STATISTICAL INFERENCE FOR NON-NORMAL POPULATIONS

山口, 和範
Interdisciplinary Graduate School of Engineering Sciences, Kyushu University

https://doi.org/10.11501/3054112

出版情報：九州大学, 1990, 理学博士, 課程博士
バージョン：
権利関係：

STATISTICAL INFERENCE FOR NON-NORMAL POPULATIONS

山口和範

STATISTICAL INFERENCE

FOR NON-NORMAL POPULATIONS

Kazunori Yamaguchi

The Colledge of Social Relations
Rikkyo University

Preface

To extract *useful* information from the real world, to manage it, and to utilize it in decision making are, I believe, the prime ends of the information science. Statistics, in this sense, is the most developed and established one among a wide variety of branches in the research field of information.

The modern statistics was founded in early twentieth century by K. Pearson, R.A. Fisher, J. Neyman, E.S. Pearson and many other applied scientists. The role of models in statistical and scientific work has attracted its importance even in those days and has now become generally recognized. The choice of model is, however, *aesthetic* and arbitrary, although based on knowledge and experience of the field of application. The beauty is just skin deep; we must not judge only by outward appearances, say, simplicity, lucidity and the case in representing with mathematical formulas, in obtaining the results of analysis and in interpreting them. The most essential feature of a model may be its helpfulness; by means of models, we clarify our thoughts and intentions towards the data or phenomena we are facing; through models, we strain the data and get the soupy essence of information we need; and, I am willing to admit that beautifulness strongly correlates helpfulness.

The normal distribution is, without doubt, one of the most *beautiful* entities in the world of statistics. In fact a great many statistical models together with the associate procedures have been established on this distribution since the days of our great pioneers mentioned before. Such procedures perform well under ideal conditions of normality but even under slight departures from the ideal they may lose their *helpfulness*. Experience with data has suggested that a proper degree of *robustness* is commonly desired.

There are two main approaches towards the robustness. One is to avoid restrictive assumptions about secondary aspects of a problem; which may be the practical motivation for consideration of non-parametric and semi-parametric models. The other

is to manage *outliers* in broad sense in order to get the results less affected by them; such as the use of trimmed means instead of the sample mean and outlier-rejection procedures. This article gives a detailed review of my own contributions to robust inference from both approaches.

I wish to express my sincere gratitude to Professor Chooichiro Asano for his constant instruction, extremely valuable suggestions and kind encouragement, throughout the period of my study at Kyushu University.

I am very grateful to Dr. Naoto Niki for his valuable comments and advice. I am greatly indebted to Dr. Michiko Watanabe as the coauthor of the papers concerning Sections 3 and 5. Thanks are also due to all my friends who participated in discussions ; among these, Dr. Zhi Geng and Mr. Atsuhiro Hayashi.

Finally, I would like to express my appreciation to the staff of the Interdisciplinary Graduate School of Engineering Sciences and the Research Institute of Fundamental Information Science for their support during my study at Kyushu University.

Kazunori Yamaguchi
The College of Social Relations
Rikkyo University, Japan 1990

Contents

1 Introduction

Much of the classical statistical methodology in multivariate analysis is concerned with models in which the underlying observation vectors are assumed to independently follow multivariate normal distributions. We, however, are often faced to the cases that the normality assumption is not appropriate for the data. There are some typical cases the normal model dose not appear to fit. One is the case all of observations follow the same distribution but it is not normal and the other important case is that most of observations follow the normal distribution but that a few observations are generated by a different mechanism.

Observations of failure time in survival analysis can be regarded as positive random variables with asymmetric distribution. If some suitable transformations for normality can be formed, *eg.* the Box-Cox transformation (Box and Cox 1964), we would apply the methods for the normal model. Otherwise, it is much more difficult to derive an inference method for obtaining optimal parameter values and to examine its properties under such distribution than under the normal distribution. In this case, non-parametric or semi-parametric methods are often applied to survival data, for the sake of robustness. Among these, Cox (1972)'s proportional hazard model may be one of the most useful semi-parametric model for survival data. As extension of Cox's to multivariate case, Yamaguchi (1986) considered several models for association in bivariate data with the proportional hazards assumption. Because of being semi-parametric, these models include some unspecified parts treated as nuisance functions or nuisance parameters and which forces us to construct conditional inferences for objective parameters (see Kalbfleisch and Sprott 1970 ; Kalbfleisch and Prentice 1973 ; Cox 1975).

Another non-normal case may happen when the data include some *outliers*. An intuitive definition of outliers would be "observations which deviate so much from another observations as to arouse suspicions that they are generated by a different

mechanism". In the context of the presence of outliers two main problems arise. One is the detection of outliers and the other is the construction of procedures which are not heavily affected by outliers, that is, robust procedures.

Many workers empirically have pointed out that the distribution of really observed error terms often have heavier-tail than that of the normal distribution. In particular, Jeffreys (1961) claimed that most error distributions could be approximated by the t distribution with 5, 6 or 7 degrees of freedom.

Difference in the tail parts, is so influential to the results of estimation that when the normality assumption is not appropriate for given data, the t model or mixture model is often applied instead of the normal error model. (see Kariya and Shinha 1989, for the robustness of statistical tests.) The scale mixtures of normal distributions are one of the most popular easily handled families of symmetric distributions with heavy-tails relative to the normal distribution, which are suitable as error distributions used instead of the normal when the data may include *outliers*. We might be faced to the problem of model selection, in deciding which distribution would be most appropriate. Then we shall use AIC, the remarkably useful tool for this problem.

Many statistical significant tests for detection of outliers were proposed. Barnett and Lewis (1978) listed 44 tests concerned with normal distribution. In this article, however, we do not use any tests to detect outliers. The detection of outliers shall be also treated as the problems concerned with model building and model selection.

The EM algorithm is introduced by Dempster *et al.* (1977) for computing maximum likelihood estimates from incomplete data. This EM algorithm is also available for pseudo incomplete data. When the error are assumed to follow a scale mixture of normals, we consider unobservable random variables in place of the unknown mixing parameters and treat the additional variables as missing values. One remarkable merit of use the EM algorithm is that we can handle ordinary missing values contained in the data simultaneously. Knowledge, or absence of knowledge, of the mechanisms that led to certain values being missing is a key element in choosing an appropriate analysis

method and in interpreting the results. In many cases, however, the mechanism leading to missing data does not enter explicitly; then, an assumption is being made that the mechanism is ignorable. We do not discuss the mechanism leading to missing data in detail and the missing values are assumed to be *missing at random* (Rubin 1976 ; Little and Rubin 1987) through this article.

This article includes five Sections. Section 2 gives a detailed review of work done in Yamaguchi (1988a). We consider several semi-parametric models in survival analysis with the proportional hazards assumption. Those models include some nuisance parts which force us to estimate objective parameters by means of partial likelihood (Cox, 1985; Wong, 1986). A test for the proportional hazards hypothesis is also proposed.

Section 3 deals with linear model with symmetric and heavy-tailed error distributions (Yamaguchi 1989). The scale mixtures of normals are used there as the error distributions, presenting a new method of maximum likelihood estimation through the contaminated normal error model (Yamaguchi 1990c). Model selection and detection of outliers are also considered.

In Section 4, the analysis of repeated measures data is studied in the situation we have some suspicious observations as *outliers* (Yamaguchi 1989a, Yamaguchi 1990b). For these data, we utilize the model given by Jennrich and Schluchter (1986) and the results derived in Section 3. Two numerical examples of real data are illustrated.

Section 5 gives a new robust method in factor analysis (Watanabe and Yamaguchi 1989, Yamaguchi and Watanabe 1990a) and demonstrates the robustness of these methods in comparison with usual normal factor analysis by a Monte Carlo simulation (Yamaguchi and Watanabe 1990b). Finally, model selection and detection of *outliers* in factor analysis are discussed with the data due to Mardia *et al.* (1979) as numerical example.

2 Model for Association in Bivariate Survival Data

The construction of bivariate or multivariate distribution has interested many statisticians from the early days. One approach is to extend a univariate distribution remaining its properties. The multivariate normal distribution is one of the most popular example. In survival analysis, we know the multivariate shock model introduced by Marshall and Olkin (1967), which is a generalization of the lack of memory property of the exponential distribution. In particular, in the bivariate case, Plackett (1965) considered to make a class of bivariate distributions with a parameter which was a measure of association, when two marginal distributions were given.

In survival analysis, especially, analysis of the survival times of fatal patients, its main purpose is to evaluate variation of risk with respect to time and real-time control rather than prediction of survival time. From this point of view, modeling on the hazard function stands on actual meanings and needs. We also construct a model for association in bivariate survival data on the hazard function.

When we observe pairs of failure times, one of our main interests is whether one failure affects the other failure time. This question can be answered by hazard regression model analysis (Cox, 1972, 1975), regarding one failure times as time-dependent covariates which possibly influence the hazard for the other failures. For another approach, Clayton (1978) introduced a model for association in bivariate survival data with a parameter which has a simple interpretation. Let s and t be two survival times and $f(s,t)$ be its joint probability density function. Then Clayton's model satisfies

$$f(s,t)\int_s^\infty\int_t^\infty f(u,v)dudv = \theta\int_s^\infty f(u,t)du\int_t^\infty f(s,v)dv.$$

Representing by the hazard function, for any s_0, t_0 (>0),

$$\frac{h_s(s_0 \mid t=t_0)}{h_s(s_0 \mid t>t_0)} = h_t(t_0 \mid s=s_0)h_t(t_0 \mid s>s_0) = \theta,$$

where $h(.\mid .)$ denotes the conditional hazard function. Clayton applied this model to familial studies of disease incidence. Parametric and semi-parametric estimation

within this model is discussed by Clayton (1978) and Oakes (1982), although a fully satisfactory non-parametric procedure has not yet been found (Cox and Oakes, 1984).

Now we consider the following case. A pair of individuals are exposed to risk at the same time and they always affect one another. In this case, a failure of the partner may change the hazard function of the remainder, without delay. Reflecting such situation, we construct models with parameters which are interpreted as same concept of odds ratio of contingency tables.

2.1 Bivariate survivor function

In the case of univariate, if one of the probability density function, cumulative distribution function, survivor function, and hazard function is uniquely determined the other three functions are automatically determined. Now such relationship in bivariate case is shown.

Let $T^{(1)}$, $T^{(2)}$ be two positive random variable. We define the following four hazard functions.

$$h^{(i)}(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(T^{(i)} < t + \Delta \mid T^{(1)} \geq t, T^{(2)} \geq t),$$

$$h^{*(i)}(t \mid t') = \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(T^{(i)} < t + \Delta \mid T^{(i)} \geq t, T^{(j)} = t'),$$

where

$$t \geq t', i = 1, 2, i + j = 3.$$

These four functions determine the joint distribution of $T^{(1)}$, $T^{(2)}$, if these are continuous. The joint probability density function is, for $t^{(1)} \leq t^{(2)}$,

$$\begin{aligned} f(t^{(1)}, t^{(2)}) &= \exp\{-\int_0^{t^{(1)}} [h^{(1)}(t) + h^{(2)}(t)]dt - \int_{t^{(1)}}^{t^{(2)}} h^{*(2)}(t \mid t^{(1)})dt\} \\ &\quad \times h^{(1)}(t^{(1)}) h^{*(2)}(t^{(2)} \mid t^{(1)}), \end{aligned} \tag{1}$$

for $t^{(1)} \geq t^{(2)}$,

$$\begin{aligned} f(t^{(1)}, t^{(2)}) &= \exp\{-\int_0^{t^{(2)}} [h^{(1)}(t) + h^{(2)}(t)]dt - \int_{t^{(2)}}^{t^{(1)}} h^{*(1)}(t \mid t^{(2)})dt\} \\ &\quad \times h^{(2)}(t^{(2)}) h^{*(1)}(t^{(1)} \mid t^{(2)}). \end{aligned} \tag{2}$$

Clearly, $T^{(1)}$ is statistical independent of $T^{(2)}$, if and only if, for any t, $t' > 0$,

$$\begin{aligned} h^{*(1)}(t \mid t') &= h^{(1)}(t) \\ h^{*(2)}(t \mid t') &= h^{(2)}(t). \end{aligned}$$

Let $S(t^{(1)}, t^{(2)})$ denote the joint survivor function, then

$$h^{(i)}(t) = -\frac{1}{S(t,t)} \frac{\partial S(t^{(1)}, t^{(2)})}{\partial t^{(i)}} \mid_{t^{(1)}=t, t^{(2)}=t},$$

$$h^{*(i)}(t \mid t') = -\frac{\frac{\partial^2 S(t^{(1)}, t^{(2)})}{\partial t^{(1)} \partial t^{(2)}} \mid_{t^{(i)}=t, t^{(j)}=t'}}{\frac{\partial S(t^{(1)}, t^{(2)})}{\partial t^{(i)}} \mid_{t^{(i)}=t, t^{(j)}=t'}}.$$

2.2 Model for association

The purpose of analysis of two dimensional contingency table is to know the relationship between two qualitative variables. Measures of association are numerical assessments of the strength of the statistical dependence of two qualitative variables.

While we consider a model for association in bivariate failure times, at first we start from the standpoint of contingency tables. Divide a interval $[0, \infty)$ into q subintervals such that

$$I_j = [t_j, t_{j+1}), \quad j = 1, 2, ..., q,$$

where

$$0 = t_1 < t_2 < \cdots < t_{q+1} = \infty.$$

We get a $q \times q$ contingency table if we collect data in each subinterval I_j. Clayton (1974) gave a method for analysis of two ordered categorical variables. The method assumes that for each of the $(q-1) \times (q-1)$ possible ways of collapsing the table into a 2×2 table the odds ratio is constant. Under this assumption, the inference about the constant odds ratio is performed. When the odds ratio is unit, this model is independent model. The unconditional maximum likelihood, conditional (partial) likelihood and Mantel-Haenszel estimators have been studied as estimator of common

odds ratio of several 2×2 tables. Breslow (1981) gave the properties of these estimators under a large sample scheme in which the number of tables increases but the possible marginal configurations remain fixed. In this situation, the unconditional maximum likelihood estimator is not consistent. The asymptotic relative efficiency of the Mantel-Haenszel and conditional likelihood estimators was given by Hauck (1988). Hauck *et al.* (1982) examined small sample properties of these estimators and, Yamaguchi (1988) derived a extended Mantel-Haenszel estimator and demonstrated the properties of these estimators.

