \beta' X_i$ plays the role of μ in Sections 2 — 6. Therefore, the conditional expectation calculations for the linear model are essentially the same as those in Sections 2 — 6. For example,

$$\hat{W}_{i,i} \equiv E(\tau_i|\mathbf{Y}_{\text{obs}}, \beta, \Psi, \nu) = \frac{\nu + p_i}{\nu + \delta_{i,\text{obs}}}, \tag{38}$$

where, as before, $\delta_{i,\text{obs}} = (Y_{i,\text{obs}} - \mu_{i,\text{obs}})'\Psi_{i,\text{obs}}^{-1}(Y_{i,\text{obs}} - \mu_{i,\text{obs}})$ and p_i is the dimension of $Y_{i,\text{obs}}$ and $\mu_{i,\text{obs}}$, but now $\mu_{i,\text{obs}}$ refers to the components of $\mu_i = \beta' X_i$ corresponding to the components of $Y_{i,\text{obs}}$, i.e., the components of Y_i that are observed, rather than simply the components of the common μ.

Because it is so much more efficient, as will be seen in Section 8, we only give the ECME algorithm for finding the MLE of the parameters (β, Ψ, ν). As in

Section 6, we use the partition $\theta_1 = (\beta, \Psi)$ and $\theta_2 = \nu$, and maximize the actual likelihood over ν given (β, Ψ).

At iteration $t+1$ with input $(\beta^{(t)}, \Psi^{(t)})$ and $\nu^{(t)}$,

E-step: Calculate

$$\hat{W}_{i,i}^{(t+1)} \equiv E(\tau_i|\mathbf{Y}_{\rm obs}, \beta^{(t)}, \Psi^{(t)}, \nu^{(t)}) = \frac{\nu^{(t)} + p_i}{\nu^{(t)} + \delta_{i,\rm obs}^{(t)}},$$
$$\hat{Y}_i^{(t)} \equiv E(Y_i|\mathbf{Y}_{\rm obs}, \beta^{(t)}, \Psi^{(t)}, \nu^{(t)}),$$

and $\Psi_i^{(t)}$ in (22), where $\delta_{i,\rm obs}^{(t)} = (Y_{i,\rm obs} - \mu_{i,\rm obs})'(\Psi_{i,\rm obs}^{(t)})^{-1}(Y_{i,\rm obs} - \mu_{i,\rm obs})$, and thereby calculate the sufficient statistics:

$$S_{\tau XX}^{(t+1)} = \sum_{i=1}^{n} \hat{W}_{i,i} X_i X_i', \quad S_{\tau XY}^{(t+1)} = \sum_{i=1}^{n} \hat{W}_{i,i} X_i \left(\hat{Y}_i^{(t)}\right)',$$
$$\text{and} \quad S_{\tau YY}^{(t+1)} = \sum_{i=1}^{n} \hat{W}_{i,i} \hat{Y}_i^{(t)} \left(\hat{Y}_i^{(t)}\right)' + \sum_{i=1}^{n} \Psi_i^{(t)}.$$

CM-step 1: Calculate

$$\hat{\beta}^{(t+1)} = (S_{\tau XX}^{(t+1)})^{-1} S_{\tau XY}^{(t+1)} = \left(\mathbf{X}'\hat{W}^{(t+1)}\mathbf{X}\right)^{-1} \mathbf{X}' W \hat{\mathbf{Y}}^{(t)}$$

and

$$\hat{\Psi}^{(t+1)} = \frac{1}{n}\left(S_{\tau YY}^{(t+1)} - (S_{\tau XY}^{(t+1)})'(S_{\tau XX}^{(t+1)})^{-1} S_{\tau XY}^{(t+1)}\right)$$
$$= \frac{1}{n}(\hat{\mathbf{Y}}^{(t+1)} - \mathbf{X}\hat{\beta}^{(t+1)})'\hat{W}^{(t+1)}(\hat{\mathbf{Y}}^{(t+1)} - \mathbf{X}\hat{\beta}^{(t+1)}) + \frac{1}{n}\sum_{i=1}^{n} \Psi_i^{(t)}.$$

CM-step 2: Use a one-dimensional search, such as that in Section 3, to find $\hat{\nu}^{(t+1)}$, the solution in ν to Equation (30) with δ_i replaced by $\delta_{i,\rm obs}(\mu_i^{(t+1)}, \Psi_{i,\rm obs}^{(t+1)})$, which is the $\Psi_{i,\rm obs}^{(t+1)}$ distance of $Y_{i,\rm obs}$ from $\mu_{i,\rm obs}^{(t+1)}$, where $\mu_i^{(t+1)} = \left(\beta^{(t+1)}\right)' X_i$.

