Journal of Multivariate Analysis **63**, 296–312 (1997)
Article No. MV971703

ML Estimation of the Multivariate *t* Distribution and the EM Algorithm

Chuanhai Liu

Bell Laboratories, Lucent Technologies

Maximum likelihood estimation of the multivariate *t* distribution, especially with unknown degrees of freedom, has been an interesting topic in the development of the EM algorithm. After a brief review of the EM algorithm and its application to finding the maximum likelihood estimates of the parameters of the *t* distribution,

than the previous procedures. Most important, the idea of this new implementation is quite general and useful for the development of the EM algorithm. Comparisons of different methods based on two datasets are presented. © 1997 Academic Press

1. INTRODUCTION

The multivariate *t* distribution is a useful distribution in statistics and data analysis for, for example, providing robust procedures for estimation (e.g., Jeffreys, 1939; 1957, pp. 64–65; Maronna, 1976; Lange *et al*., 1987; Liu, 1996). The development of likelihood-based methods for estimation of the multivariate *t* distribution has stimulated many general methods, such as the ECME algorithm (Liu and Rubin, 1994), an extension of the EM algorithm (Dempster *et al*., 1977). The EM algorithm is a very popular method in statistical computation for maximum likelihood estimation due to its simplicity and stable convergence (Dempster *et al*., 1977; Wu, 1983).

The EM algorithm is the algorithm that iteratively finds (via the E-step) and maximizes (via the M-step) a current approximation, the Q function, to the real log-likelihood function, the L function, of the parameters of a probability model. To find a Q function, the observed dataset is augmented into a *complete* dataset by the inclusion of *missing values*. A Q function is obtained as the expectation of the complete-data log-likelihood function

Received November 20, 1996.

Key words and phrases: EM, ECM, ECME, incomplete data, Newton–Raphson.

0047-259X/97 $25.00

over the missing values given the observed data and the current estimate of the parameters.

When the M-step of EM is difficult, it can be replaced with a sequence of constrained (on some functions of parameters) maximizations of the Q function, called CM-steps. This extension of the EM algorithm is called the ECM algorithm by Meng and Rubin (1993). Liu and Rubin (1994) realized that a faster converging algorithm could often be obtained by replacing some of CM-steps of ECM with CM-steps that maximize the corresponding constrained L function. For the sake of convenience, we denote by "CMQ-step" a step maximizing a constrained Q function and by "CML-step" a step maximizing a constrained L function. This extension of EM and ECM is called the ECME algorithm by Liu and Rubin (1994), with "E" for "neither." As noted by Meng and van Dyk (1997), after a sequence of CML-steps an E-step is generally required before a call to a sequence of CMQ-steps to guarantee the monotone convergence of the likelihood. Starting with the CML algorithm that sequentially maximizes constrained L functions, Fessler and Hero (1994) considered an EM-step, i.e., an iteration of the EM algorithm, for each CML-step and proposed the SAGE algorithm (for space-alternating generalized EM). Meng and van Dyk (1997) extended EM, ECM, ECME, and SAGE further to allow data-augmentation schemes as well as the constraining functions for the CM-steps to vary from a CM-step to another CM-step. They call this algorithm AECM (for alternating ECM).

The development of these extensions of the EM algorithm has a very close relationship with the study of methods for maximum likelihood estimation of the *t* distribution. A brief history of the theoretical development for ML estimation of the *t* distribution is given in Liu and Rubin (1995), who wrote:

> 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).

The history continues as follows. Kent *et al.* (1994) constructed an EM algorithm for fitting *t* distributions via a "curious likelihood identity" and found that a modification of the algorithm converges faster than the conventional EM algorithm for the *t*. This aroused Meng and van Dyk's (1997) curiosity about the method of Kent *et al.* (1994). Meng and van Dyk (1997) showed that this modification is the EM algorithm that corresponds to one of a class of data augmentation schemes for the *t* distribution using normal distributions. Liu and Rubin (1995) applied a CML-step for updating the unknown degrees of freedom. The examples and interpretations of the theoretical results of Liu and Rubin (1994, 1995) show that this ECME scheme dramatically improves the EM algorithm for the *t* distribution with unknown degrees of freedom. They also explicitly defined an MC–ECM scheme for maximum likelihood estimation of the *t* distribution with unknown degrees of freedom. Kowalski *et al.* (1996) and Meng and van Dyk (1997) demonstrated that more promising versions of ECME for the *t* can be obtained by using alternative data augmentation schemes in the CMQ-step for updating the center and scatter matrix.

Following Meng and van Dyk (1997), here we consider a class of data augmentation schemes even more general than the class of Meng and van Dyk (1997) for maximum likelihood estimation of the multivariate *t* distribution. This leads to new versions of ECME for maximum likelihood estimation of the *t* distribution with possible missing values. This algorithm is, in fact, obtained by maximizing the likelihood function over an *expanded* parameter. Most important, the idea of this paper is quite general and useful for new implementations of the EM algorithms for many statistical models. It appears that these new procedures converge even faster than the modifications of the algorithms of Liu and Rubin (1995) discussed by Meng and van Dyk (1997). The theoretical derivations show that, with a particular CML-step of ECME that updates a proportionality constant of the scatter matrix, all members of the broader class lead to some new version of ECME.