We are analyzing continuous variables. The above method, of course, can be applied by dividing the interval $[0, \infty)$ into some subintervals, but there is loss of information and the results depend on the way of construction of subintervals. We consider the limitation of the above model such that all lengths of subintervals tend to 0, as a model for continuous variables. Let

$$p_{uv} = \Pr(T^{(1)} \in I_u, T^{(2)} \in I_v), \quad u, v = 1, 2, ..., q.$$

Define $(q-1)$ odds ratios for each of $T^{(1)}$ and $T^{(2)}$, respectively, as follows,

$$\theta_i^{(1)} = \frac{\sum_{v=1}^{i} p_{iv} \sum_{u=i+1}^{q} \sum_{v=i+1}^{q} p_{uv}}{\sum_{v=i+1}^{q} p_{iv} \sum_{u=i+1}^{q} \sum_{v=1}^{i} p_{uv}},$$

$$\theta_i^{(2)} = \frac{\sum_{u=1}^{i} p_{ui} \sum_{u=i+1}^{q} \sum_{v=i+1}^{q} p_{uv}}{\sum_{u=i+1}^{q} p_{ui} \sum_{u=1}^{i} \sum_{v=i+1}^{q} p_{uv}},$$

$$i = 1, 2, ..., q-1.$$

When all lengths of subintervals tend to zero, the limits of these odds ratios are the ratios of hazard functions. Let

$$Q_{11i} = \sum_{v=1}^{i} p_{iv},$$

$$Q_{12i} = \sum_{u=i+1}^{q} \sum_{v=1}^{i} p_{uv},$$

$$Q_{21i} = \sum_{v=i+1}^{q} p_{iv},$$

and

$$Q_{22i} = \sum_{u=i+1}^{q} \sum_{v=i+1}^{q} p_{uv},$$

then

$$Q_{11i} = S(t_i, 0) - S(t_{i+1}, 0) - S(t_i, t_{i+1}) + S(t_{i+1}, t_{i+1}),$$

$$Q_{12i} = S(t_i, 0) - S(t_{i+1}, t_{i+1}),$$

$$Q_{21i} = S(t_i, t_{i+1}) - S(t_{i+1}, t_{i+1}),$$

and

$$Q_{22i} = S(t_{i+1}, t_{i+1}).$$

Now for fixed t_i, when $\Delta_i = t_{i+1} - t_i$ tends to zero,

$$\frac{Q_{11i}}{\Delta_i} \longrightarrow -\frac{\partial}{\partial t^{(1)}} S(t_i, 0) + \frac{\partial}{\partial t^{(1)}} S(t_i, t_i),$$

$$Q_{12i} \longrightarrow S(t_i, 0) - S(t_i, t_i),$$

$$\frac{Q_{21i}}{\Delta_i} \longrightarrow -\frac{\partial}{\partial t^{(1)}} S(t_i, t_i),$$

and

$$Q_{22i} \longrightarrow S(t_i, t_i).$$

For the odds ratio,

$$\begin{aligned} \theta_i^{(1)} &= \frac{Q_{11i} Q_{22i}}{Q_{12i} Q_{21i}} \\ &\longrightarrow \frac{\{-\frac{\partial}{\partial t^{(1)}} S(t_i, 0) + \frac{\partial}{\partial t^{(1)}} S(t_i, t_i)\} S(t_i, t_i)}{-\frac{\partial}{\partial t^{(1)}} S(t_i, t_i)\{S(t_i, 0) - S(t_i, t_i)\}}. \end{aligned}$$

Furthermore, let

$$h^{**(i)}(t \mid T^{(j)} < t) = \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(T^{(i)} < t + \Delta \mid T^{(i)} \geq t, T^{(j)} < t),$$

$$i = 1, 2, i + j = 3,$$

then

$$h^{**(1)}(t \mid T^{(2)} < t) = \frac{-\frac{\partial}{\partial t^{(1)}} S(t, 0) + \frac{\partial}{\partial t^{(1)}} S(t, t)}{S(t, 0) - S(t, t)}.$$

Therefore

$$\begin{aligned}\theta_i^{(1)} &= \frac{Q_{11i}Q_{22i}}{Q_{12i}Q_{21i}} \\ &\longrightarrow \frac{h^{**(1)}(t_i \mid T^{(2)} < t_i)}{h^{(1)}(t_i)}.\end{aligned}$$

Similarly,

$$\theta_i^{(2)} \longrightarrow \frac{h^{**(2)}(t_i \mid T^{(1)} < t_i)}{h^{(2)}(t_i)}.$$

As $\theta_i^{(j)}$ is constant for any i, we obtain the model for continuous variables such that

$$h^{**(i)}(t \mid T^{(j)} < t) = \theta^{(i)}h^{(i)}(t), \quad i = 1, 2, i + j = 3.$$

However, as it may be more convenient that a bivariate distribution is represented by $h^{(i)}$ and $h^{*(i)}$ than by $h^{(i)}$ and $h^{**(i)}$, we would like to use $h^{(i)}$ and $h^{*(i)}$ for our model.

$$\begin{aligned}h^{**(i)}(t \mid T^{(j)} < t) &= \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(T^{(i)} < t + \Delta \mid T^{(i)} \geq t, T^{(j)} < t) \\ &= \frac{\lim_{\Delta \to 0} \frac{1}{\Delta} \int_0^t \Pr(t \leq T^{(i)} < t + \Delta, T^{(j)} = t')dt'}{\int_0^t \Pr(t \leq T^{(i)}, T^{(j)} = t')dt'} \\ &= \frac{\int_0^t \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(t \leq T^{(i)} < t + \Delta, T^{(j)} = t')dt'}{\int_0^t \Pr(t \leq T^{(i)}, T^{(j)} = t')dt'} \\ &= \frac{\int_0^t h^{*(i)}(t \mid t') \Pr(t \leq T^{(i)}, T^{(j)} = t')dt'}{\int_0^t \Pr(t \leq T^{(i)}, T^{(j)} = t')dt'}.\end{aligned}$$

These equations yield that if we set for any t and $t' > 0$

$$h^{*(i)}(t \mid t') = \theta^{(i)}h^{(i)}(t), \quad i = 1, 2. \tag{3}$$

then

$$h^{**(i)}(t \mid T^{(j)} < t) = \theta^{(i)}h^{(i)}(t), \quad i = 1, 2, i + j = 3.$$

Thus we consider (3) as basic model for association. Clearly

$$\theta^{(1)} = \theta^{(2)} = 1$$

if and only if $T^{(1)}$ and $T^{(2)}$ are mutually statistically independent. When both of $\theta^{(1)}$ and $\theta^{(2)}$ are greater than one, there exists positive association, and when both of them are less than one, there exists negative association.

We add some restrictions about to the model (3) in order to give simple interpretations. At first we consider the case that the bivariate distribution is symmetric. That is, for any $t > 0$

$$h^{(1)}(t) = h^{(2)}(t), \quad \theta^{(1)} = \theta^{(2)}.$$

Model 1:

$$\begin{cases} h^{(1)}(t) = h(t) \\ h^{*(1)}(t \mid t') = \theta h(t) \\ h^{(2)}(t) = h(t) \\ h^{*(2)}(t \mid t') = \theta h(t), \end{cases}$$

where θ is any positive constant and $h(t)$ is any integrable positive function.

In this case, from the equations (1), (2), the joint probability function $f(t^{(1)}, t^{(2)})$ is for $t^{(1)} \leq t^{(2)}$

$$\begin{aligned} f(t^{(1)}, t^{(2)}) &= \exp\{-\int_0^{t^{(1)}} 2h(u)du - \int_{t^{(1)}}^{t^{(2)}} \theta h(u)du\}\theta h(t^{(1)})h(t^{(2)}) \\ &= \theta S(t^{(1)})^{1-\theta} S(t^{(2)})^{\theta-1} f(t^{(1)}) f(t^{(2)}), \end{aligned}$$

for $t^{(1)} \geq t^{(2)}$

$$\begin{aligned} f(t^{(1)}, t^{(2)}) &= \exp\{-\int_0^{t^{(2)}} 2h(u)du - \int_{t^{(2)}}^{t^{(1)}} \theta h(u)du\}\theta h(t^{(1)})h(t^{(2)}) \\ &= \theta S(t^{(1)})^{\theta-1} S(t^{(2)})^{1-\theta} f(t^{(1)}) f(t^{(2)}), \end{aligned}$$

where

$$S(t) = \int_0^t h(u)du \text{ and } f(t) = -\frac{dS(t)}{dt}.$$

Let

$$X = \min\{T^{(1)}, T^{(2)}\}, Y = \max\{T^{(1)}, T^{(2)}\},$$

and $f_{X,Y}$, f_X and f_Y be the joint probability density function of X and Y, the marginal density function of X and that of Y, respectively, then

$$f_{X,Y}(x, y) = 2\theta S(x)^{1-\theta} S(y)^{\theta-1} f(x) f(y),$$

$$f_X(x) = 2\theta S(x) f(x), \tag{4}$$

and

$$f_Y(y) = \frac{2\theta}{2-\theta}\{S(y)^{\theta-1} - S(y)\} f(y).$$

Next in the case of asymmetric case, we assume the proportionality of two marginal hazard functions, that is

$$h^{(1)}(t) = h(t),$$

$$h^{(2)}(t) = \alpha h(t).$$

Model 2:

$$\begin{cases} h^{(1)}(t) = h(t) \\ h^{*(1)}(t \mid t') = \theta^{(1)} h(t) \\ h^{(2)}(t) = \alpha h(t) \\ h^{*(2)}(t \mid t') = \alpha\theta^{(2)} h(t), \end{cases}$$

In this case, the equations (1) and (2) yield the joint probability function $f(t^{(1)}, t^{(2)})$, for $t^{(1)} \leq t^{(2)}$

$$\begin{aligned} f(t^{(1)}, t^{(2)}) &= \exp\{-\int_0^{t^{(1)}} (1+\alpha)h(u)du - \int_{t^{(1)}}^{t^{(2)}} \alpha\theta^{(2)} h(u)du\} \\ &\quad \times \alpha\theta^{(2)} h(t^{(1)}) h(t^{(2)}) \\ &= \alpha\theta^{(2)} S(t^{(1)})^{\alpha - \alpha\theta^{(2)}} S(t^{(2)})^{\alpha\theta^{(2)}-1} f(t^{(1)}) f(t^{(2)}), \end{aligned}$$

for $t^{(1)} \geq t^{(2)}$

$$\begin{aligned} f(t^{(1)}, t^{(2)}) &= \exp\{-\int_0^{t^{(2)}} (1+\alpha)h(u)du - \int_{t^{(2)}}^{t^{(1)}} \alpha\theta^{(1)} h(u)du\} \\ &\quad \times \alpha\theta^{(1)} h(t^{(1)}) h(t^{(2)}) \\ &= \alpha\theta^{(1)} S(t^{(1)})^{\theta^{(1)}-1} S(t^{(2)})^{\alpha - \theta^{(1)}} f(t^{(1)}) f(t^{(2)}). \end{aligned}$$

Similarly,

$$\begin{aligned} f_{X,Y}(x,y) &= (1+\alpha) f(x) f(y) \\ &\quad \times \{\theta^{(1)} S(x)^{\alpha - \theta^{(1)}} S(y)^{\theta^{(1)}-1}\}^{1-\delta} \\ &\quad \times \{\alpha\theta^{(2)} S(x)^{\alpha - \alpha\theta^{(2)}} S(y)^{\alpha\theta^{(2)}-1}\}^{\delta} \end{aligned}$$

$$
\begin{aligned}
f_X(x) &= (1+\alpha)S(x)^{\alpha} f(x), \\
f_Y(y) &= (1+\alpha)f(y) \\
&\quad \times [\frac{\theta^{(1)}}{1+\alpha-\theta^{(1)}}\{S(y)^{\theta^{(1)}-1} - S(y)^{\alpha}\}]^{1-\delta} \\
&\quad \times [\frac{\alpha\theta^{(2)}}{1+\alpha-\alpha\theta^{(2)}}\{S(y)^{\alpha\theta^{(2)}-1} - S(y)^{\alpha}\}]^{\delta},
\end{aligned}
$$

where

$$
\delta = \begin{cases} 0 & \text{if } T^{(1)} > T^{(2)} \\ 1 & \text{if } T^{(1)} < T^{(2)} \end{cases} .
$$

2.3 Estimation of parameters

As there is a nuisance function $h(t)$ in our models, the method for estimation is based on the partial likelihood (Cox, 1975), which is generalization of conditional and marginal likelihood. Let the data $\{(t_i^{(1)}, t_i^{(2)}), i = 1, 2, ..., n\}$ be arranged in ascending order of failure time as follows,

$$
t_{(1)} < t_{(2)} < \cdots < t_{(2n)},
$$

and V_j and W_j be the following events,

V_j : The individual (j) fails at time t_j.

W_j : There is no failure in an interval $[t_{j-1}, t_j)$ and an individual fails at time t_j.

The probability of the event $\{(V_j, W_j), j = 1, 2, ..., 2n\}$ is

$$
\prod_{j=1}^{2n} \Pr(W_j \mid W^{(j-1)}, V^{(j-1)}) \Pr(V_j \mid W^{(j)}, V^{(j-1)}),
$$

where

$$
W^{(j)} = \{W_1, W_2, \cdots, W_j\},
$$

$$
V^{(j)} = \{V_1, V_2, \cdots, V_j\},
$$

and the partial likelihood L_p is

$$\begin{aligned} L_p &= \prod_{j=1}^{2n} \Pr(V_j \mid W^{(j)}, V^{(j-1)}) \\ &= \prod_{j=1}^{2n} \frac{h_{(j)}(t_{(j)})}{\sum_{l \in R_j} h_l(t_{(l)})}. \end{aligned} \tag{5}$$

Data can be also represented as Table 1 at the failure times.