8. Numerical Examples

To illustrate the use of the EM, MCECM and ECME algorithms, we first present an artificial example that demonstrates their essential features and then present a real data example.

An Artificial Example: Table 1 is an artificial bivariate data set, where the symbols "?" denote missing values, created by appending four extreme values to the artificial dataset used in Murray (1977). We suppose that the data follow the bivariate **t** and the missing-data mechanism is ignorable. Using a variety of

starting values $\mu^{(0)}$, $\Psi^{(0)}$ and $\nu^{(0)}$ for the iterative algorithms, we found three stationary points of the likelihood, which are displayed in Table 2. Table 2 also gives three stationary points under the multivariate normal model.

Table 1. A bivariate data set with missing values

x_1	-1	-1	1	1	-2	-2	2	2	?	?	?	?	-12	12	?	?
x_2	-1	1	-1	1	?	?	?	?	-2	-2	2	2	?	?	-12	12

Table 2. Stationary points of the parameters in the artificial example

Model	μ	$(\Psi_{1,1},$	$\Psi_{2,2},$	$\Psi_{2,1})$	ν	Loglikelihood
	$(0,0)$	(5.6565,	5.6565,	4.3883)	1.4030	-45.5543
t	$(0,0)$	(5.6565,	5.6565,	-4.3883)	1.4030	-45.5543
	$(0,0)$	(4.2347,	4.2347,	0.0000)	1.2527	-45.7713
	$(0,0)$	(38.0000,	38.0000,	36.9865)	—	-47.4185
Normal	$(0,0)$	(38.0000,	38.0000,	-36.9865)	—	-47.4185
	$(0,0)$	(38.8000,	38.8000,	0.0000)	—	-51.2066

To compare the convergence rates of the algorithms described in this article, we started all three, that is, EM = ECM, multicycle ECM, and ECME = multicycle ECME, from the same value. Specifically, we let $\mu^{(0)} = (0,0)$, $\Psi_{1,1}^{(0)} = \Psi_{2,2}^{(0)} = 38.0000$, $\Psi_{1,2}^{(0)} = 36.9865$, that is, the maximum likelihood estimate of μ and Ψ with ν fixed at ∞, and let $\nu^{(0)} = 1000.00$. For judging convergence, we set the tolerance parameter for each component of $\theta^{(t+1)} - \theta^{(t)}$ to 1.0E-05. The sequences of $\nu^{(t)}$ $(t = 1, 2, \ldots)$ and the corresponding loglikelihoods $L(\mu^{(t)}, \Psi^{(t)}, \nu^{(t)}|Y_{\text{obs}})$ using EM, multicycle ECM and ECME are displayed in Figure 1, and these demonstrate a significantly faster convergence rate for the ECME algorithm in this artificial example. With respect to the actual computational times of the different methods, Table 3 shows that ECME has at least a seven-fold advantage in this case over EM and multicycle ECM. We expect such advantages to be typical, although not guaranteed, because of the freedom to maximize over ν the constrained actual likelihood rather than the expected complete-data loglikelihood, which has the memory of the previous value of the parameter ν through the expected complete-data sufficient statistics; Liu and Rubin (1994) provide explicit large sample results for the one-dimensional t-distribution.

