九州大学学術情報リポジトリ
Kyushu University Institutional Repository

ANALYSIS OF REPEATED-MEASURES DATA WITH OUTLIERS

Yamaguichi, Kazunori
Interdisciplinary Graduate SchooI of Engineering Sciences, Kyushu University

https://doi.org/10.5109/13411

出版情報：Bulletin of informatics and cybernetics. 24 (1/2), pp.71-80, 1990-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 24, No. 1 ~ 2, 1990

ANALYSIS OF REPEATED-MEASURES DATA WITH OUTLIERS

By

Kazunori YAMAGUCHI*

Abstract

The analysis of repeated-measures data is studied in the situation with some suspicious data which deviate relatively from other data and which might be generated by different mechanisms. If we assume the alternative error distributions which are heavy-tailed relative to the normal distribution, the estimates might be relatively unaffected by outliers. From this point of view, we assume scale mixtures of multivariate normal distributions as the error distributions, in order to reduce the influence of outliers. The case with missing observations is jointly considered and the method for computing the maximum likelihood estimates is given by applying the EM algorithm. Model selection and detection of outliers are also discussed with real data.

1. Introduction

Being engaged on statistical data analysis, we are often faced to missing data and suspicious data like outliers. On the other hand, most of classical methods of statistical analysis are based on the normal distribution assumptions. Therefore the results due to such methods are very sensitive to some outliers, and their reliability highly depends on the normality assumption.

Instead of the normal distribution, the assumption of alternative distribution relatively heavy-tailed, reflects interest in estimates which are relatively unaffected by extreme observations. Dempster *et al.* (1980) discussed the linear regression model with such error distributions named by the normal/independent distributions. Rubin (1983) gave the methods for robust estimation of mean vector and covariance matrix. Little (1988) extended Rubin's methods in the case with missing data.

In the present paper, the analysis of incomplete repeated-measures data is studied in the situation with outliers. Regarding this analysis, Jennrich and Schluchter (1986) discussed on analysis of the unbalanced repeated-measures data with the normality assumption. We now discuss the analysis under an assumption of the scale mixtures of multivariate normal distributions, instead of the normality assumption, in order to reduce the influence of extreme observations. These distributions keep symmetry and include contaminated multivariate normal distributions and the multivariate t distributions. Our discussion is confined to models with such distributions, according to

* Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Japan

Dempster *et al.* (1980)'s assertion that *within the class of symmetric error distributions, the limitation is not especially stringent in practice, because the class of normal/independent distributions is rich in long-tailed symmetric examples.*

For incomplete observations with missing values, the EM algorithm is available, as shown by Dempster *et al.* (1977), and is a general approach to iterative computation of maximum likelihood estimates when the observations can be regarded as incomplete data.

Notations and basic model are given in section 2. Even if interest is mainly to estimate regression parameters, efficiency of their estimates may be improved considerably by modeling the covariance matrix parsimoniously. Therefore we may also consider models with structured covariance matrices, where Jennrich and Schluchter (1986) listed several important structural models. In Section 3, we propose the methods for computing maximum likelihood estimates of the regression and covariance parameters, by applying the EM algorithm. As concerns the random-effects model, which is the special case of the structured covariance models, but we discuss it particularly.

For the analysis of repeated-measures data, our main interest may be whether the specified effects are zero or not. However, it may be impossible to apply the classical ANOVA method for this interest. Furthermore we need to obtain the confidence intervals of some specific means. Under such requirements, we derive the asymptotic variances of the regression parameters on the basis of the asymptotic normality of the maximum likelihood estimates. Finally in section 4, model selection and detection of outliers are discussed with real data.

2. The Model

We consider the situation where a fixed number T of measurements, corresponding to different times or experimental conditions, are to be collected on each of n subjects, but not all of the subjects' responses are observed, where we assume that missing is at random. Furthermore it is possible that data might include some extreme observations. We now assume the scale mixtures of multivariate normal distributions instead of the normality assumption in order to reduce the influence of extreme observations.