Table 1 : Data at the failure time $t_{(j)}$

	Partner	Died	Survival	Total
Type (1)	Survival	$d_j^{(1s)}$	$r_j^{(1s)} - d_j^{(1s)}$	$r_j^{(1s)}$
	Died	$d_j^{(1d)}$	$r_j^{(1d)} - d_j^{(1d)}$	$r_j^{(1d)}$
Type (2)	Survival	$d_j^{(2s)}$	$r_j^{(2s)} - d_j^{(2s)}$	$r_j^{(2s)}$
	Died	$d_j^{(2d)}$	$r_j^{(2d)} - d_j^{(2d)}$	$r_j^{(2d)}$
Total		1	$n_j - 1$	n_j

2.3.1 Model 1

By (5), the partial likelihood L_p based on Model 1 is

$$L_p = \prod_{j=1}^{2n} \frac{1^{d_j^{(1s)}+d_j^{(2s)}} \theta^{d_j^{(1d)}+d_j^{(2d)}}}{r_j^{(1s)} + r_j^{(2s)} + \theta(r_j^{(1d)} + r_j^{(2d)})}.$$

Then the log partial likelihood ℓ_p is

$$\begin{aligned} \ell_p &= \log L \\ &= \sum_{j=1}^{n} [(d_j^{(1d)} + d_j^{(2d)}) \log \theta - \log\{r_j^{(1s)} + r_j^{(2s)} + \theta(r_j^{(1d)} + r_j^{(2d)})\}]. \end{aligned}$$

The solution of $\partial \ell_p / \partial \theta = 0$, $\widehat{\theta_p}$ satisfies the following equation,

$$\widehat{\theta_p} = \frac{\sum_{j=1}^{2n} (d_j^{(1d)} + d_j^{(2d)})}{\sum_{j=1}^{2n} \frac{r_j^{(1d)} + r_j^{(2d)}}{r_j^{(1s)} + r_j^{(2s)} + \widehat{\theta_p}(r_j^{(1d)} + r_j^{(2d)})}}. \tag{6}$$

When the observations $\{(x_i, y_i), i = 1, 2, ..., n\}$ are given, the full likelihood L is,

$$L = \prod_{i=1}^{n} \{2\theta S(x_i)^{1-\theta} S(y_i)^{\theta-1} f(x_i) f(y_i)\}.$$

Here let

$$p_i = H(x_i) \text{ and } q_i = H(y_i),$$

where $H(t)$ is the cumulative hazard function, that is

$$H(t) = \int_0^t h(u)du.$$

Then the log likelihood ℓ is

$$\begin{aligned} \ell &= \log L \\ &= \sum_{i=1}^{n} \{\log 2 + \log \theta - (1-\theta)p_i - (\theta - 1)q_i\}. \end{aligned}$$

This yields

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} (\frac{1}{\theta} + p_i - q_i),$$

and

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} (q_i - p_i)}.$$

For the estimation of $H(t)$, we can use the fact that $\{x_i, i = 1, 2, ..., n\}$ is a sample with size n from the distribution with the hazard function $2h(x)$, that is, we use the sample cumulative hazard function as the estimate of $H(x)$,

$$H_n(t) = \sum_{(t)} \frac{1}{2(n+1-i)},$$

where $\sum_{(t)}$ denotes sum over i, $x_{(i)} \leq t$, and $\{x_{(i)}\}$ is the ordered statistics of $\{x_i\}$.

Theorem 1

$$\widehat{\theta}_c = \frac{n}{\sum_{i=1}^{n} \{H_n(y_i) - H_n(x_i)\}}$$

is a consistent estimate of θ.

Proof

$$\begin{aligned}\frac{1}{\widehat{\theta_c}} &= \frac{\sum_{i=1}^{n}\{H_n(y_i) - H_n(x_i)\}}{n} \\ &= \int H_n(t)d\{F_n(t) - G_n(t)\} \\ &\longrightarrow \int H(t)d\{F(t) - G(t)\},\end{aligned}$$

where

$$F_n(t) = \#\{i; x_i > t\}/n,$$

$$G_n(t) = \#\{i; y_i > t\}/n,$$

and F, G are the survivor functions of X, Y, respectively.

By (4),

$$\begin{aligned}\int H(t)d\{F(t) - G(t)\} &= \int H(t)\frac{2\theta}{2-\theta}\{S(t)^{\theta-1} - S(t)\}f(t)dt - \int H(t)2\theta S(t)f(t)dt \\ &= \frac{2+\theta}{2\theta} - \frac{1}{2} \\ &= \frac{1}{\theta}.\end{aligned}$$

As we derived a consistent estimator $\widehat{\theta_c}$ of θ, we would like to derive a two-step estimator $\widehat{\theta_2}$ from the equation (6), which is asymptotically equivalent to the maximum partial likelihood estimator.

$$\widehat{\theta_2} = \frac{\sum_{j=1}^{2n}(d_j^{(1d)} + d_j^{(2d)})}{\sum_{j=1}^{2n}\frac{r_j^{(1d)}+r_j^{(2d)}}{r_j^{(1s)}+r_j^{(2s)}+\widehat{\theta_c}(r_j^{(1d)}+r_j^{(2d)})}}.$$

2.3.2 Model 2

In the case of Model 2, by (5), the partial likelihood L_p is

$$L_p = \prod_{j=1}^{2n}\frac{1^{d_j^{(1s)}}\theta^{(1)d_j^{(1d)}}\alpha^{d_j^{(2s)}}(\alpha\theta^{(2)})^{d_j^{(2d)}}}{r_j^{(1s)} + \theta^{(1)}r_j^{(1d)} + \alpha r_j^{(2s)} + \alpha\theta^{(2)}r_j^{(2d)}},$$

and

$$\begin{aligned}\ell_p &= \log L_p \\ &= \sum_{j=1}^{2n}\{d_j^{(1d)}\log\theta^{(1)} + d_j^{(2s)}\log\alpha + d_j^{(2d)}\log(\alpha\theta^{(2)}) \\ &\quad -\log(r_j^{(1s)} + \theta^{(1)}r_j^{(1d)} + \alpha r_j^{(2s)} + \alpha\theta^{(2)}r_j^{(2d)})\}.\end{aligned}$$

From this

$$\begin{aligned}\frac{\partial\ell}{\partial\theta^{(1)}} &= \sum_{j=1}^{2n}\{\frac{d_j^{(1d)}}{\theta^{(1)}} - \frac{r_j^{(1d)}}{Q_j}\}, \\ \frac{\partial\ell}{\partial\theta^{(2)}} &= \sum_{j=1}^{2n}\{\frac{d_j^{(2d)}}{\theta^{(2)}} - \frac{r_j^{(2d)}}{Q_j}\}, \\ \frac{\partial\ell}{\partial\alpha} &= \sum_{j=1}^{2n}\{\frac{d_j^{(2s)} + d_j^{(2d)}}{\alpha} - \frac{r_j^{(2d)} + \theta^{(2)}r_j^{(2d)}}{Q_j}\},\end{aligned} \tag{7}$$

where

$$Q_j = r_j^{(1s)} + \theta^{(1)}r_j^{(1d)} + \alpha r_j^{(2s)} + \alpha\theta^{(2)}r_j^{(2d)}.$$

The maximum partial likelihood estimates are obtained as the solutions that the above three equations simultaneously equal to zero.

Now we would like to derive consistent estimators in the same way of Model 1. When the observations $\{(x_i, y_i, \delta_i), i = 1, 2, ..., n\}$ are given, the full likelihood L is,

$$\begin{aligned}L &= \prod_{i=1}^{n}[(1+\alpha)f(x_i)f(y_i)\{\theta^{(1)}S(x_i)^{\alpha-\theta^{(1)}}S(y_i)^{\theta^{(1)}-1}\}^{1-\delta_i} \\ &\quad \{\alpha\theta^{(2)}S(x_i)^{\alpha-\alpha\theta^{(2)}}S(y_i)^{\alpha\theta^{(2)}-1}\}^{\delta_i}]\end{aligned}$$

Here let

$$p_i = H(x_i) \text{ and } q_i = H(y_i),$$

then the log likelihood ℓ is

$$\begin{aligned}\ell &= \sum_{i=1}^{n}[(1-\delta_i)\{\log\alpha + \log\theta^{(1)} - (\alpha-\theta^{(1)})p_i - (\theta^{(1)}-1)q_i\} \\ &\quad +\delta_i\{\log\alpha + \log\theta^{(2)} - (\alpha-\alpha\theta^{(2)})p_i - (\alpha\theta^{(2)}-1)q_i\} - p_i - q_i],\end{aligned}$$

and, the ML estimates are

$$\widehat{\theta^{(1)}} = \frac{n^{(0)}}{\sum_{i=1;\delta=0}^{n}(q_i - p_i)},$$

$$\widehat{\theta^{(2)}} = \frac{n^{(1)}\sum_{i=1}^{n} p_i}{n^{(0)}\sum_{i=1;\delta=1}^{n}(q_i - p_i)},$$

$$\widehat{\alpha} = \frac{n^{(0)}}{\sum_{i=1}^{n} p_i}.$$

In this case, for the estimation of $H(t)$, we can use the fact that $\{x_i; \delta = 1\}$ is a sample with size $n^{(1)}$ from the distribution with the hazard function $h(x)$, that is, we use the sample cumulative hazard function as the estimate of $H(x)$.

$$H_n^*(t) = \sum_{(t)} \frac{1}{(n^{(1)} + 1 - i)},$$

where $\sum_{(t)}$ denotes sum over i, $x_{(i)}^* \leq t$, and $\{x_{(i)}^*\}$ is the ordered statistics of $\{x_i; \delta = 1\}$. Using this sample cumulative hazard function, we obtain consistent estimators,

$$\widehat{\theta_c^{(1)}} = \frac{n^{(0)}}{\sum_{i=1;\delta=0}^{n}\{H_n^*(y_i) - H_n^*(x_i)\}},$$

$$\widehat{\theta_c^{(2)}} = \frac{n^{(1)}\sum_{i=1}^{n} H_n^*(x_i)}{n^{(0)}\sum_{i=1;\delta=1}^{n}\{H_n^*(y_i) - H_n^*(x_i)\}},$$

$$\widehat{\alpha_c} = \frac{n^{(0)}}{\sum_{i=1}^{n} H_n^*(x_i)}.$$

Theorem 2 *$\widehat{\theta_c^{(1)}}$, $\widehat{\theta_c^{(2)}}$, $\widehat{\alpha_c}$ are consistent estimators of $\theta^{(1)}$, $\theta^{(2)}$ and α, respectively.*

Proof

At first, for $j = 1, 2$ let

$$F_n^{(j)}(t) = \#\{i; x_i > t, \delta = j\}/n^{(0)},$$

$$G_n^{(j)}(t) = \#\{i; y_i > t, \delta = j\}/n^{(0)},$$

and $F^{(j)}$ and $G^{(j)}$ be survivor functions of x and y given $\delta = j$, respectively.

(1) $\widehat{\theta_c^{(1)}}$:

$$\begin{aligned}
\frac{1}{\widehat{\theta_c^{(1)}}} &= \frac{\sum_{i=1;\delta=0}^{n}\{H_n^*(y_i) - H_n^*(x_i)\}}{n^{(0)}} \\
&= \int H_n^*(t)d\{F_n^{(0)}(t) - G_n^{(0)}(t)\} \\
&\longrightarrow \int H(t)d\{F^{(0)}(t) - G^{(0)}(t)\} \\
&= \int H(t)\frac{(1+\alpha)\theta^{(1)}}{1+\alpha-\theta^{(1)}}\{S(t)^{\theta^{(1)}-1} - S(t)^{\alpha}\}f(t)dt - \int H(t)(1+\alpha)S(t)^{\alpha}f(t)dt \\
&= \frac{1+\alpha+\theta^{(1)}}{(1+\alpha)\theta^{(1)}} - \frac{1}{1+\alpha} \\
&= \frac{1}{\theta^{(1)}}.
\end{aligned}$$

(2) $\widehat{\theta_c^{(2)}}$:

$$\begin{aligned}
\frac{1}{\widehat{\theta_c^{(2)}}} &= \frac{n^{(0)}\sum_{i=1;\delta=1}^{n}\{H_n^*(y_i) - H_n^*(x_i)\}}{n^{(1)}\sum_{i=1}^{n} H_n^*(x_i)} \\
&= \frac{n^{(0)}\int H_n^*(t)d\{F_n^{(1)}(t) - G_n^{(1)}(t)\}}{-n\int H_n^*(t)dF_n(t)}.
\end{aligned}$$

While

$$\frac{n^{(0)}}{n} \longrightarrow \Pr(\delta = 0) = \frac{\alpha}{1+\alpha},$$

$$\begin{aligned}
&\int H_n^*(t)d\{F_n^{(1)}(t) - G_n^{(1)}(t)\} \\
\longrightarrow\ &\int H(t)d\{F^{(1)}(t) - G^{(1)}(t)\} \\
=\ &\int H(t)\frac{\alpha\theta^{(2)}}{1+\alpha-\alpha\theta^{(2)}}\{S(t)^{\alpha\theta^{(2)}-1} - S(t)^{\alpha}\}f(t)dt - \int H(t)(1+\alpha)S(t)^{\alpha}f(t)dt \\
=\ &\frac{1+\alpha+\alpha\theta^{(2)}}{(1+\alpha)\alpha\theta^{(2)}} - \frac{1}{1+\alpha} \\
=\ &\frac{1}{\alpha\theta^{(2)}}.
\end{aligned}$$

$$-\int H_n^*(t)dF_n(t) \longrightarrow -\int H(t)DF(t)$$

$$= \int H(t)(1+\alpha)S(t)^{\alpha} f(t)dt$$
$$= \frac{1}{1+\alpha}.$$

By the above equations,

$$\frac{1}{\widehat{\theta_c^{(2)}}} \longrightarrow \frac{1}{\theta^{(2)}}.$$

(3) $\widehat{\alpha_c}$:

$$\begin{aligned}
\frac{1}{\widehat{\alpha_c}} &= \frac{\sum_{i=1}^{n} H_n^*(x_i)}{n^{(0)}} \\
&= -\frac{n}{n^{(0)}} \int H_n^*(t) dF_n(t) \\
&\longrightarrow \frac{1+\alpha}{\alpha} \int H(t) dF(t) \\
&= \frac{1+\alpha}{\alpha} \int H(t)(1+\alpha)S(t)^{\alpha} f(t)dt \\
&= \frac{1+\alpha}{\alpha} \frac{1}{1+\alpha} \\
&= \frac{1}{\alpha}.
\end{aligned}$$

Also in this case, we can easily obtain the two-step estimators using these consistent estimators.