An analogy with the Gibbs sampler may help to clarify the source of the expected gains. Consider the simple case with missing random variable Y_{mis} and

random parameter $\theta = (\theta_1, \theta_2)$. ECM employs an E-step using the distribution $(Y_{\rm mis}|\theta_1, \theta_2)$, a first CM-step using the distribution $(\theta_1|Y_{\rm mis}, \theta_2)$, and a second CM-step using the distribution $(\theta_2|Y_{\rm mis}, \theta_1)$; the analogous Gibbs sampler takes draws from these three conditional distributions. In contrast, ECME uses the same E-step and the same first CM-step, but uses the distribution $(\theta_2|\theta_1)$ for the second CM-step, where the E-step and second CM-step distributions are factors of the joint distribution $(Y_{\rm mis}, \theta_2|\theta_1)$; the analogous Gibbs sampler takes draws from these three distributions, which is equivalent to the two-step Gibbs sampler drawing $(Y_{\rm mis}, \theta_2|\theta_1)$ in one step and $(\theta_1|Y_{\rm mis}, \theta_2)$ in the other. Liu, Wong and Kong (1994) and Liu (1994) give results supporting the expected more rapid convergence with the two-step Gibbs sampler.

ln(Iteration) ln(Iteration)

Figure 1. The sequences of $\nu^{(t)}$ $(t = 1, 2, \ldots)$ in (a) and the corresponding actual log-likelihood in (b) found by multicycle ECM (solid line), EM (dotted line that is indistinguishable from the solid line) and ECME (dashed line) in the artificial example. Note the use of the logarithmic scale for the number of iterations.

Table 3. CPU times of different algorithms until convergence for the artificial example

Algorithm	CPU time per iteration (sec)	# of iterations	Total CPU time (min:sec)
EM = ECM	0.26	7,061	30:27
Multicycle ECM	0.26	7,053	30:11
ECME	1.92	130	4:10

Table 4. Data for the Creatinine Clearance Example (Shih and Weisberg (1986))

No.	WT	SC	Age	CR
1	71.0	0.71253	38	132.0
2	69.0	1.48161	78	53.0
3	85.0	2.20545	69	50.0
4	100.0	1.42505	70	82.0
5	59.0	0.67860	45	110.0
6	73.0	0.75777	65	100.0
7	63.0	1.11969	76	68.0
8	81.0	0.91611	61	92.0
9	74.0	1.54947	68	60.0
10	87.0	0.93873	64	94.0
11	79.0	0.99528	66	105.0
12	93.0	1.07445	49	98.0
13	60.0	0.70122	43	112.0
14	70.0	0.71253	42	125.0
15	83.0	0.99528	66	108.0
16	70.0	2.52212	78	30.0
17	73.0	1.13100	35	111.0
18	85.0	1.11969	34	130.0
19	68.0	1.37982	35	94.0
20	65.0	1.11969	16	130.0
21	53.0	0.97266	54	59.0
22	50.0	1.60602	73	38.0
23	74.0	1.58339	66	65.0
24	67.0	1.40244	31	85.0
25	80.0	0.67860	32	140.0
26	67.0	1.19886	21	80.0
27	68.0	7.60001	81	4.3
28	72.2	2.10001	43	43.2
29	NA	1.35719	78	75.0
30	NA	1.05183	38	41.0
31	107.0	NA	62	120.0
32	75.0	NA	70	52.0
33	62.0	NA	63	73.0
34	52.0	NA	68	57.0

The table reports the results of a clinical trial on Endogenous creatinine clearance of 34 male patients, conducted overseas by Merck Sharp and Dohme Research Laboratories. WT = body weight, SC = *serum* creatinine concentration, CR = Endogenous creatinine clearance, NA = not available.

Table 5. Mahalanobis distances and weights evaluated at MLE in the clinical example under the t model where the p-values are calculated based on the approximation $\delta_{i,\text{obs}}/p_i \sim F_{p_i,\nu}$.

Case i	$\hat{\delta}_{i,\text{obs}}$	p-value	$\hat{w}_i$
1	2.1531	0.2863	1.3995
2	2.5139	0.3413	1.2982
3	3.9687	0.5258	1.0049
4	3.6915	0.4953	1.0501
5	3.1494	0.4294	1.1514
6	2.8120	0.3841	1.2249
7	3.9468	0.5235	1.0083
8	2.0130	0.2641	1.4433
9	0.9163	0.0870	1.9108
10	2.8648	0.3914	1.2128
11	1.9731	0.2577	1.4562
12	4.0596	0.5354	0.9909
13	2.7458	0.3748	1.2405
14	1.9759	0.2582	1.4553
15	2.0294	0.2667	1.4380
16	6.0978	0.7006	0.7551
17	2.7311	0.3727	1.2440
18	3.8146	0.5091	1.0295
19	4.9598	0.6190	0.8708
20	9.2908	0.8375	0.5501
21	4.2477	0.5545	0.9631
22	6.7422	0.7373	0.7023
23	1.4412	0.1707	1.6543
24	4.3536	0.5648	0.9482
25	4.8966	0.6137	0.8783
26	5.5908	0.6671	0.8026
27	86.3933	0.9993	0.0728
28	8.5917	0.8157	0.5849
29	3.2238	0.5767	0.9607
30	15.1689	0.9605	0.3101
31	4.9377	0.7305	0.7384
32	1.4378	0.2924	1.3998
33	1.2040	0.2430	1.4889
34	4.6031	0.7068	0.7733

