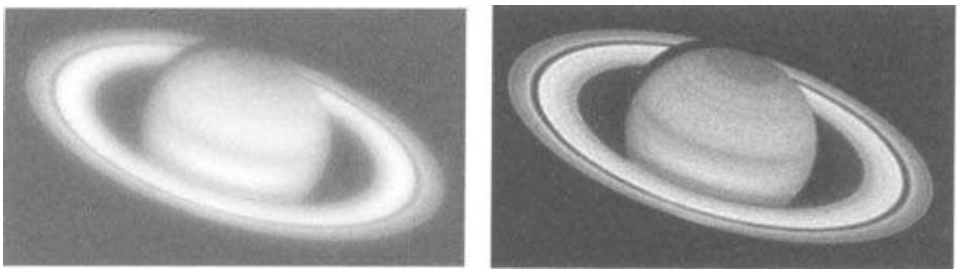

(a) (b)

Figure 2.2 (a) Blurred Image of Saturn obtained with WF/PC camera of Hubble Space Telescope reproduced with permission from Space Telescope Science Institute (STSCI); (b) its Reconstructed Image using the Richardson-Lucy Algorithm from Molina et al. (2001) reproduced with permission from IEEE ©IEEE.

2.6 EXAMPLE 2.6: MULTIVARIATE t-DISTRIBUTION WITH KNOWN DEGREES OF FREEDOM

2.6.1 ML Estimation of Multivariate t-Distribution

The multivariate t-distribution has many potential applications in applied statistics; see Kotz and Nadarajah (2004). A p-dimensional random variable $\boldsymbol{W}$ is said to have a multivariate t-distribution $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with location $\boldsymbol{\mu}$, positive definite inner product matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom if given the weight u,

$$\boldsymbol{W} \mid u \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/u), \tag{2.36}$$

where the random variable U corresponding to the weight u is distributed as

$$U \sim \text{gamma}(\tfrac{1}{2}\nu, \tfrac{1}{2}\nu). \tag{2.37}$$

The gamma (α, β) density function $f(u; \alpha, \beta)$ is given by

$$f(u; \alpha, \beta) = \{\beta^{\alpha} u^{\alpha-1}/\Gamma(\alpha)\} \exp(-\beta u) I_{[0,\infty)}(u); \qquad (\alpha, \beta > 0).$$

On integrating out u from the joint density function of $\boldsymbol{W}$ and U that can be formed from (2.36) and (2.37), the density function of $\boldsymbol{W}$ is given by

$$f_p(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+p}{2})\,|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu}{2})\{1 + \delta(\boldsymbol{w}, \boldsymbol{\mu}; \boldsymbol{\Sigma})/\nu\}^{\frac{1}{2}(\nu+p)}}, \tag{2.38}$$

where

$$\delta(\boldsymbol{w}, \boldsymbol{\mu}; \boldsymbol{\Sigma}) = (\boldsymbol{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w} - \boldsymbol{\mu})$$

denotes the Mahalanobis squared distance between $\boldsymbol{w}$ and $\boldsymbol{\mu}$ (with $\boldsymbol{\Sigma}$ as the covariance matrix). As ν tends to infinity, U converges to one with probability one, and so $\boldsymbol{W}$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

We consider now the application of the EM algorithm for ML estimation of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the t-density (2.38) in the case where the degrees of freedom ν is known. For

instance, ν can be assumed to be known in statistical analyses where different specified degrees of freedom ν are used for judging the robustness of the analyses; see Lange, Little, and Taylor (1989) and Lange and Sinsheimer (1993). For $\nu < \infty$, ML estimation of $\boldsymbol{\mu}$ is robust in the sense that observations with large Mahalanobis distances are downweighted. This can be clearly seen from the form of the equation (2.48) to be derived for the MLE of $\boldsymbol{\mu}$. As ν decreases, the degree of downweighting of outliers increases.

The EM algorithm for known ν is given in Rubin (1983), and is extended to the case with missing data in $\boldsymbol{W}$ in Little (1988) and in Little and Rubin (1987,2002, pages 257–264). Liu and Rubin (1994, 1995) and Little and Rubin (2002, pages 183–184) have shown how the MLE can be found much more efficiently by using the ECME algorithm. The use of the latter algorithm for this problem is to be described in Chapter 5, where the general case of unknown ν is to be considered. As cautioned by Liu and Rubin (1995), care must be taken especially with small or unknown ν, because the likelihood function can have many spikes with very high likelihood values but little associated posterior mass under any reasonable prior. In that case, the associated parameter estimates may be of limited practical interest by themselves, even though formally they are 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; see Gelman and Rubin (1992).