By (7), the maximum partial likelihood estimators satisfy the equations

$$\widehat{\theta_p^{(1)}} = \frac{\sum_{j=1}^{2n} d_j^{(1d)}}{\sum_{j=1}^{2n} \frac{r_j^{(1d)}}{r_j^{(1s)} + \widehat{\theta_p^{(1)}} r_j^{(1d)} + \widehat{\alpha}_p r_j^{(2s)} + \widehat{\alpha}_p \widehat{\theta_p^{(2)}} r_j^{(2d)}}},$$

$$\widehat{\theta_p^{(2)}} = \frac{\sum_{j=1}^{2n} d_j^{(2d)}}{\sum_{j=1}^{2n} \frac{r_j^{(2d)}}{r_j^{(1s)} + \widehat{\theta_p^{(1)}} r_j^{(1d)} + \widehat{\alpha}_p r_j^{(2s)} + \widehat{\alpha}_p \widehat{\theta_p^{(2)}} r_j^{(2d)}}},$$

$$\widehat{\alpha_p} = \frac{\sum_{j=1}^{2n} \{d_j^{(1d)} + d_j^{(2d)}\}}{\sum_{j=1}^{2n} \frac{r_j^{(2s)} + \widehat{\theta_p^{(2)}} r_j^{(2d)}}{r_j^{(1s)} + \widehat{\theta_p^{(1)}} r_j^{(1d)} + \widehat{\alpha}_p r_j^{(2s)} + \widehat{\alpha}_p \widehat{\theta_p^{(2)}} r_j^{(2d)}}}.$$

A single iteration from a consistent estimators produces the two-step estimators which are asymptotically equivalent to the maximum partial likelihood estimators,

$$\widehat{\theta_2^{(1)}} = \frac{\sum_{j=1}^{2n} d_j^{(1d)}}{\sum_{j=1}^{2n} \frac{r_j^{(1d)}}{r_j^{(1s)}+\widehat{\theta_c^{(1)}} r_j^{(1d)}+\widehat{\alpha_c} r_j^{(2s)}+\widehat{\alpha_c}\widehat{\theta_c^{(2)}} r_j^{(2d)}}},$$

$$\widehat{\theta_2^{(2)}} = \frac{\sum_{j=1}^{2n} d_j^{(2d)}}{\sum_{j=1}^{2n} \frac{r_j^{(2d)}}{r_j^{(1s)}+\widehat{\theta_c^{(1)}} r_j^{(1d)}+\widehat{\alpha_c} r_j^{(2s)}+\widehat{\alpha_c}\widehat{\theta_c^{(2)}} r_j^{(2d)}}},$$

$$\widehat{\alpha_2} = \frac{\sum_{j=1}^{2n} \{d_j^{(1d)} + d_j^{(2d)}\}}{\sum_{j=1}^{2n} \frac{r_j^{(2s)}+\widehat{\theta_c^{(2)}} r_j^{(2d)}}{r_j^{(1s)}+\widehat{\theta_c^{(1)}} r_j^{(1d)}+\widehat{\alpha_c} r_j^{(2s)}+\widehat{\alpha_c}\widehat{\theta_c^{(2)}} r_j^{(2d)}}}.$$

2.4 Test of the proportional hazards assumption

The proportional hazard model is popular and widely available to survival analysis with censored data. The model for association in bivariate survival data is also based on the proportional hazards assumption. An important problem arising in applying these models is to assess the proportionality of the hazards. However, the analytical results are mostly discussed on the existence of the proportionality. In fact, Andersen (1982) has mentioned "a number of worked examples of analyses of survival data using this model have been published, but surprisingly little attention has been paid to the problem of model checking." In view of such a necessity of model checking, we propose a test of the proportionality of the hazards.

The observations from sample i $(i = 1, 2)$ are (X_{ij}, d_{ij}) $j = 1, ..., n_i$, where $X_{ij} = \min(X_{ij}^0, U_{ij})$, $d_{ij} = I(X_{ij} = X_{ij}^0)$, $I(\cdot)$ is the indicator function, X_{ij}^0 is the true survival time of j-individual of i-sample with a continuous distribution function F_i , and U_{ij} is the corresponding censoring variable. We assume that U_{ij} are independent identically distributed random variable with distribution function L_i , and that the variables U_{ij} are independent of the variables X_{ij}.

Furthermore, we assume that $n_i/n \longrightarrow r_i$, $0 < r_i < 1$, as $n \longrightarrow \infty$, where $n = n_1 + n_2$. Let the cumulative hazard function of F_i be denoted by $H_i(t)$, that is $H_i(t) = -\log\{1 - F_i(t)\}$. Also let

$$N_i(t) = \#\{j; x_{ij} \leq t \text{ and } d_{ij} = 1\},$$

$$Y_i(t) = \#j; x_{ij} \geq t,$$

$$y_i(t) = \{1 - F_i(t)\}\{1 - L_i(t-)\}.$$

We assume that the supports of $y_i(t)$ are $[0, \infty]$.

We intend to test the null hypothesis that $H_1(t) = \theta H_2(t)$, for $t \geq 0$ and some unknown constant θ.

Now we can write the logarithm of the Cox's partial likelihood as follows,

$$\int_0^\infty \log\theta dN_i(s) - \int_0^\infty \log\{Y_1(s)\theta + Y_2(s)\}d\{N_1(s) + N_2(s)\},$$

and let

$$C_n(\theta; t) = \int_0^t \log\theta dN_1(s) - \int_0^t \log\{Y_1(s)\theta + Y_2(s)\}d\{N_1(s) + N_2(s)\}.$$

The estimator $\hat{\theta}$ is defined as the solution to the likelihood equation

$$\frac{\partial}{\partial\theta} C_n(\theta; \infty) = 0.$$

The asymptotic properties of this estimator have been studied by many workers, for example Tsiatis (1981) and Wong (1986).

Furthermore let $B_n(\theta; t)$ be the product of the parameter θ and the derivative of $C_n(\theta; t)$ with respect to θ. Then,

$$B_n(\theta; t) = \int_0^t dN_1(s) - \int_0^t \frac{Y_1(s)\theta}{Y_1(s)\theta + Y_2(s)} d\{N_1(s) + N_2(s)\}.$$

This can be interpreted as the residual at the time t. The first term of left hand side is the observed number of deaths from sample 1 until time t. The second term is the corresponding expected number under the null hypothesis.

Because the parameter θ is unknown, we replace θ by $\widehat{\theta}$ in B_n and let,

$$W_n(t) = n^{-1/2} B_n(\widehat{\theta}; t).$$

It is clear by definition that $W_n(0) = W_n(\infty) = 0$.

In the next section, we propose a test by using $W_n(t)$.

2.4.1 Newly proposed test

At first we give two lemmas.

Lemma 1 *Under $H_1(t) = \theta_0 H_2(t)$ for any $t \geq 0$ and some θ_0 , the estimator $\widehat{\theta}$ is strongly consistent.*

Proof

$B_n(\theta; t)$ is clearly a non-increasing function of θ. For fixed θ, as $n \longrightarrow \infty$,

$$\frac{1}{n} B_n(\theta; t) \longrightarrow B(\theta; t), \quad \text{a.s.},$$

where

$$B(\theta; t) = \int_0^t \frac{r_1 y_1(s) r_2 y_2(s)}{\theta r_1 y_1(s) + r_2 y_2(s)} d\{H_1(s) - \theta H_2(s)\}.$$

If $H_1(t) = \theta_0 H_2(t)$ for any $t \geq 0$, then $B(\theta_0; \infty) = 0$. $B(\theta; t)$ is also a strictly decreasing function on a neighborhood of θ_0 . Therefore,

$$\Pr(\lim_{n \to \infty} \widehat{\theta} = \theta_0) = 1.$$

Let

$$V_n(t) = \{\widehat{\theta} Q_n(\infty)\}^{-1/2} W_n(t),$$

$$Q_n(t) = n^{-1} \int_0^t \frac{Y_1(s) Y_2(s)}{\{\widehat{\theta} Y_1(s) + Y_2(s)\}^2} d\{N_1(s) N_2(s)\},$$

$$Q(t) = r_1 r_2 \int_0^t \frac{y_1(s) y_2(s)}{\theta_1 y_1(s) + r_2 y_2(s)} dH_2(s),$$

$$G_n(t) = Q_n(t)/Q_n(\infty),$$

and

$$G(t) = Q(t)/Q(\infty).$$

Lemma 2 *Under $H_1(t) = \theta_0 H_2(t)$ for any $t > 0$ and some θ_0, the process $\{V_n(t); 0 \leq t \leq \infty\}$ converges in law to the process $\{W^0(G(t)); 0 \leq t \leq \infty\}$, where $W^0(\cdot)$ is the tied-down Brownian motion process.*

Proof

See Theorem 1 of Wei (1984).

Let

$$T_n = \int_0^\infty V_n(s)^2 dG_n(s).$$

This is our proposed test statistic with the properties which are represented by following two theorems.

The above two lemmas yield the following theorem.

Theorem 3 *Under the null hypothesis, the asymptotic distribution of T_n is same to the distribution of*

$$\int_0^1 W^0(t)^2 dt.$$

The consistent property of this test is obtained by the next theorem.

Theorem 4 *When the null hypothesis is not hold, for any $c > 0$,*

$$\lim_{n\to\infty} \Pr(T_n > c) = 1.$$

Proof

From the proof of lemma 1, the estimator $\widehat{\theta}$ converges to a constant θ_A , even if the assumption of proportionality of hazards is violated. On the other hand, for fixed t,

$$n^{-1/2} W_n(t) \longrightarrow \int_0^t \frac{r_1 y_1(s) r_2 y_2(s)}{\theta_A r_1 y_1(s) + r_2 y_2(s)} d\{H_1(s) - \theta_A H_2(s)\}.$$

Clearly there exists some $t^* > 0$ such that

$$\int_0^{t^*} \frac{r_1 y_1(s) r_2 y_2(s)}{\theta_A r_1 y_1(s) + r_2 y_2(s)} d\{H_1(s) - \theta_A H_2(s)\} \neq 0.$$

Therefore, $W_n(t^*) \longrightarrow \infty$ in probability, for some t^* and $T_n \longrightarrow \infty$ in probability.

2.4.2 Remarks

The tables of the critical points of the distribution of $\int_0^1 W^0(t)^2 dt$ were given in some papers, for example, Schumacher (1984).

A general form of the integral-type test statistics of this problem can be represented as follows,

$$T_{n,\psi} = \int_0^\infty \psi(s) V_n(s)^2 dG_n(s),$$

where ψ is a weight function. We can obtain various test statistics by selection of the weight function ψ. For example, when the estimated variance function of $V_n(t)$ is selected as $\psi(t)^{-1}$, the test statistic $T_{n,\psi}$ is an Anderson-Darling type test statistic. The asymptotic distributions of these statistics were studied in Durbin (1973).

3 Linear Model with Heavy-tailed Error Distributions

Error terms in most statistical models are assumed to be the random variables following the normal distribution, and under this assumption, the maximum likelihood estimation is carried out. In this case, the theoretical validity of the results is guaranteed only when data satisfy the assumption of the normality. In the general case of applying such a method to actual data, its robustness becomes a problem.

As a model of error distribution other than the normal distribution, a scale mixture of normals might be used, which has a relatively heavier tail than that of the normal and is unimodal and symmetrical distribution. The family of scale mixtures of the normal distribution includes, in particular, the t-distribution, double exponential distribution and logistic distribution.

The assumption of a heavier-tailed distribution reflects interest in estimates which are relatively unaffected by outliers. In particular, the t-distribution has been frequently used in analysis of real data (Zellner 1976, Sutradhar and Ali 1986 and so on.), when they considered that data included some outliers. Aitkin and Wilson (1980) treated several types of mixture models of two normals. In this section, we do not confine the t family or contaminated normal family, but instead employ the family of scale mixtures of the normals and give a general method for parameter estimation. At that time two problems would arise. One is the identification of error distribution in the family and the other is the detection of outliers. As mentioned in Chapter 1, we treat both problems as the model selection with the help of the AIC.

3.1 Linear model and iteratively reweighted least squares

We begin with the linear model and iteratively reweighted least squares (IRLS). The data consist of an $n \times 1$ response vector $\boldsymbol{Y}$ and an $n \times m$ design matrix $\boldsymbol{X}$. It is

assumed that

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{\beta}$ is a vector of parameters and e is a vector such that the components s_i of $\sigma^{-1}e$ are independently and identically distributed with known density $f(s_i)$ on $-\infty < s_i < \infty$. In the context of ordinary least squares we do not use the assumption of error distribution. The weigthed least squares estimate of $\boldsymbol{\beta}$ is chosen to minimize

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{8}$$

for a particular given $\boldsymbol{W}$, where $\boldsymbol{W}$ is a positive definite diagonal matrix. We assume that $X'WX$ is full rank, so that the unique solution which attains the minimum of (8) exists, and it can be written

$$\boldsymbol{b}(\boldsymbol{W}) = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}).$$

As the weighted least squares estimate depends on the weight matrix $\boldsymbol{W}$, we have to select proper weight matrix. When the weight matrix is not fixed, IRLS is used. IRLS is a process of obtaining the sequence $\boldsymbol{b}^{(0)}, \boldsymbol{b}^{(1)}, ...,$ and $\boldsymbol{b}^{(l+1)}$ for $l > 0$ is a weighted least squares estimate corresponding to a weight matrix $\boldsymbol{W}^{(l)}$, where $\boldsymbol{W}^{(l+1)}$ depends on $\boldsymbol{b}^{(l)}$. To define a specific version of IRLS, we need to define a sequence of the weight matrices.

A general statistical justification for IRLS arises from the fact that it can be viewed as a ML estimation.

The log likelihood is

$$\ell(\boldsymbol{\beta}, \sigma) = -n\log\sigma + \sum_{i=1}^{n} \log f(s_i), \tag{9}$$

where

$$s_i = (y_i - \boldsymbol{X}_i\boldsymbol{\beta})/\sigma.$$

Let

$$w(s) = \begin{cases} -\frac{df(s)/ds}{sf(s)}, & \text{for } s \neq 0 \\ -\lim_{s\to 0} \frac{df(s)/ds}{sf(s)}, & \text{for } s = 0. \end{cases} \tag{10}$$

We assume in (10) that $f(s) > 0$ for all s, that $df(s)/ds$ exists for $z \neq 0$ and that $w(s)$ has a finite limit as $z \to 0$. Also, since $w(s)$ is selected as the weight function we must assume that $df/ds \leq 0$ for $s > 0$ and $df/ds \geq 0$ for $s < 0$, hence $f(s)$ is unimodal with a mode at $s = 0$. Furthermore, to simplify the theory we assume $df(0)/ds = 0$.

Dempster *et al.*(1980) gave the following Lemma and Theorem concerned with the connection between IRLS process and the log likelihood function (9).

Lemma 3 *For $(\boldsymbol{\beta}, \sigma)$ such that $\sigma > 0$ and $w(z_i)$ is finite for all i, the equations derived from the log likelihood (9) are given by*

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} = 0, \tag{11}$$

and

$$-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + n\sigma^2 = 0, \tag{12}$$

where $\boldsymbol{W}$ is a diagonal matrix with elements $w(s_1), w(s_2), ..., w(s_n)$.

Lemma 3 suggests an IRLS procedure. Since $\boldsymbol{W}$ depends on $\boldsymbol{\beta}$ and σ, we cannot immediately solve the equations (11), (12). Thus we might derive an iterative procedure: at each iteration substitute the temporary values of $\boldsymbol{\beta}$ and σ into the expression for $\boldsymbol{W}$; then, holding $\boldsymbol{W}$ fixed, solve (11) and (12) to obtain the next values of $\boldsymbol{\beta}$ and σ, that is, we take

$$\hat{\boldsymbol{\beta}}^{(l+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(l)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(l)}\boldsymbol{Y}, \tag{13}$$

and

$$\hat{\sigma}^{(l+1)} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l+1)})'\boldsymbol{W}^{(l)}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(l+1)})/n. \tag{14}$$

Theorem 5 *If an instance of an IRLS algorithm defined by (13) and (14) converged to $(\boldsymbol{\beta}^*, \sigma^*)$ where the weight are all finite and $\sigma^* > 0$, then $(\boldsymbol{\beta}^*, \sigma^*)$ is a stationary point of $\ell(\boldsymbol{\beta}, \sigma)$.*

3.2 Scale mixtures of normal distributions

If u is a standard normal random variable with density

$$(2\pi)^{-1/2}\exp(-\frac{1}{2}u^2), \quad (-\infty < u < \infty)$$

and q is a positive random variable distributed independently of u with distribution function $\mathrm{M}(q)$, then the random variable $z = uq^{-1/2}$ is called to have a scale mixture of normal distributions. Andrew *et al.* (1972) said it had a normal/independent distribution.