A Clinical Trial Example: The data in Table 4, reproduced from Table 1 in Shih and Weisberg (1986), are from a clinical trial on 34 male patients with 3 covariates, body weight (WT) in kg, serum creatinine (SC) concentration in mg/deciliter, and age in years, and one outcome variable, endogenous creatinine (CR) clearance. Of the 34 male patients, two had no recorded WT, and four were missing SC; the missingness is assumed to be ignorable as in Shih and Weisberg (1986). A typical model recommended in many pharmacokinetics textbooks for modelling CR as a function of WT, SC and Age (see Shih and Weisberg (1986)) is of the form

$$E(\ln(\mathrm{RC})) = \beta_0 + \beta_{\mathrm{WT}} \ln(\mathrm{WT}) + \beta_{\mathrm{SC}} \ln(\mathrm{SC}) + \beta_{\mathrm{Age}} \ln(140 - \mathrm{Age}). \tag{39}$$

Shih and Weisberg (1986) use this data to illustrate their method for assessing influence in multiple linear regression using the multivariate normal distribution with incomplete data. Assuming that the response ln(RC) and the three predictors ln(WT), ln(SC), and ln(140-Age) are jointly multivariate normal, they found that one observation, case 27, is influential, and another case, case 30 (with WT missing), has relatively large influence on the regression coefficient $\beta = (\beta_0, \beta_{\mathrm{WT}}, \beta_{\mathrm{SC}}, \beta_{\mathrm{Age}})$ but is not "significant".

Here, we apply the multivariate **t** distribution to this data set. Starting from various places, there appears to be only one mode, with $\hat{\nu} = 6.51$. The likelihood ratio statistic G^2 for the **t** relative to the normal is 10.0694, which suggests that the **t** distribution is more appropriate for these data than the normal distribution.

Table 5 shows the corresponding estimated Mahalanobis distances of all cases with $\theta = \hat{\theta}$, the MLE. Because, given θ, $\delta_{i,\mathrm{obs}}/p_i \sim F_{p_i,\nu}$, where $F_{p_i,\nu}$ denotes the F-distribution with degrees of freedom (p_i, ν), we have approximately

$$P(\max_{1 \le i \le 34} \delta_{i,\mathrm{obs}} > 86.3933) = 0.0235,$$

which is the chance that the largest Mahalanobis distance among 34 cases independently drawn from $\mathbf{t}_4(\hat{\mu}, \hat{\Psi}, \hat{\nu})$ is larger than $\hat{\delta}_{27,\mathrm{obs}} = 86.3933$.

Table 6 gives the MLEs of the regression coefficients in Equation (39) based on all patients and based on all patients except patient 27 under the **t** distribution and normal distribution, respectively. The estimate of the treatment effect, β_{SC}, is less sensitive to the deletion of patient 27 under the **t** model than under the normal model.

Table 6. The MLEs of the regression coefficients for the model of the clinical trial example

Model	All patients				
	ν	β_0	β_{WT}	β_{SC}	β_{Age}
$\mathbf{t}$	6.51	−2.96	1.02	−0.82	0.70
Normal	(∞)	−3.34	1.17	−1.08	0.65
	Patient 27 deleted				
	ν	β_0	β_{WT}	β_{SC}	β_{Age}
$\mathbf{t}$	∞	−3.25	1.07	−0.77	0.71
Normal	(∞)	−3.25	1.07	−0.77	0.71