Let Y_i^* be a $T \times 1$ complete data vector for subject i, where $i = 1, \ldots, n$. The Y are assumed to follow the model

$$Y_i^* = X_i^* \beta + e_i^* \tag{1}$$

where X_i^* is a known matrix, β is a $p \times 1$ vector of unknown regression parameters, and the e_i^* are mutually independent. We assume that conditional on unobserved q_i, e_i is normally distributed with mean vector $\mathbf{0}$ and covariance matrix Σ/q_i, where q_i is a positive random variable with known probability (density) function $M(q_i)$.

But not all of the subjects' responses are not observed, then let Y_i be a $t_i \times 1$ vector containing the observed responses for subject i. According to (1), the Y_i follows the model $T_i = X_i\beta + e_i$, where X_i and e_i are a submatrix of X_i^* and a subvector of e_i^*, corresponding to the observed responses, respectively. Conditional on q_i, e_i is distributed as $N(\mathbf{0}, \Sigma_i/q_i)$, where Σ_i is a submatrix of Σ. We consider the estimation in

both cases unstructured Σ and structured Σ. In the structured case, we assume the elements of Σ are known functions of m unknown parameters contained in the vector θ. The regression parameter β vary independently of the covariance parameters θ. When we wish to emphasize that Σ depends on θ, we shall write Σ as $\Sigma(\theta)$.

3. Maximum Likelihood Estimates

We now discuss the algorithm for the maximum likelihood estimators, and, throughout this section, apply the EM algorithm to get ML estimators, treating q_i as missing data in addition to ordinary missing values.

3.1 Unstructured Σ

If all elements of Y_i^* and X_i^*, and q_i were observed, the likelihood is shown by

$$l = \prod_{i=1}^{n} M(q_i) \frac{1}{\sqrt{2\pi}^T} \exp\left(-\frac{1}{2} q_i e_i^{*\prime} \Sigma^{-1} e_i^*\right)$$

and the log-likelihood λ is

$$\begin{aligned} \lambda &= \log l \\ &= \text{Const.} - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} \{\operatorname{tr}(q_i \Sigma^{-1} e_i^* e_i^{*\prime})\} . \end{aligned} \tag{2}$$

Let e_i^* be partitioned into two subvectors for each subject,

$$e_i^* = \begin{pmatrix} e_i^{(1)} \\ e_i^{(2)} \end{pmatrix} \qquad i = 1, \ldots, n ,$$

where we assume $e_i^{(2)}$ is unobserved and $e_i^{(1)}$ is observed, then $e_i^{(1)} = Y_i - X_i\beta$, and also

$$\Sigma_i = \begin{pmatrix} \Sigma_{11i} & \Sigma_{12i} \\ \Sigma_{21i} & \Sigma_{22i} \end{pmatrix}$$

is similarly into submatrices. Then λ is

$$\begin{aligned} \lambda = \text{Const.} &- \frac{1}{2} \sum_{i=1}^{n} \log |\Sigma_{11i}| - \frac{1}{2} \sum_{i=1}^{n} q_i (Y_i - X_i\beta)' \Sigma_{11i}^{-1} (Y_i - X_i\beta) \\ &- \frac{1}{2} \sum_{i=1}^{n} \log |\tilde{\Sigma}_{22i}| \\ &- \frac{1}{2} \sum_{i=1}^{n} q_i \{e_i^{(2)} - \Sigma_{21i}\Sigma_{11i}^{-1}(Y_i - X_i\beta)\}' \tilde{\Sigma}_{22i}^{-1} \{e_i^{(2)} - \Sigma_{21i}\Sigma_{11i}^{-1}(Y_i - X_i\beta)\} \end{aligned} \tag{3}$$

where

$$\tilde{\Sigma}_{22i} = \Sigma_{22i} - \Sigma_{12i}\Sigma_{11i}^{-1}\Sigma_{12i} .$$

By (2), we get an estimate of Σ,

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} q_i e_i^* e_i^{*\prime} . \tag{4}$$

For the regression parameter vector β, differentiating (3) with respect to β, we have

$$\frac{\partial\lambda}{\partial\beta} = \sum_{i=1}^{n} q_i X_i' \Sigma_{11i}^{-1} e_i^{(1)} - \sum_{i=1}^{n} q_i X_i' \Sigma_{11i}^{-1} \Sigma_{12i} \tilde{\Sigma}_{22i}^{-1} \{e_i^{(2)} - \Sigma_{21i}\Sigma_{11i}^{-1} e_i^{(1)}\} .$$

Then, the ML estimator of β is obtained as $\partial\lambda/\partial\beta = 0$.