Section 2 defines a class of data augmentation schemes for the multivariate t distribution, which includes the class used by Meng and van Dyk (1997) as a subset. Under this class of data augmentation schemes, Section 3 gives the CMQ-step for updating the center and the scatter matrix (up to a proportionality constant) of the multivariate *t* distribution with fixed degrees of freedom. The theoretical results show that all the CMQ-steps generated with the data augmentation schemes in this class are the same, which implies that the conventional data augmentation scheme for the *t* distribution results in as good a CMQ-step for the center and the scatter matrix up to a proportionality constant as those using the larger class of data augmentation schemes. Section 4 discusses techniques for updating the proportionality constant with a CML-step, which of course does not depend on any data augmentation scheme. A new approach to the maximum likelihood estimates of the *t* distribution using ECME is thus proposed for the case of fixed degrees of freedom. This approach also motivates two more versions of the ECME algorithm for the multivariate *t* distribution with unknown degrees of freedom. One version consists of a CMQ-step for the center and the scatter matrix up to a proportionality constant, a CML-step for the proportionality constant, and another CML-step for the degrees of freedom. The other version uses a single CML-step to update the proportionality constant and the degrees of freedom simultaneously. The detailed procedures are given in Section 5. Comparisons of the different versions of MC-ECM and ECME for maximum likelihood estimation of the *t* distribution are presented in Section 6 based on numerical results for two datasets. Section 7 gives a concluding discussion.

2. A CLASS OF DATA AUGMENTATION SCHEMES FOR THE MULTIVARIATE *t* DISTRIBUTION

The multivariate *t* distribution is the distribution with the density function

$$f(Y\,|\,\theta)=\frac{\Gamma\left(\dfrac{v+p}{2}\right)|\Psi|^{-1/2}}{(\pi v)^{p/2}\,\Gamma\left(\dfrac{v}{2}\right)\left[1+\dfrac{1}{v}(Y-\mu)'\,\Psi^{-1}(Y-\mu)\right]^{(v+p)/2}}, \tag{1}$$

where $\theta=(\mu, \Psi, v)\in\Theta=\{(\mu, \Psi, v)\colon \Psi>0, v>0\}$ and μ, Ψ, and v are called respectively the center, scatter matrix, and degrees of freedom. The conventional EM approach to maximum likelihood estimation of the parameters θ is to augment the observed data using the following model (e.g., Liu and Rubin, 1995):

$$Y\,|\,\theta, \tau \sim N_p(\mu, \Psi/\tau) \tag{2}$$

and

$$\tau\,|\,\theta \sim \text{Gamma}(v/2, v/2) \qquad (v>0) \tag{3}$$

with density function

$$f(\tau\,|\,\theta)\,d\tau=\frac{1}{\Gamma(v/2)}\left(\frac{v}{2}\right)^{v/2}\tau^{v/2-1}\exp\left\{-\frac{v\tau}{2}\right\}d\tau \qquad (\tau>0).$$

The marginal distribution of Y specified by Eqs. (2) and (3) is uniquely defined as in (1) when Ψ/τ in (2) is replaced with $a(\theta)\,\Psi/w$, where $a(\theta)$ is any positive function of θ and $w=a(\theta)\,\tau$. This leads to a class of data augmentation schemes, including the scheme specified by (2) and (3), for the multivariate t distribution, which is described by the following models:

$$Y \mid \theta, \tau \sim N_p\left(\mu, \frac{a(\theta)}{w}\,\Psi\right) \tag{4}$$

and

$$w/a(\theta) \mid \theta \sim \text{Gamma}(\nu/2, \nu/2) \qquad (\nu > 0), \tag{5}$$

that is,

$$f(w \mid \theta)\,dw = \frac{1}{a(\theta)\,\Gamma(\nu/2)}\left(\frac{\nu}{2}\right)^{\nu/2}\left(\frac{w}{a(\theta)}\right)^{\nu/2-1}\exp\left\{-\frac{\nu w}{2a(\theta)}\right\}dw \qquad (w > 0).$$

Meng and van Dyk (1997) considered the subclass of this class of data augmentation schemes with the constraint $a(\theta) = |\Psi|^{-\alpha}$, where α, which does not involve Ψ, is called a "working parameter". The idea of replacing τ in (2) with $w = a(\theta)\,\tau$ is to allow the fractional missing-data information of the data augmentation scheme, which dominates the rate of convergence of EM, to vary according to $a(\theta)$. The fastest converging algorithm can be obtained by finding the $a(\theta)$ with "smallest" fractional missing-data information. For a general $a(\theta)$, this approach will generally result in more difficult CMQ-steps because the parameters μ, Ψ, and ν are no longer distinct in the sense that the parameters θ_1 and θ_2 in the factorization $f(Y, w \mid \theta) = f(Y \mid w, \theta_1) \cdot f(w \mid \theta_2)$ are not independent of each other. We will focus on the class with $a(\theta)$ in (4) and (5) being a function of $|\Psi|$ and ν. In the next section, we consider possible alternatives of the previous versions of the EM algorithms under this broader class of data augmentation schemes.

3. UPDATING THE CENTER AND SCATTER MATRIX UP TO A PROPORTIONALITY CONSTANT WITH FIXED DEGREES OF FREEDOM VIA A CMQ-STEP

Given n independent observations from a multivariate t distribution with unknown parameters θ, with possible ignorable missing values, $Y_i = \{Y_{i,\,\text{obs}}, Y_{i,\,\text{mis}}\}$, where $Y_{i,\,\text{obs}}$ and $Y_{i,\,\text{mis}}$ are respectively the observed and missing components of Y_i, for $i = 1, ..., n$, we have n additional missing values w_i ($i=1, ..., n$), where (Y_i, w_i) follows the model specified by Eqs. (4) and (5). The complete-data likelihood function is then

$$\prod_{i=1}^{n} \left|\frac{a(\theta)}{w_i}\Psi\right|^{-1/2} \exp\left\{-\frac{w_i}{2a(\theta)}(Y_i-\mu)'\,\Psi^{-1}(Y_i-\mu)\right\}$$
$$\times\frac{(v/2)^{v/2}}{a(\theta)\,\Gamma(v/2)}\left(\frac{w_i}{a(\theta)}\right)^{v/2-1} \exp\left\{-\frac{vw_i}{2a(\theta)}\right\}.$$