9. Discussion

The EM, ECM and ECME algorithms can all be used to find the MLEs of the parameters of the multivariate $\mathbf{t}$ distribution. The associated large sample variances of the MLEs based on second derivatives of the log-likelihood at the MLEs can be obtained, for example, using SEM (Meng and Rubin (1991)) and its extensions SECM (van Dyk, Meng and Rubin (1995)), or using the explicit formulas given in Lange, Little and Taylor (1989). Due to the fact that the likelihood of the multivariate $\mathbf{t}$ distribution can have many modes, the full posterior distribution is far more reliable inferentially than the MLEs and associated variances, especially with small samples. However, the MLEs and associated variances of the parameters of the $\mathbf{t}$ distribution can still be very useful. First, they can provide sensible values of the degrees of freedom that can be used, for example, in sensitivity analyses of models. Second, and perhaps more important in practice, they can be used to create an approximate starting distribution for multiple sequences of the Gibbs sampler (Gelman and Rubin (1992)), which in turn can track the entire posterior distribution.

The Gibbs sampler can be viewed as a stochastic version of EM type algorithms (EM, ECM, ECME), where the parameters are considered random variables, and different versions of the Gibbs sampler exist corresponding to the EM, ECM and ECME algorithms. When data contain missing values, an especially efficient method is provided in Liu (1995), which extends the monotone data augmentation of Rubin and Schafer (1990) for multivariate normal data to multivariate $\mathbf{t}$ data and generalizes the results of Liu (1993) to implement efficiently monotone data augmentation for multivariate $\mathbf{t}$ data.

Acknowledgements

We gratefully acknowledge NSF SES-9207456 and the U.S. Census Bureau for partially supporting this work, and thank the editors and referees for many helpful comments.

Appendix

Proposition 1 shows that Equation (17) has a unique solution in the interval $(0, +\infty]$. To find this solution, one can use a one-dimensional search such as the half-interval method (Carnahan, Luther and Wilkes (1969)). To calculate the Digamma function $\phi(\cdot)$, we use

$$\phi(x+1) = \phi(x) + \frac{1}{x}, \quad x > 0, \tag{40}$$

(see (6.3.5) of Davis (1965)) and the integral of

$$\phi(x) = -\gamma + \int_0^1 \frac{1 - y^{x-1}}{1-y} \mathrm{d}y \tag{41}$$

for $2 \leq x < 3$ (Davis (1965), eq. 6.3.22), where γ is Euler's constant. Although (41) is true for all $x > 0$, we use (41) with $2 \leq x < 3$ because $(1 - y^{x-1})/(1-y)$ is quite smooth (in fact, almost linear) as a function of y in the interval $(0, 1)$ given $x \in [2, 3)$.

Proposition 1. *Let*

$$u(x) = -\phi(x) + \ln(x) + S_{\tau * \tau}/n + 1, \tag{42}$$

where $S_{\tau*\tau}$ is given in (14)*. Then $u(x)$ has the following properties:*

(1) *The function $u(x)$ is concave over $(0, +\infty)$,* i.e.*, $u''(x) < 0$ for all $x \in (0, +\infty)$.*

(2) *$-\phi(x) + \ln(x)$, is strictly decreasing over $(0, +\infty)$, with limits*

$$\lim_{x \to 0+} (-\phi(x) + \ln(x)) = +\infty \quad \textit{and} \quad \lim_{x \to +\infty} (-\phi(x) + \ln(x)) = 0.$$

(3) *$S_{\tau*\tau}/n + 1 \leq 0$, and $S_{\tau*\tau}/n + 1 = 0$ iff $\nu = +\infty$. When $S_{\tau*\tau}/n + 1 < 0$, the maximization of $u(x)$ has a unique solution that is determined by $u'(x) = 0$. When $S_{\tau*\tau}/n + 1 = 0$, $u(x)$ is strictly increasing in $(0, +\infty)$ and $\sup_{x \in (0,+\infty)} u(x) = u(+\infty)$.*

Proof. From Equation (42), conclusion (1) will be proved by showing $\phi'(x) - (1/x) > 0$ for all x in $(0, +\infty)$. From (6.4.12) of Davis (1965), i.e., as $x \to +\infty$,

$$\phi'(x) \approx \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} - \frac{1}{30x^5} + \frac{1}{42x^7} - \frac{1}{30x^9} + \cdots,$$

we have

$$\lim_{x\to+\infty} x^2\left(\phi'(x)-\frac{1}{x}\right)=\frac{1}{2}.$$

Thus, there exists a positive constant $N_0>1$ such that $\phi'(x)-(1/x)>0$ for all $x\geq N_0$. Suppose that this inequality holds for all $x\geq m>1$; then from (40), for all x in $[m-1,m)$, $\phi'(x)-(1/x)=\phi'(x+1)+(1/x^2)-(1/x)>\phi'(x+1)-1/(x+1)>0$. By induction, we have $\phi'(x)-(1/x)>0$ for all $x\in(0,+\infty)$. Thus, (1) is proved and the first part of (2) follows from (1).