Since q_i and $e_i^{(2)}$ are not observed, we have to calculate the conditional expectations of the sufficient statistics given $e_i^{(1)}$. First we let

$$w_i = \mathrm{E}(q_i | e_i^{(1)}) , \tag{5}$$

where the specific form of w_i depends on the distribution of q_i, that is M (see examples.).

$$\mathrm{E}(q_i e_i^{(2)} | e_i^{(1)}) = \mathrm{E}\{q_i E(e_i^{(2)} | q_i, e_i^{(1)}) | e_i^{(1)}\} = w_i \hat{e}_i^{(2)} , \tag{6}$$

$$\mathrm{E}(q_i e_i^{(2)} e_i^{(2)\prime} | e_i^{(1)}) = \mathrm{E}\{q_i E(e_i^{(2)} e_i^{(2)\prime} | q_i, e_i^{(1)}) | e_i^{(1)}\} = w_i \hat{e}_i^{(2)} \hat{e}_i^{(2)\prime} + \tilde{\Sigma}_{22i} , \tag{7}$$

where

$$\hat{e}_i^{(2)} = \Sigma_{21i}\Sigma_{11i}^{-1} e_i^{(1)} .$$

Then

$$\mathrm{E}\left(\frac{\partial\lambda}{\partial\beta} | e_1^{(1)}, \ldots, e_n^{(1)}\right) = \sum_{i=1}^{n} w_i X_i' \Sigma_{11i}^{-1} e_i^{(1)} ,$$

therefore

$$\hat{\beta} = \left(\sum_{i=1}^{n} w_i X_i' \Sigma_{11i}^{-1} X_i\right)^{-1} \sum_{i=1}^{n} w_i X_i' \Sigma_{11i}^{-1} Y_i , \tag{8}$$

and from (4) and (7),

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \{w_i \hat{e}_i^* \hat{e}_i^{*\prime} + \tilde{\Sigma}_{22i}\} . \tag{9}$$

where

$$\hat{e}_i^* = \begin{pmatrix} e_i^{(1)} \\ \hat{e}_i^{(2)} \end{pmatrix} .$$

We can summarize the EM algorithm as follows;

E-step: to calculate the conditional expectations (5), (6) and (7).
M-step: to renew the estimates by (8) and (9).

EXAMPLE 1. Contaminated multivariate normal case
Let

$$\mathrm{M}(q_i) = \begin{cases} 1-\delta & \text{if } q_i = 1 \\ \delta & \text{if } q_i = \gamma \\ 0 & \text{otherwise} \end{cases} ,$$

then the marginal distribution of e_i is the contaminated multivariate normal distribution and

$$w_i = \frac{1 - \delta + \delta\gamma^{1+t_i/2} \exp\{(1-\gamma)d_i^2/2\}}{1 - \delta + \delta\gamma^{t_i/2} \exp\{(1-\gamma)d_i^2/2\}}$$

where

$$d_i^2 = (Y_i - X_i\beta)'\Sigma_{i11}^{-1}(Y_i - X_i\beta)\,.$$

EXAMPLE 2. Multivariate t case
When $q_i \times v$ has the chi-squared distribution with v degrees of freedom, the marginal distribution of e_i is the multivariate t distribution and

$$w_i = (v + t_i)/(v + d_i^2)\,.$$

(see Andrews *et al.* (1972) and Andrews and Mallows (1974) for another examples.)

3.2 Structured Σ

Jennrich and Schluchter (1986) discussed unbalanced repeated measures models with structured covariance matrices. We consider Jennrich and Schluchter's models in the case including some outliers. Instead of the normality assumption, we use the scale mixtures of multivariate normal distributions as the assumption of the error distributions. The model is $Y_i = X_i\beta + e_i$, where conditional on q_i, e_i is distributed as $\mathrm{N}(\mathbf{0}, \Sigma_i(\theta)/q_i)$.