The scale mixtures of normal distributions are a convenient family of symmetric distributions for components of error terms. The following examples show some familiar examples of these.

Example 1 : Contaminated normal distribution

If

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

then the distribution of z is the contaminated normal distribution with contaminated fraction δ and variance inflation factor λ, that is,

$$z \sim (1-\delta) \times \mathrm{N}(0, \Sigma) + \delta \times \mathrm{N}(0, \Sigma/\lambda).$$

Example 2 : t distribution

Let ν be a constant. If the distribution of $\nu \times q$ is χ^2 distribution with ν degrees of freedom, then the distribution of z is the t distribution with ν degree of freedom. When $\nu = 1$, it is also called the Cauchy distribution.

Example 3 : Double exponential distribution

If

$$\mathrm{M}(q) = \frac{1}{2}q^{-2}\exp(-\frac{1}{2q}),$$

z has the double exponential distribution with the probability density function

$$f(z) = \frac{1}{2}\exp(-\mid z \mid).$$

Example 4 : Logistic distribution

If

$$\mathrm{M}(q) = \sum_{k=1}^{\infty}(-1)^{k-1}k^2q^{-2}\exp(-\frac{k}{2q}),$$

z has the logistic distribution with the distribution function

$$F(z) = [1 + \exp(-x)]^{-1}.$$

Dempster *et al.* (1980) pointed out a close connection with IRLS. Knowledge of the scale factors $q_i^{-1/2}$ in each component $e_i = \sigma u_i q_i^{-1/2}$ would lead to the use of weighted least squares with weight matrix W whose diagonal elements are $q_1, q_2, ..., q_n$, and treating these weights as missing data might lead to a statistically natural derivation of IRLS.

The density function of z, $f(z)$ is

$$f(z) = \int_0^{\infty}(2\pi)^{-1/2}q^{1/2}\exp(-\frac{1}{2}qz^2)\mathrm{dM}(z). \tag{15}$$

Proposition 1 *Suppose that z is a scale mixture random variable of normal distribution with the density function (15) Then for* $0 <\mid z \mid< \infty$*;*

(i) *the conditional distribution of q given z exists,*

(ii) $\mathrm{E}(q^k \mid z) < \infty$, *for* $k > -1/2$,

(iii) $w(z) = \mathrm{E}(q \mid z)$,

(iv) $dw(z)/dz = -z\mathrm{var}(q \mid z)$,

(v) $w(z) = w(-z)$ *is finite, positive, and nonincreasing for* $z > 0$.

For $z = 0$*:*

(vi) *the conditional distribution of* q *given* z *exists if and only if* $\mathrm{E}(q^{1/2}) < \infty$,

(vii) $w(0) \geq w(z)$ *for* $z \neq 0$ *and* $w(0)$ *is finite if and only if* $\mathrm{E}(q^{3/2}) < \infty$,

(viii) $dw(0)/dz$ *is finite if and only if* $\mathrm{E}(q^{5/2}) < \infty$.

Proposition 2 *Suppose that* $u \sim \mathrm{N}(0,1)$ *and that* q *is a positive random variable distributed independently of* u *with distribution function* $M(q)$*. Then*

$$z = uq^{-1/2}$$

is equivalent to that the conditional distribution of z *given* $q = q_0$ *is* $\mathrm{N}(0, 1/q_0)$.

Proposition 3 *The kurtosis of* z *is never less than that of* u.

Proof

$$\begin{aligned}
\frac{\mathrm{E}(z^4)}{\mathrm{E}(z^2)^2} &= \frac{\mathrm{E}(u^4 q^{-2})}{\mathrm{E}(u^2 q^{-1})^2} \\
&= \frac{\mathrm{E}(u^4)\mathrm{E}(q^{-2})}{\mathrm{E}(u^2)^2 \mathrm{E}(q^{-1})^2} \\
&\geq \frac{\mathrm{E}(u^4)}{\mathrm{E}(u^2)^2}.
\end{aligned}$$

The family of scale mixtures of the normal is heavier-tailed than the normal distribution in the meaning of the kurtosis. We note that the condition of the normality of u is not necessary in the above lemma, that is, even when u is not limited to a normal random variable, the tail becomes heavier than that of the distribution of the original random variable.

3.3 Multivariate model

We now consider an extension of the above results to the multivariate case.

3.3.1 Basic statistics

Let $\boldsymbol{U}$ be a p-component random vector distributed as $\mathrm{N}(\mathbf{o}, \boldsymbol{\Sigma})$ and q be a positive random variable distributed independently of $\boldsymbol{U}$ with distribution function $\mathrm{M}(q)$. Then the random vector $\boldsymbol{Z} = \boldsymbol{U}q^{-1/2}$ has a scale mixture distribution of multivariate normal.

Proposition 4 *The density function of* $\boldsymbol{Z}$, $f(\boldsymbol{Z})$ *is*

$$f(\boldsymbol{Z}) = \int_0^\infty (2\pi)^{-1/2} q^{p/2} \mid \Sigma \mid^{-1/2} \exp(-\frac{1}{2} q \boldsymbol{Z}' \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}) \mathrm{dM}(q). \tag{16}$$

The mean vector and covariance matrix of $\boldsymbol{Z}$ are reprensented by the following lemma.

Proposition 5

$$\mathrm{E}(\boldsymbol{Z}) = 0,$$

$$\mathrm{Cov}(\boldsymbol{Z}) = \int_0^\infty q^{-1} d\mathrm{M}(q) \boldsymbol{\Sigma}.$$

The next lemma gives the same result in the multivariate case of lemma 3. The multivariate kurtosis $\kappa_{2,p}$ is defined by Mardia (1970) as follows:

> Let $\boldsymbol{X}$ be a arbitrary p-dimensional random vector, $\boldsymbol{\mu}$ be its $p \times 1$ mean vector, and $\boldsymbol{\Sigma}$ be its $p \times p$ covariance matrix. Then
>
> $$\kappa_{2,p} = \mathrm{E}\{[(\boldsymbol{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})]^2\}.$$

Proposition 6 *The multivariate Kurtosis of Z is not less than that of U.*

Proof

$$\begin{aligned}
\mathrm{E}[(\boldsymbol{Z}'\{\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}')\}^{-1}\boldsymbol{Z})^2] &= \mathrm{E}[(\frac{\boldsymbol{Z}'}{q^{1/2}}\{E(\frac{1}{q}\boldsymbol{Z}\boldsymbol{Z}')\}^{-1}\frac{\boldsymbol{Z}}{q^{1/2}})^2] \\
&= \mathrm{E}[(\frac{1}{q}\boldsymbol{Z}'\{E(\frac{1}{q})\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}')\}^{-1}\boldsymbol{Z})^2] \\
&= \mathrm{E}[\boldsymbol{Z}'\{\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}')\}^{-1}\boldsymbol{Z}]^2\mathrm{E}(\frac{1}{q^2})\{\mathrm{E}(\frac{1}{q})\}^{-2} \\
&\geq \mathrm{E}[\boldsymbol{Z}'\{\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}')\}^{-1}\boldsymbol{Z}]^2
\end{aligned}$$

Proposition 7

$$\mathrm{E}(q \mid \boldsymbol{Z}) = -\boldsymbol{Z}'\frac{df(\boldsymbol{Z})}{d\boldsymbol{Z}}/\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}f(\boldsymbol{Z}),$$

and if $\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} = \boldsymbol{Z}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_0$ *then* $\mathrm{E}(q \mid \boldsymbol{Z}) = \mathrm{E}(q \mid \boldsymbol{Z}_0)$.

Let $s^2 = s^2(\boldsymbol{Z}) = \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}$ and $w(s^2) = \mathrm{E}(q \mid \boldsymbol{Z})$, because w has the same value if s^2 is same.

Proposition 8 $w(s^2)$ *is finite, positive, and nonincreasing for* $s^2 \neq 0$.

Proof

$$\begin{aligned}
\frac{dw}{ds^2} &= -\frac{1}{2}\frac{\int_0^\infty q^2(2\pi)^{-1/2}q^{1/2} \mid \boldsymbol{\Sigma} \mid^{-1/2}\exp(-\frac{1}{2}qs^2)d\mathrm{M}(q)}{\int_0^\infty(2\pi)^{-1/2}q^{1/2} \mid \boldsymbol{\Sigma} \mid^{-1/2}\exp(-\frac{1}{2}qs^2)d\mathrm{M}(q)} \\
&\quad +\frac{1}{2}\{\frac{\int_0^\infty q(2\pi)^{-1/2}q^{1/2} \mid \boldsymbol{\Sigma} \mid^{-1/2}\exp(-\frac{1}{2}qs^2)d\mathrm{M}(q)}{\int_0^\infty(2\pi)^{-1/2}q^{1/2} \mid \boldsymbol{\Sigma} \mid^{-1/2}\exp(-\frac{1}{2}qs^2)d\mathrm{M}(q)}\}^2 \\
&\leq 0
\end{aligned}$$

We now consider the multivariate regression to give the relation between the conditional expectation of q given $\boldsymbol{Z}$ and IRLS.

Suppose $\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_n$ are a set of n observations, $\boldsymbol{Y}_i$ following the model,

$$\boldsymbol{Y}_i = \boldsymbol{\beta}\boldsymbol{X}_i + \boldsymbol{e}_i, \tag{17}$$

where $\boldsymbol{\beta}$ is a $p \times m$ matrix of parameters, $\boldsymbol{X}_i$ is a known design matrix and $\boldsymbol{e}_i/\sigma$ is a vector with the density function $f(\boldsymbol{e}_i)$ and $\boldsymbol{\Sigma}' = \sigma^2 \mathbf{I}_p$ for the sake of simplicity. Then the log likelihood function is

$$\ell(\boldsymbol{\beta}) = -\frac{np}{2}\log\sigma^2 + \sum_{i=1}^{n}\log f(\boldsymbol{e}_i). \tag{18}$$

Let

$$w_{ij}^* = -\frac{1}{e_{ij}f(\boldsymbol{e}_i)}\frac{df(\boldsymbol{e}_i)}{d\boldsymbol{e}_i}$$

and

$$\boldsymbol{W}_i^* = \mathrm{diag}\{w_{i1}^*, w_{i2}^*, ..., w_{ip}^*,\}$$

where e_{ij} is the j-th component of $\boldsymbol{e}_i$. Then the likelihood equations from (18) are

$$\sum_{i=1}^{n}\boldsymbol{W}_i^*\boldsymbol{e}_i\boldsymbol{X}_i' = \mathbf{o},$$

and

$$np\sigma^2 - \sum_{i=1}^{n}\boldsymbol{e}_i'\boldsymbol{W}_i^*\boldsymbol{e}_i = \mathbf{o}.$$

Proposition 9 *If $\boldsymbol{e}_i$ has the density function (10), then*

$$\mathrm{E}(q_i \mid \boldsymbol{e}_i) = \boldsymbol{e}_i'\boldsymbol{W}_i^*\boldsymbol{e}_i/\boldsymbol{e}_i'\boldsymbol{e}_i.$$

While w_{ij}^* is a weight of the j-th component of the i-th individual, $\mathrm{E}(q_i \mid \boldsymbol{e}_i)$ might be regarded as a weight of the i-th individual which is regarded as a weighted average of w_{ij}^*.

3.4 ML estimation and EM algorithm

We now establish a concrete procedure of ML estimation using the EM algorithm. It is assumed that $\boldsymbol{Y}_i = \boldsymbol{\beta}\boldsymbol{X}_i + \boldsymbol{e}_i$ $(i = 1, .., n)$, and $\boldsymbol{e}_i$ is independently identically distributed from a scale mixture of multivariate normal. Namely, there exist n mutually independent positive random variables q_i, which follow the distribution function $\mathrm{M}(q_i)$, and conditional on q_i, $\boldsymbol{e}_i$ follows $\mathrm{N}(\mathbf{o}, \boldsymbol{\Sigma}/q_i)$. According to the method of description of the EM algorithm by Dempster *et al.* (1977), $\{\boldsymbol{Y}_i\}$ represent the directly observed data, called incomplete data because there is assumed to exist further potential data $\{q_i\}$ which are not observed, and we denote by $\{\boldsymbol{O}_i\} = \{\boldsymbol{Y}_i, q_i\}$ a representation of the complete data, including both observed and unobserved. The log likelihood of $\{\boldsymbol{O}_i\}$ is,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = Const. - \frac{n}{2}\log \mid \boldsymbol{\Sigma} \mid -\frac{1}{2}\sum_{i=1}^{n} q_i(\boldsymbol{Y}_i - \boldsymbol{\beta}\boldsymbol{X}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\beta}\boldsymbol{X}_i). \qquad (19)$$

The evaluation of the conditional expectation of $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ (19) is realized in E-step, and the maximization of $\mathrm{E}(\ell \mid \boldsymbol{Y}_i)$ with respect to the objective parameters is realized in M-step, respectively.

E-step: With the observed data $\boldsymbol{Y}_i$ and the temporary values of parameters $\boldsymbol{\beta}^{(l)}$, $\boldsymbol{\Sigma}^{(l)}$, the conditional expectation of ℓ is evaluated. In this case it is none other than determining the conditional expectation of q_i, which would be determined if M(.) is specified.

M-step: Assuming that $q_i^{(l+1)}$ determined in the E-step was given, the estimated values of the parameters are renewed such that these values maximizes the temporary log likelihood. In this case,

$$\hat{\boldsymbol{\beta}}^{(l+1)} = \sum_{i=1}^{n} q_i^{(l+1)}\boldsymbol{Y}_i\boldsymbol{X}_i'(\sum_{i=1}^{n} q_i^{(l+1)}\boldsymbol{X}_i\boldsymbol{X}_i')^{-1},$$

$$\widehat{\boldsymbol{\Sigma}}^{(l+1)} = \sum_{i=1}^{n} q_i^{(l+1)}(\boldsymbol{Y}_i - \hat{\boldsymbol{\beta}}^{(l+1)}\boldsymbol{X}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\beta}}^{(l+1)}\boldsymbol{X}_i)'/n.$$

The above E-step and M-step are repeatedly carried out, taking the proper initial values, and the ML estimation are obtained. The E-step is materialized by specifying the distribution of q. Hereinafter, its several examples are shown.

Example 5 : Contaminated multivariate normal distribution

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

where $\lambda < 0$ and $0 < \delta < 1$.