From (40) we have

$$\lim_{x\to+0}\left(-\phi(x)+\ln(x)\right)=\lim_{x\to+0}\left(-\phi(x+1)+\frac{1}{x}+\ln(x)\right)=+\infty.$$

From (6.3.18) of Davis (1965) we have

$$\lim_{x\to+\infty}\left(-\phi(x)+\ln(x)\right)=\lim_{x\to+\infty}\left(\frac{1}{2x}+\frac{1}{12x^2}-\frac{1}{120x^4}+\frac{1}{252x^6}-\cdots\right)=0.$$

Thus (2) is proved. The first part of (3) is an immediate result of the inequality $\ln(\tau_i)-\tau_i+1\leq 0$ for all $\tau_i>0$ and $\ln(\tau_i)-\tau_i+1=0$ iff $\tau_i=1$. The second part of (3) follows from (2).

References

Carnahan, B., Luther, H. A. and Wilkes, J. O. (1969). Applied Numerical Methods. John Wiley, New York.

Davis, P. J. (1965). Gamma function and related functions. In Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (Edited by Abramowitz M. and Stegun, I. A.). National Bureau of Standards. Applied Mathematics Series **55**, 253-293.

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

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. In Multivariate Analysis (Edited by Krishnaiah-V), 35-57. Amsterdam, North-Holland.

van Dyk, D. A., Meng, X. L. and Rubin, D. B. (1995). Maximum likelihood estimation via the ECM algorithm: Computing the asymptotic variance. Statistica Sinica **5**, 000-000.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). Statist. Sci. **7**, 457-511.

Goodnight, J. H. (1979). A tutorial on the sweep operator. Amer. Statist. **33**, 149-158.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using the **t** distribution. J. Amer. Statist. Assoc. **84**, 881-896.

Lange, K. and Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression. J. Comput. & Graph. Statist. **2**, 175-198.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. Appl. Statist. **37**, 23-38.

Little, R. J. A. and Rubin, D. B. (1987). Statistical Analysis with Missing Data. John Wiley, New York.

Liu, C. (1993). Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. J. Multivariate Anal. **46**, 198-206.

Liu, C. H. (1995). Missing data imputation using the multivariate t distribution. To appear in J. Multivariate Anal.

Liu, C. H. and Rubin, D. B. (1994). A simple extension of EM and ECM with faster monotone convergence. To appear in Biometrika.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a binding problem. To appear in J. Amer. Statist. Assoc. **89**.

Liu, J. S., Wong, W. and Kong, A. (1994). Correlation structure and the convergence of the Gibbs sampler. To appear in J. Roy. Statist. Soc. Ser.B **56**.

Meng, X. L. (1994). On the rate of convergence of the ECM algorithm. Ann. Statist. **22**, 326-339.

Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. J. Amer. Statist. Assoc. **86**, 899-909.

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

Murray, G. D. (1977). Comment on "Maximum likelihood estimation from incomplete data via the EM algorithm," by A. P. Dempster, N. M. Laird and D. B. Rubin, J. Roy. Statist. Soc. Ser.B **39**, 27-28.

Rubin, D. B. (1976). Inference and missing data. Biometrika **63**, 581-592.

Rubin, D. B. (1983). Iteratively reweighted least squares. In Encyclopedia of Statistical Sciences, Vol. 4, 272-275. John Wiley, New York.

Rubin, D. B. and Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. Proceedings of the Statistical computing Section of the American Statistical Association, 83-88.

Shih, W. J. and Weisberg, S. (1986). Assessing influence in multiple linear regression with incomplete data. Technometrics **28**, 231-239.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. Ann. Statist. **11**, 95-103.

Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

(Received October 1993; accepted August 1994)