To update the center and the scatter matrix with fixed v, we have the complete-data log-likelihood function:

$$L(\theta)=-\frac{n(v+p)}{2}\ln(a(\theta))-\frac{n}{2}\ln|\Psi|$$
$$-\frac{1}{2a(\theta)}\,\mathrm{trace}\left[\Psi^{-1}\sum_{i=1}^{n} w_i(Y_i-\mu)(Y_i-\mu)'\right]-\frac{v}{2a(\theta)}\sum_{i=1}^{n} w_i. \tag{6}$$

We let $p_{i,\,\mathrm{obs}}$ be the dimension of $Y_{i,\,\mathrm{obs}}$, $\mu_{i,\,\mathrm{obs}}$ be the sub-vector of μ corresponding to the observed components of Y_i, and $\Psi_{i,\,\mathrm{obs}}=\Psi_{i,\,\mathrm{obs,\,obs}}$, $\Psi_{i,\,\mathrm{obs,\,mis}}$, $\Psi_{i,\,\mathrm{mis,\,obs}}$, and $\Psi_{i,\,\mathrm{mis}}=\Psi_{i,\,\mathrm{mis,\,mis}}$ be the sub-matrices of Ψ corresponding to the observed and missing components of Y_i. We write $\tilde{\theta}$ for the previous estimate of θ and $\hat{\theta}$ for the updated (i.e., current) estimate of θ. For any $\tilde{\theta}\in\Theta$, (6) can be represented as

$$-\frac{n(v+p)}{2}\ln(a(\theta))-\frac{n}{2}\ln|\Psi|$$
$$-\frac{a(\tilde{\theta})}{2a(\theta)}\,\mathrm{trace}\left[\Psi^{-1}\sum_{i=1}^{n}\frac{w_i}{a(\tilde{\theta})}(Y_i-\mu)(Y_i-\mu)'\right]-\frac{va(\tilde{\theta})}{2a(\theta)}\sum_{i=1}^{n}\frac{w_i}{a(\tilde{\theta})}$$

and

$$w_i/a(\tilde{\theta})(\tilde{\theta},\,Y_{\mathrm{obs}})\sim\mathrm{Gamma}\left(\frac{v+p_{i,\,\mathrm{obs}}}{2},\frac{v+\tilde{\delta}_{i,\,\mathrm{obs}}}{2}\right),$$

where

$$\tilde{\delta}_{i,\,\mathrm{obs}}=(Y_{i,\,\mathrm{obs}}-\tilde{\mu}_{i,\,\mathrm{obs}})'\,\tilde{\Psi}_{i,\,\mathrm{obs}}^{-1}(Y_{i,\,\mathrm{obs}}-\tilde{\mu}_{i,\,\mathrm{obs}}).$$

Following, for example, Liu and Rubin (1995), we have that the expected complete-data log-likelihood function is

$$\begin{aligned} Q(\theta \mid \tilde{\theta}) = & -\frac{n(\nu+p)}{2} \ln(a(\theta)) - \frac{n}{2} \ln |\Psi| \\ & - \frac{a(\tilde{\theta})}{2a(\theta)} \operatorname{trace}\left[\Psi^{-1} \left(\sum_{i=1}^{n} \tilde{\tau}_i (\tilde{Y}_i - \mu)(\tilde{Y}_i - \mu)' + \sum_{i=1}^{n} \tilde{\Phi}_{i,\,\text{mis}} \right) \right] \\ & - \frac{\nu a(\tilde{\theta})}{2a(\theta)} \sum_{i=1}^{n} \tilde{\tau}_i \end{aligned} \tag{7}$$

for any $\theta \in \Theta$ and $\tilde{\theta} \in \Theta$, where $\tilde{Y}_i = E(Y_i \mid Y_{\text{obs}}, \theta = \tilde{\theta})$, $\tilde{\tau}_i = E(w_i / a(\tilde{\theta}) \mid Y_{\text{obs}}, \theta = \tilde{\theta}) = (\nu + p_{i,\,\text{obs}})/(\nu + \tilde{\delta}_{i,\,\text{obs}})$, and $\tilde{\Phi}_{i,\,\text{mis}}$ is a $(p \times p)$ matrix with the (j, k)th element being the corresponding element of $\tilde{\Psi}_{i,\,\text{mis}} - \tilde{\Psi}_{i,\,\text{mis, obs}} \tilde{\Psi}^{-1}_{i,\,\text{obs}} \tilde{\Psi}_{i,\,\text{obs, mis}}$ if both the jth element and the kth element of Y_i are missing, and zero otherwise.

For any Ψ with fixed ν, $\hat{\mu}$ is the weighted average of $\tilde{Y}_i$ with weights $\tilde{\tau}_i$ for $i = 1, ..., n$, i.e.,

$$\hat{\mu} = \frac{\tilde{S}_{\tau Y}}{\tilde{S}_{\tau}},$$

where

$$\tilde{S}_{\tau Y} = \sum_{i=1}^{n} \tilde{\tau}_i \tilde{Y}_i \qquad \text{and} \qquad \tilde{S}_{\tau} = \sum_{i=1}^{n} \tilde{\tau}_i. \tag{8}$$

To obtain $\hat{\Psi}$, we maximize

$$Q_1(\theta \mid \tilde{\theta}) = -\frac{n(\nu+p)}{2} \ln(a(\theta)) - \frac{n}{2} \ln |\Psi| - \frac{a(\tilde{\theta})}{2a(\theta)} [\operatorname{trace}(\Psi^{-1} \tilde{S}_{\tau YY}) + \nu \tilde{S}_{\tau}], \tag{9}$$

where

$$\tilde{S}_{\tau YY} = \sum_{i=1}^{n} \tilde{\tau}_i (\tilde{Y}_i - \hat{\mu})(\tilde{Y}_i - \hat{\mu})' + \sum_{i=1}^{n} \tilde{\Phi}_{i,\,\text{mis}}. \tag{10}$$