Suppose that $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ denote an observed random sample from the $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ distribution. That is,

$$\boldsymbol{y} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_n^T)^T$$

denotes the observed data vector. The problem is to find the MLE of $\boldsymbol{\Psi}$ on the basis of $\boldsymbol{y}$, where $\boldsymbol{\Psi}$ contains the elements of $\boldsymbol{\mu}$ and the distinct elements of $\boldsymbol{\Sigma}$. From (2.38), the log likelihood function for $\boldsymbol{\Psi}$ that can be formed from $\boldsymbol{y}$ is

$$\begin{aligned} \log L(\boldsymbol{\Psi}) &= \sum_{j=1}^{n} \log f_p(\boldsymbol{w}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu), \\ &= -\tfrac{1}{2}np\log(\pi\nu) + n\{\log\Gamma(\tfrac{\nu+p}{2}) - \log\Gamma(\tfrac{1}{2}\nu)\} \\ &\quad -\tfrac{1}{2}n\log|\boldsymbol{\Sigma}| + \tfrac{1}{2}n(\nu+p)\log\nu \\ &\quad -\tfrac{1}{2}(\nu+p)\sum_{j=1}^{n}\log\{\nu + \delta(\boldsymbol{w}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma})\}, \end{aligned} \tag{2.39}$$

which does not admit a closed form solution for the MLE of $\boldsymbol{\Psi}$.

In the light of the definition (2.36) of this t-distribution, it is convenient to view the observed data $\boldsymbol{y}$ as incomplete. The complete-data vector $\boldsymbol{x}$ is taken to be

$$\boldsymbol{x} = (\boldsymbol{y}^T, \boldsymbol{z}^T)^T,$$

where

$$\boldsymbol{z} = (u_1, \ldots, u_n)^T.$$

The missing variables $u_1, \ldots, u_n$ are defined so that

$$\boldsymbol{W}_j \mid u_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/u_j), \tag{2.40}$$

independently for $j = 1, \ldots, n$, and

$$U_1, \ldots, U_n \overset{\text{iid}}{\sim} \text{gamma}(\tfrac{1}{2}\nu, \tfrac{1}{2}\nu). \tag{2.41}$$

Here in this example, the missing data vector $\boldsymbol{z}$ consists of variables that would never be observable as data in the usual sense.

Because of the conditional structure of the complete-data model specified by (2.36) and (2.37), the complete-data likelihood function can be factored into the product of the conditional density of $\boldsymbol{W}$ given $\boldsymbol{z}$ and the marginal density of $\boldsymbol{Z}$. Accordingly, the complete-data log likelihood can be written as

$$\log L_c(\boldsymbol{\Psi}) = \log L_{1c}(\boldsymbol{\Psi}) + a(\boldsymbol{z}),$$

where

$$\begin{aligned} \log L_{1c}(\boldsymbol{\Psi}) &= -\tfrac{1}{2}np\log(2\pi) - \tfrac{1}{2}n\log|\boldsymbol{\Sigma}| \\ &\quad -\tfrac{1}{2}\sum_{j=1}^{n} u_j(\boldsymbol{w}_j - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}_j - \boldsymbol{\mu}) \end{aligned} \tag{2.42}$$

and

$$\begin{aligned} a(\boldsymbol{z}) &= -n\log\Gamma(\tfrac{1}{2}\nu) + \tfrac{1}{2}n\nu\log(\tfrac{1}{2}\nu) \\ &\quad +\tfrac{1}{2}\nu\sum_{j=1}^{n}(\log u_j - u_j) - \sum_{j=1}^{n}\log u_j. \end{aligned} \tag{2.43}$$

Now the E-step on the $(k+1)$th iteration of the EM algorithm requires the calculation of $Q(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k+1)})$, the current conditional expectation of the complete-data log likelihood function $\log L_c(\boldsymbol{\Psi})$. In the present case of known ν, we need focus only on the first term $\log L_{1c}(\boldsymbol{\Psi})$ in the expression (2.42) for $\log L_c(\boldsymbol{\Psi})$ (since the other term does not involve unknown parameters). As this term is linear in the unobservable data u_j, the E-step is effected simply by replacing u_j with its current conditional expectation given w_j.