Based on Jennrich and Schluchter's hybrid EM scoring algorithm, the algorithm for computing maximum likelihood estimators is easily obtained. The steps of the algorithm are in the following way.

(i) Compute some conditional expectations (5), (6) and (7)

(ii) Compute updated estimates $\hat{\beta}$ of β, using equation (8) and

$$S = \frac{1}{n}\sum_{i=1}^{n}\{w_i\hat{e}_i^*\hat{e}_i^{*\prime} + \tilde{\Sigma}_{22i}\}\,.$$

(iii) Compute the update θ by Jennrich and Schluchter's scoring step:

The update θ is $\theta + \Delta$, where $\Delta = H^{-1}G$,

$$[G]_i = \tfrac{1}{2}\,\mathrm{tr}\,\Sigma^{-1}(S - \Sigma)\Sigma^{-1}\dot{\Sigma}_i\,, \qquad [H]_{ij} = \tfrac{1}{2}\,\mathrm{tr}\,\Sigma^{-1}\dot{\Sigma}_i\Sigma^{-1}\dot{\Sigma}_j\,,$$

$$\dot{\Sigma}_i = \partial\Sigma(\theta)/\partial\theta_i\,.$$

3.3 Random-effects model

The random-effects model is $Y_i = X_i\beta + Z_ib_i + e_i$, where β is a $p \times 1$ vector of unknown population parameters, b_i is a $k \times 1$ vector of unknown individual effects and X_i and Z_i are a known $t_i \times p$ design matrix linking β to Y_i and a $t_i \times k$ one linking b_i to Y_i, respectively. Conditional on q_i, $e_i \sim \mathrm{N}(0, \sigma^2/q_iI_{t_i})$ and $b_i \sim \mathrm{N}(0, D/q_i)$, that is,

$$\begin{pmatrix} Y_i \\ b_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \frac{1}{q_i}\begin{pmatrix} Z_iDZ_i' + \sigma^2 I_{t_i} & Z_iD \\ DZ_i' & D \end{pmatrix}\right). \tag{10}$$

This model is a special case of the structured Σ such that $\Sigma_i = Z_i D Z_i' + \sigma^2 I_{t_i}$. But in this case treating b_i and q_i as missing data, we give the method based on the EM algorithm in order to obtain the maximum likelihood estimators.

If b_i and q_i were observed,

$$\hat{\beta} = \left(\sum_{i=1}^{n} q_i X_i' X_i \right)^{-1} \sum_{i=1}^{n} X_i'(Y_i - Z_i b_i) , \tag{11}$$

$$\hat{\sigma} = \frac{1}{\sum t_i} \sum_{i=1}^{n} q_i (Y_i - X_i \hat{\beta} - Z_i b_i)'(Y_i - X_i \hat{\beta} - Z_i b_i) \tag{12}$$

and

$$\hat{D} = \frac{1}{n} \sum_{i=1}^{n} q_i b_i b_i' . \tag{13}$$

In this case, we have to calculate the following conditional expectations given Y_i;

$$w_i = \mathrm{E}(q_i | Y_i) , \tag{14}$$

$$\mathrm{E}(q_i b_i | Y_i) = \mathrm{E}\{q_i \mathrm{E}(b_i | q_i, Y_i) | Y_i\} = w_i \hat{b}_i , \tag{15}$$

$$\mathrm{E}(q_i b_i b_i' | Y_i) = \mathrm{E}\{q_i \mathrm{E}(b_i b_i' | q_i, Y_i) | Y_i\} = w_i \hat{b}_i \hat{b}_i' + \tilde{D} , \tag{16}$$

and

$$\begin{aligned} \mathrm{E}(q_i e_i' e_i | Y_i) &= \mathrm{E}\{\mathrm{tr}\,(q_i e_i e_i') | Y_i\} \\ &= \mathrm{tr}\,[\mathrm{E}\{q_i \mathrm{E}(e_i e_i' | q_i, Y_i) | Y_i\}] \\ &= \mathrm{tr}\,\{w_i \hat{e}_i \hat{e}_i' + C\} , \end{aligned} \tag{17}$$

where

$$\hat{b}_i = D Z_i' (Z_i D Z_i + \sigma^2 I_{t_i})^{-1} (Y_i - X_i \beta) ,$$

$$\tilde{D} = D - D Z_i' (Z_i D Z_i' + \sigma^2 I_{t_i})^{-1} Z_i D ,$$

$$\hat{e} = Y_i - X_i \beta - Z_i \hat{b}_i , \qquad C = Z_i \tilde{D} Z_i' .$$

The EM algorithm can be summarized in the following way.