Differentiating (9) with respect to Ψ^{-1}, we have

$$\begin{aligned} \frac{\partial Q_1(\theta \mid \tilde{\theta})}{\partial \Psi^{-1}} = & \left\{ -\frac{n(\nu+p)}{2a(\theta)} \frac{\partial a(\theta)}{\partial |\Psi^{-1}|} + [\operatorname{trace}(\Psi^{-1} \tilde{S}_{\tau YY}) + \nu \tilde{S}_{\tau}] \right. \\ & \left. \times \frac{a(\tilde{\theta})}{2a^2(\theta)} \frac{\partial a(\theta)}{\partial |\Psi^{-1}|} - \frac{n}{2 |\Psi^{-1}|} \right\} \\ & \times |\Psi^{-1}| \, [2\Psi - \operatorname{Diag}(\Psi)] - \frac{a(\tilde{\theta})}{2a(\theta)} [2\tilde{S}_{\tau YY} - \operatorname{Diag}(\tilde{S}_{\tau YY})]. \end{aligned} \tag{11}$$

From (11), we see that $\hat{\Psi}$ is proportional to $\tilde{S}_{\tau YY}$ for any given function $a(\theta)$. Letting

$$\hat{\Psi}^* = \frac{1}{n}\tilde{S}_{\tau YY}, \tag{12}$$

we have

$$\hat{\Psi} = \frac{1}{k}\hat{\Psi}^*, \tag{13}$$

where k is a scalar proportionality constant. Letting $k=1$, Eq. (13) is the formula for updating the estimate of Ψ using the conventional data augmentation. Letting $k=\tilde{S}_\tau/n$, (13) gives the formula for updating the estimate of Ψ using the data augmentation scheme with $a(\theta)=|\Psi|^{-1/(p+v)}$ used by Meng and van Dyk (1997).

For a general $a(\theta)$, a function $|\Psi|$ and v, there is no closed form solution for k from Eq. (11). However, we can treat k as an unknown parameter to be further estimated. In other words, we consider a CM-step with the parameters constrained so that $\Psi = \hat{\Psi}^*/k$. To *kill two birds at one shot*, we will determine the *expanded* parameter k using a CML-step instead of a CMQ-step because a CML-step avoids solving Eq. (11) for k and provides a faster converging algorithm.

4. UPDATING THE PROPORTIONALITY CONSTANT OF THE SCATTER MATRIX WITH FIXED DEGREES OF FREEDOM VIA A CML-STEP

To update k using a CML-step with constraints $\mu=\hat{\mu}$, $\Psi=\hat{\Psi}^*/k$, and $v=\hat{v}$, we try to find $\hat{k}$ such that $\hat{\Psi}=\hat{\Psi}^*/\hat{k}$ maximizes the constrained actual likelihood, i.e.,

$$L_1(k) = \sum_{i=1}^{n}\frac{p_{i,\,\mathrm{obs}}}{2}\ln(k) - \sum_{i=1}^{n}\frac{v+p_{i,\,\mathrm{obs}}}{2}\ln(v+\tilde{\Delta}_{i,\,\mathrm{obs}}k), \tag{14}$$

where

$$\tilde{\Delta}_{i,\,\mathrm{obs}} = (Y_{i,\,\mathrm{obs}} - \hat{\mu}_{i,\,\mathrm{obs}})'\,(\hat{\Psi}^*_{i,\,\mathrm{obs}})^{-1}\,(Y_{i,\,\mathrm{obs}} - \hat{\mu}_{i,\,\mathrm{obs}}).$$

This can be done using the Newton–Raphson method, which requires the computation of the following quantities:

$$\frac{\partial L_1(k)}{\partial k}=\frac{1}{2k}\sum_{i=1}^{n} p_{i,\,\mathrm{obs}}-\frac{1}{2}\sum_{i=1}^{n}\frac{(v+p_{i,\,\mathrm{obs}})\,\tilde{\Delta}_{i,\,\mathrm{obs}}}{v+\tilde{\Delta}_{i,\,\mathrm{obs}}k} \tag{15}$$

and

$$\frac{\partial^2 L_1(k)}{(\partial k)^2}=-\frac{1}{2k^2}\sum_{i=1}^{n} p_{i,\,\mathrm{obs}}+\frac{1}{2}\sum_{i=1}^{n}\frac{(v+p_{i,\,\mathrm{obs}})(\tilde{\Delta}_{i,\,\mathrm{obs}})^2}{(v+\tilde{\Delta}_{i,\,\mathrm{obs}}k)^2}. \tag{16}$$

Apparently, there are problems in using the Newton–Raphson algorithm with Eqs. (15) and (16) because (16) can be positive. Alternatively, we can find $\hat{k}$ such that $\partial L_1(k)/\partial k|_{k=\hat{k}}=0$ using the iterative formula:

$$k_{j+1}=\sum_{i=1}^{n} p_{i,\,\mathrm{obs}}\Big/\sum_{i=1}^{n}\frac{(v+p_{i,\,\mathrm{obs}})\,\tilde{\Delta}_{i,\,\mathrm{obs}}}{v+\tilde{\Delta}_{i,\,\mathrm{obs}}k_j} \qquad \text{for} \quad j=0,1,\ldots. \tag{17}$$

As shown in Proposition 1 in the Appendix, (17) monotonically converges to the unique root of Eq. (15), which maximizes the constrained L function (14). Although the solution for k from Eq. (15) is not in closed form, numerical results suggest that the number of iterations for (17) in practice is typically small, one or two steps, with the starting point being the previous estimate.