When M(q) is specified as described above, it is to assume the distribution of the mixture of $\mathrm{N}(\mathbf{o}, \boldsymbol{\Sigma})$ and $\mathrm{N}(\mathbf{o}, \boldsymbol{\Sigma}/\lambda)$ in the ratio of $1-\delta$ to δ.

Hereupon, the conditional distribution of q when $\boldsymbol{Y}$, $\boldsymbol{X}$ and the temporary values of the parameters $\widehat{\boldsymbol{\beta}}^{(l)}$, $\widehat{\boldsymbol{\Sigma}}^{(l)}$ are given is concretely evaluated, and

$$\begin{aligned} w^{(l+1)} &= \mathrm{E}(q \mid \boldsymbol{e}) \\ &= \frac{1-\delta+\delta\lambda^{1+p/2}\exp\{(1-\lambda)d^2/2\}}{1-\delta+\delta\lambda^{p/2}\exp\{(1-\lambda)d^2/2\}} \end{aligned} \tag{20}$$

is obtained, where

$$d^2 = (\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^{(l)}\boldsymbol{X})'\widehat{\boldsymbol{\Sigma}}^{(l)-1}(\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^{(l)}\boldsymbol{X}).$$

Example 6 : Multivariate t distribution

If $q \times \nu$ has the chi-squared distribution with ν degrees of freedom, the marginal distribution is the multivariate t distribution (Cornish, 1954). At this time,

$$\begin{aligned} w^{(l+1)} &= \mathrm{E}(q \mid \boldsymbol{e}) \\ &= (\nu+p)/(\nu+d^2). \end{aligned} \tag{21}$$

Both models downweight observations with large d^2. However, the curve of the weights is quite different for the two models, the multivariate t model producing relatively smoothly declining weights with increasing d^2, and the contaminated normal model tending to concentrate the low weights in a few outlying observations.

3.4.1 Estimation of mixing parameters

If the model of the distribution function of q includes some unknown parameters, for example the degrees of freedom ν for the multivariate t model, the contamination fraction δ and variance inflation factor λ for the contaminated normal model and so on, we have to estimate such parameters.

The distribution of p-variate random vector $\boldsymbol{Y}$ is assumed such that the conditional distribution of $\boldsymbol{Y}$ given positive random variable q is $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/q)$. Let $f(q;\boldsymbol{\theta})$ be the probability density function of q with unknown parameter vector $\boldsymbol{\theta}$ (mixing parameters), and $g(\boldsymbol{Y};\boldsymbol{\mu},\boldsymbol{\Sigma})$ denote the normal density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then the joint density function of $\boldsymbol{Y}$ and q is $f(q;\theta)g(\boldsymbol{Y};\boldsymbol{\mu},\boldsymbol{\Sigma}/q)$. As $\boldsymbol{\theta}$ is not included in $g(\cdot)$ but in $f(\cdot)$, the log likelihood concerned with $\boldsymbol{\theta}$, based on complete data $\{(\boldsymbol{Y}_i, q_i), i = 1, 2, .., n\}$ is

$$\text{Const.} + \sum_{i=1}^{n} \log f(q_i;\boldsymbol{\theta}). \tag{22}$$

ML estimation of $\boldsymbol{\theta}$ is performed via the EM algorithm: the evaluation of the conditional expectation of (22) given ovservation $\{\boldsymbol{Y}_i, i = 1, ..., n\}$ and temporary values of parameters, is realized in Estep. The maximization of the expected log likelihood obtained in E-step, with respect to $\boldsymbol{\theta}$ is realized in M-step.

We now illustrate concrete algorithm for t model. (see Lange *et al.*, 1989). Given l-th estimates $\boldsymbol{\mu}^{(l)}$, $\boldsymbol{\Sigma}^{(l)}$, $\nu^{(l)}$, in the E-step we compute $w_i^{(l)}$ using (21) with $\nu = \nu^{(l)}$, and

$$\begin{aligned} v_i &= \mathrm{E}(\log q_i \mid \boldsymbol{e}_i) \\ &= \psi(\nu^{(l)}/2 + p/2) - \log(\nu^{(l)}/2 + d^2/2), \end{aligned}$$

where $\psi(x) = \frac{d}{dx}\log\{\Gamma(x)\}$, the digamma function (psi function). In the M-step we compute $\boldsymbol{\mu}^{(l+1)}$ and $\boldsymbol{\Sigma}^{(l+1)}$ and find $\nu^{(l+1)}$ that maximizes

$$\ell_1(\nu) = \frac{n\nu}{2}\log(\nu/2) - n\log\{\Gamma(\nu/2)\} + (\frac{\nu}{2} - 1)\sum_{i=1}^{n} v_i^{(l)} - \frac{\nu}{2}\sum_{i=1}^{n} w_i^{(l)}.$$

It is easy to find the value of ν that maximizes ℓ_1 using a one dimensional search, for example Newton's method.

For t model, another method are considered. We calculate the maximized log likelihood for a fixed ν, which is

$$\begin{aligned}\ell_2 &= -\frac{1}{2}\log|\widehat{\boldsymbol{\Sigma}}| - \frac{1}{2}(\nu+p)\log(1+d_i^2/\nu) \\ &\quad -\frac{1}{2}p\log(\nu/2) + \log[\Gamma\{(\nu+p)/2\}/\Gamma\{\nu/2\}].\end{aligned}$$

We can regard the maximized log likelihood as a function of the degrees of freedom ν, and select the value of ν as the estimate, which attain the maximum over a grid of values of ν.

The case of the contaminated normal model is more complicated, because the model includes two parameters (Little, 1988). When the variance inflation factor λ is fixed in advance, it is easy to estimate λ simultaneously with μ and Σ by general method described in the top of this section, that is, we have only to add the calculation of $\mathrm{E}\{I(q_i = \lambda) \mid \boldsymbol{e}_i\}$ to the E-step and

$$\delta^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{I(q_i = \lambda) \mid \boldsymbol{e}_i^{(l)}\}$$

to the M-step, where I(.) is a index function and

$$\mathrm{E}\{I(q_i = \lambda) \mid \boldsymbol{Y}_i, \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)}\}\frac{\delta\lambda^{p/2}\exp\{(1-\lambda)d^2/2\}}{1-\delta+\delta\lambda^{p/2}\exp\{(1-\lambda)d^2/2\}}.$$

When λ is treated as a parameter, the simultaneous estimation of λ, δ with β and Σ can not be directly derived. Because it is meaningless to estimate λ when q_i (or w_i) are given. Thus the estimation of λ is performed based on the log likelihood ℓ_3 from the marginal distribution of Y_i, which is

$$\ell_3 = \text{Const.} - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}d_i^2 + \sum_{i=1}^{n}\log[1-\delta+\delta\lambda^{p/2}\exp\{(1-\lambda)d_i^2/2\}].$$

Then

$$\delta^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{I(q_i = \lambda^{(l)}) \mid \boldsymbol{Y}_i, \boldsymbol{\mu}^{(l+1)}, \boldsymbol{\Sigma}^{(l+1)}\} \tag{23}$$

and $\lambda^{(l+1)}$ is obtained as a solution of equation

$$\lambda^{(l+1)} = \frac{p\sum_{i=1}^{n} \mathrm{E}\{I(q_i = \lambda^{(l+1)}) \mid \boldsymbol{Y}_i, \boldsymbol{\mu}^{(l+1)}, \boldsymbol{\Sigma}^{(l+1)}\}}{\sum_{i=1}^{n} d_i^{2(l+1)} \mathrm{E}\{I(q_i = \lambda^{(l+1)}) \mid \boldsymbol{Y}_i, \boldsymbol{\mu}^{(l+1)}, \boldsymbol{\Sigma}^{(l+1)}\}}. \tag{24}$$

Note that the equation (24) for $\lambda^{(l+1)}$ depends on $\delta^{(l+1)}$, $\boldsymbol{\mu}^{(l+1)}$ and $\boldsymbol{\Sigma}^{(l+1)}$ not $\delta^{(l)}$, $\boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\Sigma}^{(l)}$.

3.4.2 On the convergence property

Before demonstrating property of the above method, we briefly rewrite outline of GEM algorithm. Instead of the "complete data" $\boldsymbol{x}$, we observe the "incomplete data" $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$. Let the density functions of $\boldsymbol{x}$, $\boldsymbol{y}$ be $f(\boldsymbol{x};\boldsymbol{\phi})$, $g(\boldsymbol{y};\boldsymbol{\phi})$, respectively. Furthermore, let $k(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\phi}) = f(\boldsymbol{x};\boldsymbol{\phi})/g(\boldsymbol{y};\boldsymbol{\phi})$ be the conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$. Then the log likelihood can be written in the following form

$$L(bphi') = \log g(\boldsymbol{y};\boldsymbol{\phi}') = Q(\boldsymbol{\phi}' \mid \boldsymbol{\phi}) - H(\boldsymbol{\phi}' \mid \boldsymbol{\phi}),$$

where

$$Q(\boldsymbol{\phi}' \mid \boldsymbol{\phi}) = \mathrm{E}\{\log f(\boldsymbol{x};\boldsymbol{\phi}') \mid \boldsymbol{y}, \boldsymbol{\phi}\},$$

$$H(\boldsymbol{\phi}' \mid \boldsymbol{\phi}) = \mathrm{E}\{\log k(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\phi}') \mid \boldsymbol{y}, \boldsymbol{\phi}\},$$

and these are assumed to exist for any $(\boldsymbol{\phi}, \boldsymbol{\phi}')$.

In general, $Q(\boldsymbol{\phi}' \mid \boldsymbol{\phi}) - Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}) \geq 0$ implies $L(\boldsymbol{\phi}') - L(\boldsymbol{\phi}) \geq 0$. Therefore, for any sequence $\{\boldsymbol{\phi}^{(p)}\}$ generated by GEM algorithm,

$$L(\boldsymbol{\phi}^{(p+1)}) \geq L(\boldsymbol{\phi}^{(p)}). \tag{25}$$

This is essential to the convergence property of GEM algorithm, and hybrid GEM algorithm must keep this property.

Let us show that the above method can generate sequences such that $L(\boldsymbol{\phi}^{(p+1)}) \geq L(\boldsymbol{\phi}^{(p)})$. Let $\boldsymbol{\psi}$ be unknown parameters except λ. Given $\boldsymbol{\psi}^{(p)}$,$\lambda^{(p)}$, from the step of GEM algorithm for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and (23), we obtain $\boldsymbol{\psi}^{(p+1)}$. Then clearly,

$$Q(\boldsymbol{\psi}^{(p+1)}, \lambda^{(p)} \mid \boldsymbol{\psi}^{(p)}, \lambda^{(p)}) \geq Q(\boldsymbol{\psi}^{(p)}, \lambda^{(p)} \mid \boldsymbol{\psi}^{(p)}, \lambda^{(p)}),$$

which yields

$$L(\boldsymbol{\psi}^{(p+1)}, \lambda^{(p)}) \geq L(\boldsymbol{\psi}^{(p)}, \lambda^{(p)}).$$

$\lambda^{(p+1)}$ determined by the equation (24) satisfies

$$L(\boldsymbol{\psi}^{(p+1)}, \lambda^{(p+1)}) \geq L(\boldsymbol{\psi}^{(p+1)}, \lambda^{(p)}) \geq L(\boldsymbol{\psi}^{(p)}, \lambda^{(p)}),$$

and therefore (25).

3.5 Model selection and detection of outliers

In the above section, we derive the method for calculating ML estimates under the given model. Next we have to select the best model among models under consideration. Now one of the most important purpose of the statistical analysis is the representation of stochastic phenomena by statistical models based on a set of observations. When we have a set of observations, we consider several models for the data. Thus we have to evaluate each model or compare models. The AIC is a leading criterion and it enables us to evaluate the validity of statistical models.

The model selection is performed with the help of the AIC, which is given by

$$\text{AIC} = -2 \times (\text{log likelihood}) + 2 \times (\text{the number of parameters}).$$

The model with the least AIC among the models under consideration is selected.

When the non-normal model, in particular, the multivariate t model with small degrees of freedom or the contaminated normal model with large variance inflation factor, is selected for the given data set, it is of interest to detect which observations would be considered outliers in the context of the normal assumption. Because better fit of non-normal model may suggest that the data set includes some values deviating from the majority of data, which are consider as outliers. w_i can give us the information on the detection of outliers. From a Baysian point of view, w_i can be regarded as the posterior mean of q_i given the data and parameter values when q_i has a prior

distribution $M(q_i)$. If the value of w_i is close to one then the observation is compatible with the normal assumption, however, if this value is close to zero then the observation can be classified as an outlier. However it may be difficult to set the threshold value in the case of the multivariate t model, because the multivariate t model produces relatively smoothly declining weights with increasing d^2. To the contrary, the contaminated model specifies clearly some extreme observations. The concrete performances are shown in numerical examples in Section 3 and Section 4 with the aid of real data.

4 Analysis of Repeated Measures Data with Outliers

A broad range of statistical investigations can be regarded as repeated measurements studies. Their essential feature is that each subject is observed at several different times or under different experimental conditions. The subjects are primary sampling units randomly selected to represent various strata or randomly assigned to levels of a grouping factor. The responses measured under the respective conditions constitute the observational units; because within each subject such responses constitute a profile of inherently multivariate data, their covariance structure plays an important role in the formulation of statistical methods for their analysis.

In many case, some observations are missing or the design is unbalanced for some other reason – for example, the presence of time-varying covariates. A systematic approach to the analysis of incomplete and unbalanced data is to specify a model and to estimate parameters of the model using maximum likelihood method. The original focus of much of the work on ML estimation with incomplete data was centered around the mixed model and the multivariate normal model with unstructured covariance matrix. More recent work has extended both models to allow arbitrary linear models to describe the mean structure and intermediate types of covariance structure. Laird and Ware (1982) studied on ML estimation procedures under general random-effects models for incomplete data and Ware (1985) discussed ML estimation under a similar model with three types of covariance structures: multivariate, random-effects, and autoregressive time series. Jennrich and Schluchter (1986) gave a general model where the expected values of the responses are described as arbitrary linear functions of unknown regression parameters, and the within-subject covariances are modeled as arbitrary functions of a set of unknown covariance parameters. The assumption of error distributions is the normal distribution in all models. In this Chapter, in order to reduce the influence of outliers we replace the normal distributions by the scale

mixtures of normals in Jennrich and Schluchter's model and give a method for ML estimation of parameters.