E-step: to calculate (14), (15), (16) and (17).
M-step: to renew the estimates by (11), (12) and (13).

3.4. Asymptotic variance of $\hat{\beta}$

We give the asymptotic variance of $\hat{\beta}$, which is used in the test for the regression parameters β and which is also used when computing the confidence intervals of some specific means. Louis (1982) devised a procedure to extract the observed information matrix when using the EM algorithm. Following Louis, we can obtain the observed information for the observed data. Let

$$\mathscr{I} = -\sum_{i=1}^{n} \mathrm{E}\left(\frac{\partial^2 \lambda_i}{\partial\beta\partial\beta'}\middle| e_i^{(1)}\right) - \sum_{i=1}^{n} \mathrm{E}\left\{\left(\frac{\partial\lambda_i}{\partial\beta}\right)\left(\frac{\partial\lambda_i}{\partial\beta}\right)'\middle| e_i^{(1)}\right\}$$
$$-\sum_{i\neq j} \mathrm{E}\left\{\left(\frac{\partial\lambda_i}{\partial\beta}\right)'\middle| e_i^{(1)}\right\}\mathrm{E}\left\{\left(\frac{\partial\lambda_j}{\partial\beta}\right)'\middle| e_j^{(1)}\right\}',$$

where

$$\lambda_i = -\tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}\operatorname{tr}(q_i\Sigma^{-1}e_i^* e_i^{*\prime}).$$

Then the asymptotic variance of $\hat{\beta}$ is given by A. Var $(\hat{\beta}) = \mathscr{I}^{-1}$. Now

$$\mathrm{E}\left(\frac{\partial^2\lambda_i}{\partial\beta\partial\beta'}\middle| e_i^{(1)}\right) = -w_i\{X_i'\Sigma_{11i}^{-1}X_i + X_i'\Sigma_{11i}^{-1}\Sigma_{12i}\tilde{\Sigma}_{22i}^{-1}\Sigma_{21i}\Sigma_{11i}^{-1}X_i\},$$

and

$$\mathrm{E}\left\{\left(\frac{\partial\lambda_i}{\partial\beta}\right)\left(\frac{\partial\lambda_i}{\partial\beta}\right)'\middle| e_i^{(1)}\right\} = w_i^* X_i'\Sigma_{11i}^{-1}e_i^{(1)}e_i^{(1)\prime}\Sigma_{11i}^{-1}X_i + w_i X_i'\Sigma_{11i}^{-1}\Sigma_{12i}\tilde{\Sigma}_{22i}^{-1}\Sigma_{21i}\Sigma_{11i}^{-1}X_i,$$

where $w_i^* = \mathrm{E}(q_i^2|e_i^{(1)})$. Thus

$$-\mathrm{E}\left(\frac{\partial^2\lambda_i}{\partial\beta\partial\beta'}\middle| e_i^{(1)}\right) - \mathrm{E}\left\{\left(\frac{\partial\lambda_i}{\partial\beta}\right)\left(\frac{\partial\lambda_i}{\partial\beta}\right)'\middle| e_i^{(1)}\right\} = w_i X_i'\Sigma_{11i}^{-1}X_i - w_i^* X_i'\Sigma_{11i}^{-1}e_i^{(1)}e_i^{(1)\prime}\Sigma_{11i}^{-1}X_i$$

and

$$\mathrm{E}\left(\frac{\partial\lambda_i}{\partial\beta}\right)'\middle| e_i^{(1)}\Big) = w_i X_i'\Sigma_{11i}^{-1}e_i^{(1)}$$