5. MAXIMUM LIKELIHOOD ESTIMATION OF THE MULTIVARIATE t WITH UNKNOWN DEGREES OF FREEDOM

When the degrees of freedom are unknown, the maximum likelihood estimates of the parameters can be obtained using the various methods given in Liu and Rubin (1995). These methods can be easily modified using the method for updating Ψ given v in Meng and van Dyk (1997) or with the method discussed in the previous section. We expect that the last one converges faster, at least in terms of the number of iterations, than the others because the proportionality constant is updated via maximization of the constrained actual likelihood function, the major idea of ECME. The framework for this algorithm is as follows:

E-STEP: Compute $\tilde{S}_{\tau}$, $\tilde{S}_{\tau Y}$, and $\tilde{S}_{\tau YY}$ is Eqs. (8) and (10).

CMQ-STEP: Maximize the Q function over μ and Ψ up to a proportionality constant with fixed $v=\hat{v}$, that is, $\hat{\mu}=\tilde{S}_{\tau Y}/\tilde{S}_{\tau}$ and $\hat{\Psi}^{*}=\tilde{S}_{\tau YY}/n$.

CML-STEP 1: Maximize the L function over k, where $\Psi = \hat{\Psi}^*/k$, given $\mu = \hat{\mu}$, $\hat{\Psi}^*$, and $v = \tilde{v}$.

CML-STEP 2: Maximize the L function over v, the degrees of freedom, given $\mu = \hat{\mu}$ and $\Psi = \hat{\Psi}^*/\hat{k}$.

Another version of the ECME algorithm for the multivariate *t* with unknown degrees of freedom can be obtained by updating the proportionality constant and the degrees of freedom simultaneously via a CML-step. More specifically, this algorithm consists of an E-step and two CM-steps as described below:

E-STEP: Compute $\tilde{S}_\tau$, $\tilde{S}_{\tau Y}$, and $\tilde{S}_{\tau YY}$ in Eqs. (8) and (10).

CMQ-STEP: Maximize the Q function over μ and Ψ up to a proportionality constant with fixed $v = \hat{v}$, that is, $\hat{\mu} = \tilde{S}_{\tau Y}/\tilde{S}_\tau$ and $\hat{\Psi}^* = \tilde{S}_{\tau YY}/n$.

CML-STEP: Maximize the L function over k and v, where $\Psi = \hat{\Psi}^*/k$, given $\mu = \hat{\mu}$ and $\hat{\Psi}^*$, that is, maximize

$$L_2(k, v) = \sum_{i=1}^{n} \ln \Gamma\left(\frac{v + p_{i,\,\text{obs}}}{2}\right) + \frac{1}{2}\sum_{i=1}^{n} p_{i,\,\text{obs}} \ln(k) - n \ln \Gamma\left(\frac{v}{2}\right) + \frac{nv}{2}\ln(v) - \sum_{i=1}^{n} \frac{v + p_{i,\,\text{obs}}}{2} \ln(v + \tilde{\Delta}_{i,\,\text{obs}} k). \qquad (18)$$

The motivation for this procedure is based on the thought that the proportionality constant and the degrees of freedom are confounded in the way that, given a smaller proportionality constant k, the algorithm will result in smaller degrees of freedom and *vice versa*. The rate of convergence of the previous EM algorithms for ML estimation of the multivariate *t* distribution depends on the tradeoff between these two parameters. Instead of a CML-step for k and another CML-step for v, a CML-step for both k and v simultaneously would solve the trade-off problem at a much faster speed.

The maximum of $L_2(k, v)$ can be found using the Newton–Raphson method. The elements of the gradient and Hessian matrix of $L_2(k, v)$ in (18) are as follows:

$$\frac{\partial L_2(k, v)}{\partial k} = \frac{\partial L_1(k)}{\partial k}$$

$$\frac{\partial L_2(k, v)}{\partial v} = \frac{1}{2}\sum_{i=1}^{n} \phi\left(\frac{v + p_{i,\,\text{obs}}}{2}\right) - \frac{n}{2}\phi\left(\frac{v}{2}\right) + \frac{n}{2}\ln\left(\frac{v}{2}\right) + \frac{n}{2} - \frac{1}{2}\sum_{i=1}^{n} \ln\left(\frac{v + \tilde{\Delta}_{i,\,\text{obs}} k}{2}\right) - \frac{1}{2}\sum_{i=1}^{n} \frac{v + p_{i,\,\text{obs}}}{v + \tilde{\Delta}_{i,\,\text{obs}} k}$$

$$\frac{\partial^2 L_2(k,\, v)}{(\partial k)^2}=\frac{\partial^2 L_1(k)}{(\partial k)^2}$$

$$\frac{\partial^2 L_2(k,\, v)}{\partial k\, \partial v}=\frac{1}{2}\sum_{i=1}^{n}\frac{(p_{i,\,\text{obs}}-\tilde{\Delta}_{i,\,\text{obs}}k)\,\tilde{\Delta}_{i,\,\text{obs}}}{(v+\tilde{\Delta}_{i,\,\text{obs}}k)^2},$$

and

$$\frac{\partial^2 L_2(k,\, v)}{\partial v\, \partial v}=\frac{1}{4}\sum_{i=1}^{n}\phi'\left(\frac{v+p_{i,\,\text{obs}}}{2}\right)-\frac{n}{4}\phi'\left(\frac{v}{2}\right)+\frac{n}{2v}-\frac{1}{2}\sum_{i=1}^{n}\frac{1}{v+p_{i,\,\text{obs}}}$$
$$+\frac{1}{2}\sum_{i=1}^{n}\frac{(p_{i,\,\text{obs}}-\tilde{\Delta}_{i,\,\text{obs}}k)^2}{(v+p_{i,\,\text{obs}})(v+\tilde{\Delta}_{i,\,\text{obs}}k)^2}, \tag{19}$$

where $\phi(\cdot)$ is the digamma function and $\phi'(\cdot)$ is the trigamma function. Since $\partial^2 L_2(k, v)/(\partial v)^2$ is not necessarily negative, we can ignore the last term of the right-hand side of Eq. (19) in using the Newton–Raphson method. When $\partial^2 L_2(k, v)/(\partial k)^2 > 0$, we can apply the algorithm (17) to adjust k. These modifications are closely related to Gauss–Newton and quasi-Newton methods (see, e.g., Lange, 1995).