4.1 Model for data with outliers

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

$$\boldsymbol{Y}_i^* = \boldsymbol{X}_i^* \boldsymbol{\beta} + \boldsymbol{e}_i^* \tag{26}$$

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

But not all of the subjects' responses are not observed, then let $\boldsymbol{Y}_i$ be a $t_i \times 1$ vector containing the observed responses for subject i. According to (26), the $\boldsymbol{Y}_i$ follows the model $\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{e}_i$, where $\boldsymbol{X}_i$ and $\boldsymbol{e}_i$ are a submatrix of $\boldsymbol{X}_i^*$ and a subvector of $\boldsymbol{e}_i^*$, corresponding to the observed responses, respectively. Conditional on q_i, $\boldsymbol{e}_i$ is distributed as $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_i/q_i)$, where $\boldsymbol{\Sigma}_i$ is a submatrix of $\boldsymbol{\Sigma}$. We consider the estimation in both cases unstructured $\boldsymbol{\Sigma}$ and structured $\boldsymbol{\Sigma}$. In the structured case, we assume the elements of $\boldsymbol{\Sigma}$ are known functions of m unknown parameters contained in the vector $\boldsymbol{\theta}$. The regression parameters $\boldsymbol{\beta}$ vary independently of the covariance parameters $\boldsymbol{\theta}$. When we wish to emphasize that $\boldsymbol{\Sigma}$ depends on $\boldsymbol{\theta}$, we shall write $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

4.2 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.

4.2.1 Unstructured Σ

If all elements of $\boldsymbol{Y}_i^*$ and $\boldsymbol{X}_i^*$, and q_i were observed, the likelihood is shown by

$$\ell = \prod_{i=1}^{n} \mathrm{M}(q_i) \frac{1}{\sqrt{2\pi}^T} \exp(-\frac{1}{2} q_i \boldsymbol{e}_i^{*\prime} \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_i^*)$$

and the log-likelihood λ is

$$\begin{aligned} \lambda &= \log \ell \\ &= \text{Const.} - \frac{n}{2} \log \mid \boldsymbol{\Sigma} \mid - \frac{1}{2} \sum_{i=1}^{n} \{\mathrm{tr}(q_i \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_i^* \boldsymbol{e}_i^{*\prime})\}. \end{aligned} \tag{27}$$

Let $\boldsymbol{e}_i^*$ be partitioned into two subvectors for each subject,

$$\boldsymbol{e}_i^* = \begin{pmatrix} \boldsymbol{e}_i^{(1)} \\ \boldsymbol{e}_i^{(2)} \end{pmatrix} \quad i = 1, ..., n,$$

where we assume $\boldsymbol{e}_i^{(2)}$ is unobserved and $\boldsymbol{e}_i^{(1)}$ is observed, then $\boldsymbol{e}_i^{(1)} = \boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}$, and also

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_{11i} & \boldsymbol{\Sigma}_{12i} \\ \boldsymbol{\Sigma}_{21i} & \boldsymbol{\Sigma}_{22i} \end{pmatrix}$$

is similarly into submatrices. Then λ is

$$\begin{aligned} \lambda &= \text{Const.} - \frac{1}{2} \sum_{i=1}^{n} \log \mid \boldsymbol{\Sigma}_{11i} \mid - \frac{1}{2} \sum_{i=1}^{n} q_i (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta})' \boldsymbol{\Sigma}_{11i}^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \\ &\quad - \frac{1}{2} \sum_{i=1}^{n} \log \mid \tilde{\boldsymbol{\Sigma}}_{22i} \mid \\ &\quad - \frac{1}{2} \sum_{i=1}^{n} q_i \{\boldsymbol{e}_i^{(2)} - \boldsymbol{\Sigma}_{21i} \boldsymbol{\Sigma}_{11i}^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta})\}' \tilde{\boldsymbol{\Sigma}}_{22i}^{-1} \{\boldsymbol{e}_i^{(2)} - \boldsymbol{\Sigma}_{21i} \boldsymbol{\Sigma}_{11i}^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta})\} \end{aligned} \tag{28}$$

where

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

By (27), we get an estimate of $\boldsymbol{\Sigma}$,

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} q_i \boldsymbol{e}_i^* \boldsymbol{e}_i^{*\prime}. \tag{29}$$

For the regression parameter vector $\boldsymbol{\beta}$, differentiating (28) with respect to $\boldsymbol{\beta}$, we have

$$\begin{aligned}\frac{\partial \lambda}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} q_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)} \\ &\quad - \sum_{i=1}^{n} q_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{\Sigma}_{12i} \tilde{\boldsymbol{\Sigma}}_{22i}^{-1} \{\boldsymbol{e}_i^{(2)} - \boldsymbol{\Sigma}_{21i} \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)}\}.\end{aligned}$$

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

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

$$w_i = \mathrm{E}(q_i \mid \boldsymbol{e}_i^{(1)}), \tag{30}$$

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

$$\begin{aligned}\mathrm{E}(q_i \boldsymbol{e}_i^{(2)} \mid \boldsymbol{e}_i^{(1)}) &= \mathrm{E}\{q_i \mathrm{E}(\boldsymbol{e}_i^{(2)} \mid q_i, \boldsymbol{e}_i^{(1)}) \mid \boldsymbol{e}_i^{(1)}\} \\ &= w_i \hat{\boldsymbol{e}}_i^{(2)}, \qquad (31) \\ \mathrm{E}(q_i \boldsymbol{e}_i^{(2)} \boldsymbol{e}_i^{(2)\prime} \mid \boldsymbol{e}_i^{(1)}) &= \mathrm{E}\{q_i \mathrm{E}(\boldsymbol{e}_i^{(2)} \boldsymbol{e}_i^{(2)\prime} \mid q_i, \boldsymbol{e}_i^{(1)}) \mid \boldsymbol{e}_i^{(1)}\} \\ &= w_i \hat{\boldsymbol{e}}_i^{(2)} \hat{\boldsymbol{e}}_i^{(2)\prime} + \tilde{\boldsymbol{\Sigma}}_{22i}, \qquad (32)\end{aligned}$$

where

$$\hat{\boldsymbol{e}}_i^{(2)} = \boldsymbol{\Sigma}_{21i} \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)}.$$

Then

$$\mathrm{E}\left(\frac{\partial \lambda}{\partial \boldsymbol{\beta}} \mid \boldsymbol{e}_1^{(1)}, \ldots, \boldsymbol{e}_n^{(1)}\right) = \sum_{i=1}^{n} w_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)},$$

therefore

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} w_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{n} w_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{Y}_i, \tag{33}$$

and from (29) and (32),

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \{w_i \hat{\boldsymbol{e}}_i^* \hat{\boldsymbol{e}}_i^{*\prime} + \tilde{\boldsymbol{\Sigma}}_{22i}\}. \tag{34}$$

where

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

We can summarize the EM algorithm as follows;

E-step: to calculate the conditional expectations (30) , (31) and (32).

M-step: to renew the estimates by (33) and (34).

Example 7 : 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 $\boldsymbol{e}_i$ 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

$$d_i^2 = (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_{i11}^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}).$$

Example 8 : Multivariate t case

When $q_i \times \nu$ has the chi-squared distribution with ν degrees of freedom, the marginal distribution of $\boldsymbol{e}_i$ is the multivariate t distribution and

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

4.2.2 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 $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i$, where conditional on q_i, $\boldsymbol{e}_i$ is distributed as $\mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i(\boldsymbol{\theta})/q_i)$.

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 (30), (31) and (32)

(ii) Compute updated estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, using equation (33) and

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

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

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

$$[G]_i = \frac{1}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i,$$

$$[H]_{ij} = \frac{1}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_j,$$

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

4.2.3 Random-effects model

The random-effects model is $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown population parameters, $\boldsymbol{b}_i$ is a $k \times 1$ vector of unknown individual effects and $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are a known $t_i \times p$ design matrix linking $\boldsymbol{\beta}$ to $\boldsymbol{Y}_i$ and a $t_i \times k$ one linking $\boldsymbol{b}_i$ to $\boldsymbol{Y}_i$, respectively. Conditional on q_i, $\boldsymbol{e}_i \sim \mathrm{N}(0, \sigma^2/q_i\,\mathrm{I}_{t_i})$ and $\boldsymbol{b}_i \sim \mathrm{N}(0, \boldsymbol{D}/q_i)$, that is,

$$\begin{pmatrix} \boldsymbol{Y}_i \\ \boldsymbol{b}_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ \boldsymbol{0} \end{pmatrix}, \frac{1}{q_i}\begin{pmatrix} \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \sigma^2\mathrm{I}_{t_i} & \boldsymbol{Z}_i\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{Z}_i' & \boldsymbol{D} \end{pmatrix}\right). \tag{35}$$

This model is a special case of the structured $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \sigma^2\mathrm{I}_{t_i}$. But in this case treating $\boldsymbol{b}_i$ and q_i as missing data, we give the method based on the EM algorithm in order to obtain the maximum likelihood estimators.

If $\boldsymbol{b}_i$ and q_i were observed,

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} q_i \boldsymbol{X}_i' \boldsymbol{X}_i)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i' (\boldsymbol{Y}_i - \boldsymbol{Z}_i \boldsymbol{b}_i), \tag{36}$$

$$\hat{\sigma} = \frac{1}{\sum t_i} \sum_{i=1}^{n} q_i (\boldsymbol{Y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i \boldsymbol{b}_i)' (\boldsymbol{Y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i \boldsymbol{b}_i) \tag{37}$$

and

$$\hat{\boldsymbol{D}} = \frac{1}{n} \sum_{i=1}^{n} q_i \boldsymbol{b}_i \boldsymbol{b}_i'. \tag{38}$$

In this case, we have to calculate the following conditional expectations given $\boldsymbol{Y}_i$;

$$w_i = \mathrm{E}(q_i \mid \boldsymbol{Y}_i), \tag{39}$$

$$\begin{aligned} \mathrm{E}(q_i \boldsymbol{b}_i \mid \boldsymbol{Y}_i) &= \mathrm{E}\{q_i \mathrm{E}(\boldsymbol{b}_i \mid q_i, \boldsymbol{Y}_i) \mid \boldsymbol{Y}_i\} \\ &= w_i \hat{\boldsymbol{b}}_i, \end{aligned} \tag{40}$$

$$\begin{aligned} \mathrm{E}(q_i \boldsymbol{b}_i \boldsymbol{b}_i' \mid \boldsymbol{Y}_i) &= \mathrm{E}\{q_i \mathrm{E}(\boldsymbol{b}_i \boldsymbol{b}_i' \mid q_i, \boldsymbol{Y}_i) \mid \boldsymbol{Y}_i\} \\ &= w_i \hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i' + \tilde{\boldsymbol{D}}, \end{aligned} \tag{41}$$

and

$$\begin{aligned} \mathrm{E}(q_i \boldsymbol{e}_i' \boldsymbol{e}_i \mid \boldsymbol{Y}_i) &= \mathrm{E}\{\mathrm{tr}(q_i \boldsymbol{e}_i \boldsymbol{e}_i') \mid \boldsymbol{Y}_i\} \\ &= \mathrm{tr}[\mathrm{E}\{q_i \mathrm{E}(\boldsymbol{e}_i \boldsymbol{e}_i' \mid q_i, \boldsymbol{Y}_i) \mid \boldsymbol{Y}_i\}] \\ &= \mathrm{tr}\{w_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i' + \boldsymbol{C}\}, \end{aligned} \tag{42}$$

where

$$\hat{\boldsymbol{b}}_i = \boldsymbol{D} \boldsymbol{Z}_i' (\boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i + \sigma^2 \mathrm{I}_{t_i})^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}),$$

$$\tilde{\boldsymbol{D}} = \boldsymbol{D} - \boldsymbol{D} \boldsymbol{Z}_i' (\boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \sigma^2 \mathrm{I}_{t_i})^{-1} \boldsymbol{Z}_i \boldsymbol{D},$$

$$\hat{\boldsymbol{e}} = \boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{Z}_i \hat{\boldsymbol{b}}_i, \quad \boldsymbol{C} = \boldsymbol{Z}_i \tilde{\boldsymbol{D}} \boldsymbol{Z}_i'.$$

The EM algorithm can be summarized in the following way.

E-step: to calculate (39),(40),(41) and (42).

M-step: to renew the estimates by (36), (37) and (38).

4.2.4 Asymptotic variance

We give the asymptotic variance of $\hat{\boldsymbol{\beta}}$, which is used in the test for the regression parameters $\boldsymbol{\beta}$ 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

$$\begin{aligned} \imath &= -\sum_{i=1}^{n} \mathrm{E}\left(\frac{\partial^2 \lambda_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \mid \boldsymbol{e}_i^{(1)}\right) - \sum_{i=1}^{n} \mathrm{E}\left\{\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}\right)' \mid \boldsymbol{e}_i^{(1)}\right\} \\ &\quad - \sum_{i \neq j} \mathrm{E}\left\{\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}\right)' \mid e_i^{(1)}\right\} \mathrm{E}\left\{\left(\frac{\partial \lambda_j}{\partial \boldsymbol{\beta}}\right)' \mid e_j^{(1)}\right\}', \end{aligned}$$

where

$$\lambda_i = -\frac{1}{2} \log \mid \boldsymbol{\Sigma} \mid -\frac{1}{2} \mathrm{tr}(q_i \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_i^* \boldsymbol{e}_i^{*\prime}).$$

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

$$\mathrm{E}\left(\frac{\partial^2 \lambda_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \mid \boldsymbol{e}_i^{(1)}\right) = -w_i \{\boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i + \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{\Sigma}_{12i} \tilde{\boldsymbol{\Sigma}}_{22i}^{-1} \boldsymbol{\Sigma}_{21i} \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i\},$$

and

$$\mathrm{E}\left\{\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}\right)' \mid \boldsymbol{e}_i^{(1)}\right\} = w_i^* \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)} \boldsymbol{e}_i^{(1)\prime} \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i + w_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{\Sigma}_{12i} \tilde{\boldsymbol{\Sigma}}_{22i}^{-1} \boldsymbol{\Sigma}_{21i} \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i,$$

where $w_i^* = \mathrm{E}(q_i^2 \mid \boldsymbol{e}_i^{(1)})$. Thus

$$-\mathrm{E}\left(\frac{\partial^2 \lambda_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \mid \boldsymbol{e}_i^{(1)}\right) - \mathrm{E}\left\{\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}\right)' \mid \boldsymbol{e}_i^{(1)}\right\} = w_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i - w_i^* \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)} \boldsymbol{e}_i^{(1)\prime} \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i$$

and

$$\mathrm{E}\left(\frac{\partial \lambda_i}{\partial \boldsymbol{\beta}}\right)' \mid \boldsymbol{e}_i^{(1)}) = w_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)}$$

Therefore

$$\imath = \sum_{i=1}^{n} \{w_i \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i - w_i^* \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)} \boldsymbol{e}_i^{(1)\prime} \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{X}_i\} - \sum_{i \neq j} w_i w_j \boldsymbol{X}_i' \boldsymbol{\Sigma}_{11i}^{-1} \boldsymbol{e}_i^{(1)} \boldsymbol{e}_j^{(1)\prime} \boldsymbol{\Sigma}_{11j}^{-1} \boldsymbol{X}_j.$$

4.3 Numerical examples

The study of growth is an important topic in many biological sciences, and in statistics the term growth curve has a special meaning (see Ware, 1983). Growth curve analysis is one of the most popular example of analysis of repeated measures data.