Therefore

$$\mathscr{I} = \sum_{i=1}^{n}\{w_i X_i'\Sigma_{11i}^{-1}X_i - w_i^* X_i'\Sigma_{11i}^{-1}e_i^{(1)}e_i^{(1)\prime}\Sigma_{11i}^{-1}X_i\} - \sum_{i\neq j} w_i w_j X_i'\Sigma_{11i}^{-1}e_i^{(1))}e_j^{(1)\prime}\Sigma_{11j}^{-1}X_j.$$

4. Numerical Examples

We analyzed Potthoff and Roy (1964)'s data, also analyzed by Jennrich and Schluchter (1986) under the normality assumption. Potthoff and Roy gave a set of growth data for 11 girls and 16 boys. None of the data are missing. Jennrich and Schluchter fitted eight different models, but our discussion was confined to the following models which appeared to fit the data in Jennrich and Schluchter's analysis. Let Y_{gst} denote the response for the s-th subject in sex group g, at time t, and X_t denote the t-th time ($g = 1, 2, t = 1, 2, 3, 4$). We applied the model for means as follows;

$$\mathrm{E}(Y_{gst}) = \alpha_g + \beta_g X_t,$$

and the unstructured model and the compound symmetry model as the model for Σ.

In calculations, we considered a method to have converged when all parameters differed by less than 0.01% between successive iterations. Table 4.1 shows the $-2 \times$ log-likelihoods under several assumptions and the numbers of parameters involved

Table 4.1. Summary of models fit

Distribution	Model for Σ	Number of parameters	$-2 \times$ log-likelihood
Normal	Unstructured	14	419.48
	Compound Symmetry	6	428.64
t distr.	Unstructured	15	414.90
$v = 12$	Compound Symmetry	7	421.25
Contam.	Unstructured	16	399.30
$\delta = 0.083$, $\gamma = 0.1$	Compound Symmetry	8	409.00

Table 4.2. Maximum Likelihood Estimates

Model			Estimated age/sex specific means (SE)		Estimated covariance matrix			
Distribution	Model for Σ	Age	Girls	Boys				
		8	21.24(.59)	22.45(.49)	5.119	2.441	3.611	2.522
Normal	Unstructured	10	22.19(.55)	24.11(.45)	2.441	3.928	2.718	3.062
		12	23.14(.57)	25.76(.47)	3.611	2.718	5.980	3.824
		14	24.09(.65)	27.42(.54)	2.522	3.062	3.824	4.618
		8	21.32(.54)	22.82(.49)	4.100	2.387	3.406	2.580
t-distr.	Unstructured	10	22.27(.52)	24.33(.47)	2.387	3.540	3.188	2.805
$v = 12$		12	23.21(.56)	25.84(.50)	3.406	3.188	5.114	3.844
		14	24.15(.63)	27.35(.57)	2.580	2.805	3.844	4.336
		8	21.26(.51)	23.33(.45)	3.756	2.636	3.716	2.883
Contam.	Unstructured	10	22.22(.51)	24.68(.45)	2.636	3.694	3.520	2.808
$\delta = 0.083$		12	23.18(.55)	26.02(.48)	3.716	3.520	5.157	3.992
$\gamma = 0.1$		14	24.13(.62)	27.36(.54)	2.883	2.808	3.992	4.364
		8	21.21(.63)	22.62(.52)	4.905	3.031	3.031	3.031
Normal	Compound symmetry	10	22.17(.57)	24.18(.47)	3.031	4.905	3.031	3.031
		12	23.13(.57)	25.75(.47)	3.031	3.031	4.905	3.031
		14	24.09(.63)	27.32(.52)	3.031	3.031	3.031	4.905
		8	21.24(.59)	22.69(.54)	4.170	2.937	2.937	2.937
t-distr.	Compound symmetry	10	22.18(.56)	24.15(.50)	2.937	4.170	2.937	2.937
$v = 12$		12	23.12(.56)	25.61(.50)	2.937	2.937	4.170	2.937
		14	24.05(.61)	27.07(.54)	2.937	2.937	2.937	4.170
		8	21.21(.60)	22.90(.53)	4.162	3.185	3.185	3.185
Contam.	Compound symmetry	10	22.17(.57)	24.29(.50)	3.185	4.162	3.185	3.185
$\delta = 0.083$		12	23.13(.57)	25.69(.50)	3.185	3.185	4.162	3.185
$\gamma = 0.1$		14	24.09(.60)	27.09(.53)	3.185	3.185	3.815	4.162