To get a rough idea about the rate of convergence of this version of ECME, as in Liu and Rubin (1994), we find that the large sample rate of convergence of this algorithm for the univariate t distribution with unknown degrees of freedom and the data from $t_1(\mu_0, \Psi_0, v_0)$ is $2/(v_0+3)$, which is even smaller (better) than $3/(v_0+3)$, the large sample rate of convergence of the EM algorithm for the univariate t distribution with fixed $v=v_0$ and the data from $t_1(\mu_0, \Psi_0, v_0)$. In the next section, we provide some numerical results for comparison of the different procedures based on two datasets.

6. NUMERICAL EXAMPLES

In this section, we compare the versions of the MC–ECM and ECME algorithms for maximum likelihood estimation of the multivariate t distribution for the two datasets presented in Liu and Rubin (1995); both are given in Tables 1 and 4 of Liu and Rubin (1995), although the second is originally from Shih and Weisberg (1986). The first dataset consists of 16 observations of two variables with 12 missing values, 6 from each variable. The second dataset consists of 34 observations of four variables RC, WT, SC, and Age from a clinical trial with 2 missing values from WT and 4 missing values from SC. For the second dataset, as in Liu and Rubin (1995), we use the multivariate t distribution for the four variables ln(RC), ln(WT), ln(SC), and ln(140-Age).

Table I

The Versions of the MC-ECM and ECME Algorithms Used for Comparisons

	Parameter Partitions		
Algorithm	CMQ	CML	Primary References
MC-ECM1	$\{\mu, \Psi\}, \{v\}$		Liu & Rubin (1995)
MC-ECM2	$\{\mu, \Psi\}, \{v\}$		Liu & Rubin (1995); Meng & van Dyk (1995)
ECME1	$\{\mu, k\Psi\}, \{v\}$	$\{k\}$	Liu & Rubin (1995); This article
ECME2	$\{\mu, \Psi\}$	$\{v\}$	Liu & Rubin (1995)
ECME3	$\{\mu, \Psi\}$	$\{v\}$	Liu & Rubin (1995); Meng & van Dyk (1995)
ECME4	$\{\mu, k\Psi\}$	$\{v\}, \{k\}$	Liu & Rubin (1995); This article
ECME5	$\{\mu, k\Psi\}$	$\{v, k\}$	This article

We consider the algorithms listed in Table I. When updating the degrees of freedom and/or the proportionality constant, the ECME versions use the most current estimates of the center and scatter matrix. Accordingly, MC-ECM instead of ECM is considered here for reasonable comparison in terms of the number of iterations.

All algorithms are started with the same initial values of the center and scatter matrix, their maximum likelihood estimates under the multivariate normal distribution, which are obtained using the EM algorithm (e.g., Little and Rubin, 1987). For the first dataset, there are three sets of stationary points, which are given in Liu and Rubin (1995); we use the set with positive $\hat{\Psi}_{1,2}$. For each dataset, we consider both the case of unknown degrees of freedom and the case of fixed degrees of freedom. With unknown

Table II

The Numbers of Iterations [and CPU Times (sec)] of the Algorithms Listed in TABLE I until Convergence

	For Dataset 1				For Dataset 2			
Algorithm	Unknown d.f.		Fixed d.f. = 1.40		Unknown d.f.		Fixed d.f. = 6.51	
MC-ECM1	5903	[4.97]	116	[0.08]	317	[0.78]	15	[0.03]
MC-ECM2	5900	[4.92]	113	[0.08]	315	[0.75]	13	[0.03]
ECME1	5896	[5.07]	109	[0.10]	313	[0.79]	12	[0.03]
ECME2	130	[0.22]	—	—	30	[0.15]	—	—
ECME3	128	[0.21]	—	—	27	[0.13]	—	—
ECME4	124	[0.23]	—	—	25	[0.13]	—	—
ECME5	124	[0.30]	—	—	23	[0.16]	—	—

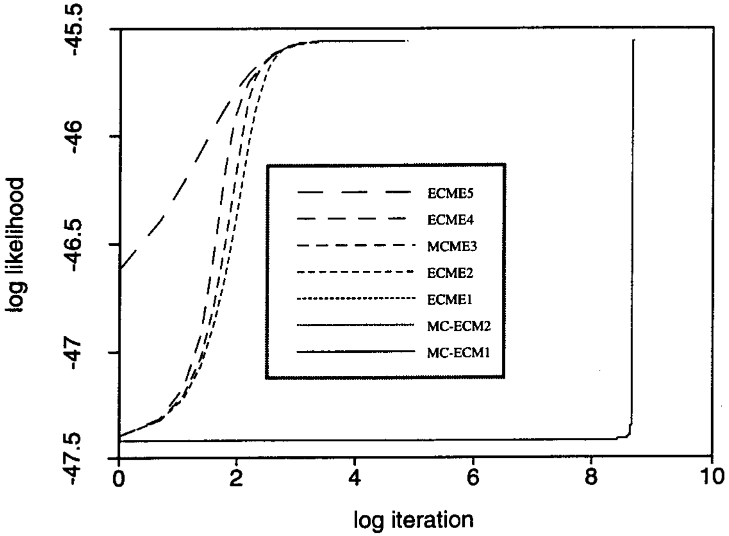

Fig. 1. Likelihood convergence of the algorithms listed in Table I for the dataset in Table I of Liu and Rubin (1995) with unknown degrees of freedom. In the plot, the MC-ECM1, MC-ECM2, and ECME1 are overlaid with each other.