Since the gamma distribution is the conjugate prior distribution for U, it is not difficult to show that the conditional distribution of U given $\boldsymbol{W} = \boldsymbol{w}$ is

$$U \mid \boldsymbol{w} \sim \text{gamma}(m_1,\ m_2), \tag{2.44}$$

where

$$m_1 = \tfrac{1}{2}(\nu + p)$$

and

$$m_2 = \tfrac{1}{2}\{\nu + \delta(\boldsymbol{w},\ \boldsymbol{\mu};\ \boldsymbol{\Sigma})\}. \tag{2.45}$$

From (2.44), we have that

$$E(U \mid \boldsymbol{w}) = \frac{\nu + p}{\nu + \delta(\boldsymbol{w},\ \boldsymbol{\mu};\ \boldsymbol{\Sigma})}. \tag{2.46}$$

Thus from (2.46),

$$E_{\boldsymbol{\Psi}^{(k)}}(U_j \mid \boldsymbol{w}_j) = u_j^{(k)},$$

where

$$u_j^{(k)} = \frac{\nu + p}{\nu + \delta(\boldsymbol{w}_j,\ \boldsymbol{\mu}^{(k)};\ \boldsymbol{\Sigma}^{(k)})}. \tag{2.47}$$

The M-step is easily implemented on noting that $L_{1c}(\boldsymbol{\Psi})$ corresponds to the likelihood function formed from n independent observations $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ with common mean $\boldsymbol{\mu}$ and

covariance matrices $\Sigma/u_1, \ldots, \Sigma/u_n$, respectively. After execution of the E-step, each u_j is replaced by $u_j^{(k)}$, and so the M-step is equivalent to computing the weighted sample mean and sample covariance matrix of $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ with weights $u_1^{(k)}, \ldots, u_n^{(k)}$. Hence

$$\boldsymbol{\mu}^{(k+1)} = \sum_{j=1}^{n} u_j^{(k)} \boldsymbol{w}_j / \sum_{j=1}^{n} u_j^{(k)} \tag{2.48}$$

and

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} u_j^{(k)} (\boldsymbol{w}_j - \boldsymbol{\mu}^{(k+1)})(\boldsymbol{w}_j - \boldsymbol{\mu}^{(k+1)})^T. \tag{2.49}$$

It can be seen in this case of known ν that the EM algorithm is equivalent to iteratively reweighted least squares. The E-step updates the weights $u_j^{(k)}$, while the M-step effectively chooses $\boldsymbol{\mu}^{(k+1)}$ and $\boldsymbol{\Sigma}^{(k+1)}$ by weighted least-squares estimation.

Liu and Rubin (1995) note that the above results can be easily extended to linear models where the mean $\boldsymbol{\mu}$ of $\boldsymbol{W}_j$ is replaced by $\boldsymbol{X}_j\boldsymbol{\beta}$, where $\boldsymbol{X}_j$ is a matrix containing the observed values of some covariates associated with $\boldsymbol{W}_j$.

In Sections 5.12.2 and 5.15, we discuss the replacement of the divisor n in (2.49) by $\sum_{j=1}^{k} u_j^{(k)}$ to improve the speed of convergence.

2.6.2 Numerical Example: Stack Loss Data

As a numerical example of the ML fitting of the t-distribution via the EM algorithm, we present an example from Lange et al. (1989), who analyzed the stack-loss data set of Brownlee (1965), which has been subjected to many robust analyses by various authors. Table 2.9 contains the data and Table 2.10 shows the slope of the regression of stack loss (y_j) on air flow (x_{1j}), temperature (x_{2j}), and acid (x_{3j}) for the linear regression model

$$y_j = \beta_0 + \sum_{i=1}^{3} \beta_i x_{ij} + e_j,$$

with t-distributed errors e_j for values of the degrees of freedom ν ranging from $\nu = 0.5$ to $\nu = \infty$ (normal). Also included in Table 2.10 are the values of four other estimators from Ruppert and Carroll (1980) of which two are trimmed least-squares ($\hat{\boldsymbol{\Psi}}_{KB}$, $\hat{\boldsymbol{\Psi}}_{PE}$), and two are the M-estimates of Huber (1964) and Andrews (1974). Also, the results are given for the ML estimate of ν, which was found to be $\hat{\nu} = 1.1$. The values of the log likelihood, $\log L(\hat{\boldsymbol{\Psi}})$, at the solutions are given in the second column for various values of ν. Twice the difference between the best fitting t and normal model log likelihoods is 5.44 which, on reference to the chi-squared distribution with one degree of freedom, suggests asymptotically a significant improvement in fit. As noted by Lange et al. (1989), the results of the fit for the ML case of $\hat{\nu} = 1.1$ are similar to those of Andrews (1974), which Ruppert and Carroll (1980) favored based on the closeness of the fit to the bulk of the data.

2.7 FINITE NORMAL MIXTURES

2.7.1 Example 2.7: Univariate Component Densities

We now extend the problem in Example 1.2 of Section 1.4.3 to the situation where the component densities in the mixture model (1.27) are not completely specified.