in the models, and the results of estimation are shown in Table 4.2. The standard errors were obtained from the results in Section 3.4. The value of degrees of freedom v for the multivariate t-distribution models was determined such that it maximized the log-likelihood of fixed model for mean vector and covariance matrix. On the other hand, for the contaminated normal model, δ and parameters for mean vector and covariance

Table 4.3. Conditional Expectations; w_i
Contaminated Multivariate Normal Distribution
(Σ: Compound Symmetry)

Girls	.997	.997	.992	.998	.998	.999	.999	.998	.998	.985
	.985									
Boys	.967	.997	.994	.971	.989	.999	.998	.964	.100	.979
	.995	.994	.100	.997	.989	.997				

Table 4.4. Result after omitting the 9-th and the 13-th boys data
(Σ: Compound Symmetry)

	Estimated age/sex specific means (SE)		Estimated covariance matrix			
Age	Girls	Boys				
8	21.21(.62)	22.95(.55)	4.476	3.516	3.516	3.516
10	22.17(.59)	24.32(.52)	3.516	4.476	3.516	3.516
12	23.13(.59)	25.70(.52)	3.516	3.516	4.476	3.516
14	24.09(.61)	27.07(.55)	3.516	3.516	3.516	4.476

matrix were estimated simultaneously for a grid of values of γ, and then the value of γ was chosen that maximized the log-likelihood of fixed model for means and covariance matrix. Strictly speaking, the estimates of parameters involved in M(.) are not ML estimates but these might be sufficient for our purpose. We perform a model selection using AIC, that is, we choose the model with the AIC less than AIC's of the others. In this case, the contaminated normal model with compound symmetry Σ was selected through AIC.

The conditional expectation w_i of q_i gives us the information of the degree of suspicion of outliers. Table 4.3 shows the values of w_i under the selected model. We wonder if the 9-th and the 13-th boys data are outliers based on this. Table 4.4 shows the result for the data without the 9-th and the 13-th boys data. This result is more similar to that under the selected model than to those under the other models.

References

[1] ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. M. and TUKEY, J. W.: *Robust Estimates of Location; Survey and Advances*, Princeton University Press., (1972).

[2] ANDREWS, D. F. and MALLOWS, C. L.: *Scale mixtures of normal distributions*, J. Roy. Statist. Soc., **B36** (1974), 99–102.

[3] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B.: *Maximum likelihood from incomplete data via the EM algorithm* (with Discussion, J. Roy. Statist. Soc., **B39** (1977), 1–38.

[4] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B.: *Iteratively reweighted least squares for linear regression when errors are normal/independent distributed*, Multivariate Analysis-V, (Krishnaiah, P. R. ed.) (1980), 35–57.

[5] JENNRICH, R. I. and SCHLUCHTER, M. D.: *Unbalanced repeated-measures models with structured covariance matrices*, Biometrics, **42** (1986), 805–820.

[6] LITTLE, R. J. A.: *Robust estimation of the mean and covariance matrix from data with missing values*, Appl. Statist., **37** (1988), 23–38.

[7] LOUIS, T. A.: *Finding the observed information matrix when using the EM algorithm*, J. Roy. Statist. Soc., **B44** (1982), 226–233.

[8] POTTHOFF, R. F. and ROY S. N.: *A generalized multivariate analysis of variance model useful especially for growth curve problems*, Biometrika, **51** (1964), 313–326.

[9] RUBIN, D. B: *Iteratively reweighted least squares*, Entry in Encyclopedia of the Statistical Sciences, (Kotz, S., Johnson, N. L. and Read, C. B., eds.), Wiley, **4** (1983).

Received August 1, 1989
Revised September 11, 1989
Communicated by Ch. Asano