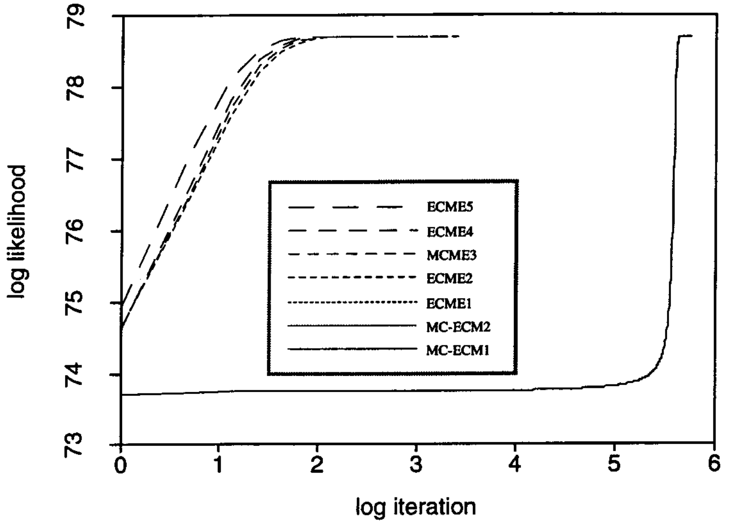

Fig. 2. Same as Fig. 1 but for the dataset in Table I of Shih and Weisberg (1986) with the transformations ln(RC), ln(WT), ln(SC), and ln(140-Age).

degrees of freedom, we let the starting value of degrees of freedom be 1000.0. With fixed degrees of freedom, we fix the degrees of freedom to be the maximum likelihood estimate of degrees of freedom, obtained by fitting the multivariate t distribution to the data with unknown degrees of freedom.

Table II gives the numbers of iterations of the algorithms until convergence, with the tolerance parameter 1.0×10^{-5} for each component of the difference between the current and previous estimates of the parameters. The sequences of the likelihood values produced by the algorithms for the two datasets with unknown degrees of freedom are displayed in Figs. 1 and 2. From Table II, we see that the ECME algorithms (ECME2, ECME3, ECME4, and ECME5) with a CML-step for the unknown degrees of freedom converge dramatically faster (in iterations) than those (MC-ECM1, MC-ECM2, and ECME1) with a CMQ-step for the degrees of freedom. The total CPU times of the ECME algorithms with a CML-step for the unknown degrees of freedom are also substantially less than those with a CMQ-step for the degrees of freedom. Comparing the results of MC-ECM1 (or ECME2) with the results of MC-ECME2 (or ECME3), we see that the algorithms (MC-ECM1 and ECME3) modified using the "optimal" data augmentation scheme of Meng and van Dyk (1997) converge somewhat faster than the corresponding algorithms (MC-ECM1 and ECME2) of Liu and Rubin (1995). The new algorithms ECME1 and ECME4 converge slightly faster respectively than the corresponding old versions. But ECME5, which updates the proportionality constant and the degrees of freedom simultaneously, appears to be the fastest algorithm, as seen clearly from Figs. 1 and 2.

7. DISCUSSION

Within the past few years, recent developments of the EM algorithms have provided about 10 ways to implement the EM, ECM, and ECME algorithms for maximum likelihood estimation of the multivariate t distribution. In a complementary way, the study of methods for maximum likelihood estimation of the multivariate t distribution has actually stimulated new ideas for generalizing and implementing EM and its extensions. This paper provides new ways to implement the ECME algorithm for the multivariate t distribution. The ideas behind the new versions of ECME of this article can lead to new implementations of the ECME algorithm for other models, such as variance components models. For a variance components model, for example, we can consider replacing the CM-step 3 of ECME version 2 in Liu and Rubin (1994) with a CML-step for updating both the residual variance and a proportionality constant of the covariance matrix of the random effects simultaneously. It is worthwhile comparing different approaches to maximum likelihood estimation of the variance components model because it is such a broadly applied model.

For the multivariate t distribution, we have seen that a faster converging algorithm can be obtained without the alternatives of the data augmentation schemes in the class defined in Section 2. Nevertheless, it is definitely a good idea to consider alternative data augmentation schemes in using the EM algorithms in practice, as emphasized by Meng and van Dyk (1997). For example, with an incomplete dataset from the multivariate normal distribution, the *monotone* EM algorithm based on the monotone data augmentation scheme described in Rubin and Schafer (1990) will certainly beat the *rectangular* EM algorithm based on the conventional data augmentation strategy, which augments the incomplete dataset into the entire rectangular dataset. When the missing-data pattern in an incomplete normal dataset is monotone, this *monotone* EM algorithm is not iterative and therefore converges in ONE iteration.

APPENDIX

PROPOSITION 1. (1) *The equation $\partial L_1(k)/\partial k = 0$ has a unique root $\hat{k}$.*

(2) *The iterative formula* (17) *monotonically converges to $\hat{k}$.*

(3) *$\hat{k}$ is the unique maximum of $L_1(k)$.*

Proof. Let

$$h(k) = 2k\,\partial L_1(k)/\partial k = \sum_{i=1}^{n} p_{i,\,\mathrm{obs}} - \sum_{i=1}^{n} \frac{(\nu + p_{i,\,\mathrm{obs}})}{\nu/(\tilde{\Delta}_{i,\,\mathrm{obs}} k) + 1},$$

then $h(k)$ is a strictly monotone function of k. As $k \to \infty$, $h(k) \to -n\nu < 0$. As $k \to 0$, $h(k) \to \sum_{i=1}^{n} p_{i,\,\mathrm{obs}} > 0$. Thus, (1) is proved.

We show that (2) is true for the case of $k_0 > \hat{k}$. For the case of $k_0 < \hat{k}$, the proof is similar. First, we show that $k_{j+1} > \hat{k}$ if $k_j > \hat{k}$. The proof is straightforward because

$$\sum_{i=1}^{n} p_{i,\,\mathrm{obs}} \bigg/ \sum_{i=1}^{n} \frac{(\nu + p_{i,\,\mathrm{obs}})\,\tilde{\Delta}_{i,\,\mathrm{obs}}}{\nu + \tilde{\Delta}_{i,\,\mathrm{obs}} k}$$

is a strictly increasing function of k. Thus

$$\frac{k_{j+1} - \hat{k}}{k_j - \hat{k}} > 0 \qquad \text{for} \quad j = 0, 1, \ldots.$$

Second, we show that for any $\varepsilon > 0$, if $k_j - \hat{k} > \varepsilon$, there exists a positive number $M(\varepsilon)$ such that

$$\frac{k_{j+1} - \hat{k}}{k_j - \hat{k}} < M(\varepsilon) < 1.$$

Simple algebraic operations lead to

$$\frac{k_{j+1} - \hat{k}}{k_j - \hat{k}} = \frac{b(k_j, \hat{k})\, k_j - \hat{k}}{k_j - \hat{k}},$$

where

$$b(k_j, \hat{k}) = \sum_{i=1}^{n} \frac{\nu + p_{i,\,\mathrm{obs}}}{\nu/(\tilde{\Delta}_{i,\,\mathrm{obs}} \hat{k}) + 1} \Bigg/ \sum_{i=1}^{n} \frac{\nu + p_{i,\,\mathrm{obs}}}{\nu/(\tilde{\Delta}_{i,\,\mathrm{obs}} k_j) + 1}.$$

Since $b(k_j, \hat{k})$ is a strictly decreasing function of k_j with fixed $\hat{k}$, $b(k_j, \hat{k}) < b(\hat{k} + \varepsilon, \hat{k})$. This implies that

$$\frac{k_{j+1} - \hat{k}}{k_j - \hat{k}} < \frac{b(\hat{k} + \varepsilon, \hat{k})\, k_j - \hat{k}}{k_j - \hat{k}} = M(\varepsilon) < \frac{b(\hat{k}, \hat{k})\, k_j - \hat{k}}{k_j - \hat{k}} = 1.$$

The proof of (2) is complete.

Since

$$\begin{aligned} k\,\frac{\partial^2 L_1(k)}{(\partial k)^2} &= -\frac{1}{2k} \sum_{i=1}^{n} p_{i,\,\mathrm{obs}} + \frac{1}{2} \sum_{i=1}^{n} \frac{(\nu + p_{i,\,\mathrm{obs}})\, \tilde{\Delta}_{i,\,\mathrm{obs}}}{\nu + \tilde{\Delta}_{i,\,\mathrm{obs}} k} \, \frac{\tilde{\Delta}_{i,\,\mathrm{obs}} k}{\nu + \tilde{\Delta}_{i,\,\mathrm{obs}} k} \\ &< -\frac{1}{2k} \sum_{i=1}^{n} p_{i,\,\mathrm{obs}} + \frac{1}{2} \sum_{i=1}^{n} \frac{(\nu + p_{i,\,\mathrm{obs}})\, \tilde{\Delta}_{i,\,\mathrm{obs}}}{\nu + \tilde{\Delta}_{i,\,\mathrm{obs}} k} \\ &= \frac{\partial L_1(k)}{\partial k}, \end{aligned}$$

we have $\partial^2 L_1(k)/(\partial k)^2|_{k=\hat{k}} < 0$. Equation (3) is verified. ∎

ACKNOWLEDGMENTS

The author thanks Professor Donald B. Rubin for his recognition of the generalities of the ideas for implementation of the EM algorithms in the article as well as for his helpful comments that greatly improved the presentation of the paper.

REFERENCES

[1] 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.

[2] 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* (V. Krishnaiah, Ed.), pp. 35–57. North-Holland, Amsterdam.

[3] Fessler, J. A., and Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Trans. Signal Process.* **42** 2664–2677.

[4] Jeffreys, S. H. (1939). *Theory of Probability*. Clarendon, Oxford.

[5] Jeffreys, S. H. (1957). *Scientific Inference* (2nd ed.). Cambridge University Press, Cambridge.

[6] Kent, J. T., Tyler, D. E., and Vardi, Y. (1994). A curious likelihood identity for the multivariate *t*-distribution. *Comm. Statist. Simulations* **23** 441–453.

[7] Kowalski, J., Tu, X. M., Day, R. S., and Mendoza-Blanco, J. (1997). On the rate of convergence of the ECME for multiple regression models with *t*-distributed errors. *Biometrika* **84** 269–281.

[8] Lange, K. L. (1995). A quasi-Newton acceleration of the EM algorithm. *Statist. Sinica* **5** 1–18.

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

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

[11] Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Appl. Statist.* **37** 23–39.

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

[13] Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *J. Amer. Statist. Assoc.* **91** 1219–1227.

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

[15] Liu, C., and Rubin, D. B. (1995). ML estimation of the multivariate *t* distribution with unknown degrees of freedom. *Statist. Sinica* **5** 19–39.

[16] Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4**, 51–67.

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

[18] Meng, X. L., and van Dyk, D. (1997). The EM algorithm—An old folk song sung to fast new tune (with discussion). *J. Roy. Statist. Soc.* **59** 511–567.

[19] Rubin, D. B. (1983). Iteratively reweighted least squares. In *Encyclopedia of Statistical Sciences*, Vol. 4, pp. 272–275. Wiley, New York.

[20] Rubin, D. B., and Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, pp. 83–88.

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

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