We analyzed two real growth data sets. One is Potthoff and Roy(1964)'s data, also analyzed by Jennrich and Schluchter(1986) under the normality assumption and another is Elston and Grizzle (1962)'s data.

4.3.1 Potthoff and Roy's data

Potthoff and Roy gave a set of growth data for 11 girls and 16 boys, listed in Table 2. 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.

Table 2 : Potthoff and Roy's data

	Age in years				Age in years			
Sex	Girls				Boys			
Subjects	8	10	12	14	8	10	12	14
1	21.0	20.0	21.5	23.0	26.0	25.0	29.0	31.0
2	21.0	21.5	24.0	25.5	21.5	(22.5)	23.0	26.5
3	20.5	(24.0)	24.5	26.0	23.0	22.5	24.0	27.5
4	23.5	24.5	25.0	26.5	25.5	27.5	26.5	27.0
5	21.5	23.0	22.5	23.5	20.0	(23.5)	22.5	26.0
6	20.0	(21.0)	21.0	22.5	24.5	25.5	27.0	28.5
7	21.5	22.5	23.0	25.0	22.0	22.0	24.5	26.5
8	23.0	23.0	23.5	24.0	24.0	21.5	24.5	25.5
9	20.0	(21.0)	22.0	21.5	23.0	20.5	31.0	26.0
10	16.5	(19.0)	19.0	19.5	27.5	28.0	31.0	31.5
11	24.5	25.0	28.0	28.0	23.0	23.0	23.5	25.0
12					21.5	(23.5)	24.0	28.0
13					17.0	(24.5)	26.0	29.5
14					22.5	25.5	25.5	26.0
15					23.0	24.5	26.0	30.0
16					22.0	(21.5)	23.5	25.0

The values in parentheses are treated as missing values in Little and Rubin (1987).

Table 3 : *Summary of models fit*

Model	Distribution	$\boldsymbol{\Sigma}$	The number of parametres	-2× log likelihood	*AIC*
(1)	MN	$\boldsymbol{\Sigma}_g = (\sigma_{ij})$	14	419.48	447.48
(2)	MN	$\boldsymbol{\Sigma}_g = (\sigma_{gij})$	24	395.02	443.02
(3)	MN	$\boldsymbol{\Sigma}_g = \sigma_1^2 \mathbf{1}\mathbf{1}' + \sigma_2^2 \mathbf{I}$	6	428.64	430.64
(4)	MN	$\boldsymbol{\Sigma}_g = \sigma_{1g}^2 \mathbf{1}\mathbf{1}' + \sigma_{2g}^2 \mathbf{I}$	8	408.81	424.81
(5)	MT	$\boldsymbol{\Sigma}_g = (\sigma_{ij})$	15	407.52	437.52
(6)	MT	$\boldsymbol{\Sigma}_g = (\sigma_{gij})$	25	395.5*	
(7)	MT	$\boldsymbol{\Sigma}_g = \sigma_1^2 \mathbf{1}\mathbf{1}' + \sigma_2^2 \mathbf{I}$	7	414.43	428.43
(8)	MT	$\boldsymbol{\Sigma}_g = \sigma_{1g}^2 \mathbf{1}\mathbf{1}' + \sigma_{2g}^2 \mathbf{I}$	9	405.75	423.75
(9)	CN	$\boldsymbol{\Sigma}_g = (\sigma_{ij})$	16	399.30	431.30
(10)	CN	$\boldsymbol{\Sigma}_g = (\sigma_{gij})$	26	387.47	439.47
(11)	CN	$\boldsymbol{\Sigma}_g = \sigma_1^2 \mathbf{1}\mathbf{1}' + \sigma_2^2 \mathbf{I}$	8	409.00	425.00
(12)	CN	$\boldsymbol{\Sigma}_g = \sigma_{1g}^2 \mathbf{1}\mathbf{1}' + \sigma_{2g}^2 \mathbf{I}$	10	402.97	422.97

MN : Multivariate Normal, MT : Maltivariate t, CN : Contaminated multivariate normal.

* The degrees of freedom is 20. The MLE of it is more than 20.

Table 4 : *Maximum Likelihood Estimates*

Model	Sex	Age	Means(S.E.)	Covariance matrix			
		8	21.21(.62)	4.470	3.880	3.880	3.880
	Girls	10	22.17(.61)	3.880	4.470	3.880	3.880
(4)		12	23.13(.61)	3.880	3.880	4.470	3.880
		14	24.09(.62)	3.880	3.880	3.880	4.470
		8	22.62(.52)	5.204	2.446	2.446	2.446
	Boys	10	24.18(.45)	2.446	5.204	2.446	2.446
		12	25.75(.45)	2.446	2.446	5.204	2.446
		14	27.32(.52)	2.446	2.446	2.446	5.204
		8	21.21(.63)	4.315	3.743	3.743	3.743
	Girls	10	22.16(.61)	3.743	4.315	3.743	3.743
(12)		12	23.12(.61)	3.743	3.743	4.315	3.743
		14	24.07(.63)	3.743	3.743	3.743	4.315
$\hat{\delta} = 0.094$		8	22.87(.51)	4.012	2.667	2.667	2.667
$\hat{\lambda} = 0.158$	Boys	10	24.28(.47)	2.667	4.012	2.667	2.667
		12	25.69(.47)	2.667	2.667	4.012	2.667
		14	27.10(.51)	2.667	2.667	2.667	4.012

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 a model for $\boldsymbol{\Sigma}$. In calculations, we considered a method to have converged when all parameters differed by less than 0.01% between successive iterations.

Table 3 shows the -2×log-likelihoods under several assumptions and the numbers of parameters involved in the models, and the results of estimation are shown in Table

4. The standard errors were obtained from the results in Section 4.2.4. The value of degrees of freedom ν for the multivariate t-distribution models, or, δ and γ for the contaminated normal model were estimated simultaneously with parameters for mean vector and covariance matrix, using the method mentioned in Section 3.4. 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.

Table 5 : *Conditional Expectations* ; w_i

Contaminated Multivariate Normal Distribution
(Σ : Compound Symmetry)

Girls	.988	.983	.933	.996	.994	.996	.997	.991	.992	.961
	.947									
Boys	.961	.994	.992	.971	.985	.997	.996	.969	.158	.947
	.992	.992	.160	.995	.987	.993				

The conditional expectation of q_i, w_i gives us the information of the degree of suspicion of outliers. Table 5 shows the values of w_i under the selected model. We wonder if the 9th and the 13th boys data are outliers based on this. Table 6 shows the result for the data without the 9th and the 13th boys data. This result is more similar to that under the selected model than to those under the other models.

Pendergast and Broffitt (1985) applied a semi-parametric method for robust estimation to this data and pointed out that the above two subjects were considered as outliers. They used the model with common unstructured covariance matrix. When we have some models to be fitted, we have to perform model selections. When we have to select one model from many models, it is more convenient that models are fully parametric than that those are semi- or non-parametric.

Table 6

Result after omitting the 9-th and the 13-th boys data

Model	Sex	Age	Means(SE)	Covariance Matrix			
		8	21.21(.62)	4.476	3.517	3.517	3.517
	Girls	10	22.17(.59)	3.517	4.476	3.517	3.517
		12	23.13(.59)	3.517	3.517	4.476	3.517
(3)		14	24.09(.62)	3.517	3.517	3.517	4.476
-2×		8	22.95(.53)				
Log-likelihood	Boys	10	24.32(.52)				
=348.4		12	25.70(.52)				
		14	27.07(.53)				
		8	21.21(.62)	4.470	3.880	3.880	3.880
	Girls	10	22.17(.61)	3.880	4.470	3.880	3.880
		12	23.13(.61)	3.880	3.880	4.470	3.880
(4)		14	24.09(.62)	3.880	3.880	3.880	4.470
-2×		8	22.92(.54)	4.480	3.231	3.231	3.231
Log-likelihood	Boys	10	24.32(.51)	3.231	4.480	3.231	3.231
=343.4		12	25.70(.51)	3.231	3.231	4.480	3.231
		14	27.07(.54)	3.231	3.231	3.231	4.480

4.3.2 Elston and Grizzle's data

Elston and Grizzle gave a simple example in which the ramus height of 20 boys was measured at 8, 8.5, 9, and 9.5 years. Table 7 lists the data used here.

The data show a steady growth with age, and we would like to fit a straight line. Let the 4×1 vector $\boldsymbol{Y}_i$ be the observations from the i-th subject and

$$\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{X}\boldsymbol{\beta}.$$

To fit a linear growth model to the data on ramus height, the matrix $\boldsymbol{X}$ is 4×2 and can be written

$$\boldsymbol{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 8.5 & 9 & 9.5 \end{pmatrix},$$

and $\boldsymbol{\beta}$ is the 2×1 vector of regression parameters. As the model of the error distribution, the multivariate normal, multivariate t, and contaminated multivariate normal distributions are used.

Table 7 : Elston and Grizzle data (Ramus height of 20 boys)

	Age in years			
Subjects	8	8.5	9	9.5
1	47.8	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.5
4	45.1	45.3	46.1	47.2
5	47.6	48.5	48.9	49.3
6	52.5	53.2	53.3	53.7
7	51.2	53.0	54.3	54.5
8	49.8	50.0	50.3	52.7
9	48.1	50.8	52.3	54.4
10	45.0	47.0	47.3	48.3
11	51.2	51.4	51.6	51.9
12	48.5	49.2	53.0	55.5
13	52.1	52.8	53.7	55.0
14	48.2	48.9	49.3	49.8
15	49.6	50.4	51.2	51.8
16	50.7	51.7	52.7	53.3
17	47.2	47.7	48.4	49.5
18	53.3	54.6	55.1	55.3
19	46.2	47.5	48.1	48.4
20	46.3	47.6	51.3	51.8

Table 8 and Table 9 summarize the results. The contaminated model is considered the best model among our models.

Table 8 : *Summary of models fit*

Error distribution	Model for Σ	Number of parameters	-2×log-likelihood
Normal	Unstructured	12	224.46
t distr.($\nu = 12$)	Unstructured	13	223.13
Contam.$\delta = 0.272, \lambda = 0.076$	Unstructured	14	205.93

Table 9 : *Maximum Likelihood Estimates*

Model			Estimated age specific means(SE)	Estimated covariance matrix			
Distribution	Model for Σ	Age					
Normal	Unstructured	8	48.65(.30)	6.014	5.880	5.488	5.271
		8.5	49.58(.29)	5.880	6.129	5.848	5.627
		9	50.52(.30)	5.488	5.848	6.575	6.599
		9.5	51.45(.34)	5.271	5.627	6.599	7.092
t-distr. $\nu = 18$	Unstructured	8	48.66(.33)	5.985	5.916	5.656	5.439
		8.5	49.52(.32)	5.916	6.163	5.988	5.750
		9	50.39(.33)	5.656	5.988	6.547	6.497
		9.5	51.26(.36)	5.439	5.750	6.497	6.818
Contam. $\delta = 0.272$ $\lambda = 0.076$	Unstructured	8	48.70(.30)	4.870	5.065	4.983	4.762
		8.5	49.43(.30)	5.065	5.405	5.351	5.078
		9	50.16(.30)	4.983	5.351	5.458	5.234
		9.5	50.89(.32)	4.762	5.078	5.234	5.128

Table 10 : *Conditional Expectations* ; w_i

Contaminated Multivariate Normal Distribution
(Σ : Unstructured)

.987	.992	.993	.988	.995	.980	.830	.076	.076	.076
.953	.076	.940	.997	.998	.996	.996	.970	.971	.076

This model suggests us that 20 boys might be classsified into two groups such that

one group consists of five subjects 8, 9, 10, 12, and 20 and another consists of the remainders. One of the most distinctive feature of these five subjects is the rapid growth, in particular, 7.0 of the 12th subject and 6.3 of the 9th subject, and at least, more than 3.1.

4.3.3 Incomplete data

Little and Rubin (1987) fitted Jennrich and Schluchter's model to the data obtained by deleting the nine values in parentheses in Table 2. Their deletion mechanism is designed to be *missing at random* (MAR) but not *missing completely at random* (MCAR). Specifically, for each gender, values at age 10 are deleted for cases with low values at age 8.

We used two methods for treating missing observaions : (i) the methods for incomplete data given in Section 3.2 ; (ii) ML methods using only the 18 complete observations, assuming the following three distributions as the error distribuions: the multivariate normal (MN), multivariate t with unknown degrees of freedom (MT) and contaminated multivariate normal distribution with two unknown parameters (MC). In this example, we only use common covariance matrices model such that $\boldsymbol{\Sigma} = \sigma_1^2 \mathbf{1}\mathbf{1}' + \sigma_2^2 \mathbf{I}$. Table 11 shows the results from the data with incomplete observations and Table 12 shows those from the data without incomplete observations. As the deletion mechanism is not MCAR, results from the data without incomplete observations are seriously biased such that means are overestimated.

Table 11 : Results with incomplete observations

Distribution			Mean(SE)	Covariance Matrix			
		8	21.13(.65)	5.113	3.095	3.095	3.095
	Girls	10	22.11(.59)	3.095	5.113	3.095	3.095
MN		12	23.09(.58)	3.095	3.095	5.113	3.095
		14	24.07(.64)	3.095	3.095	3.095	5.113
		8	22.60(.54)				
-2×	Boys	10	24.17(.49)				
log likelihood		12	25.74(.48)				
=401.31		14	27.32(.53)				
		8	21.25(.59)	3.833	2.739	2.739	2.739
	Girls	10	22.18(.56)	2.739	3.833	2.739	2.739
MT		12	23.12(.57)	2.739	2.739	3.833	2.739
		14	24.06(.61)	2.739	2.739	2.739	3.833
$\widehat{\nu} = 5.02$		8	22.71(.56)				
-2×	Boys	10	24.12(.52)				
log likelihood		12	25.54(.52)				
=388.57		14	26.96(.56)				
		8	21.12(.61)	4.286	3.260	3.260	3.260
	Girls	10	22.10(.58)	3.260	4.286	3.260	3.260
CN		12	23.09(.58)	3.260	3.260	4.286	3.260
$\widehat{\delta} = 0.0854$		14	24.07(.61)	3.260	3.260	3.260	4.286
$\widehat{\lambda} = 0.099$		8	22.91(.54)				
-2×	Boys	10	24.31(.51)				
log likelihood		12	25.70(.51)				
=383.15		14	27.10(.54)				

Table 12 : Results without incomplete observations

Distribution			Mean(SE)	Covariance Matrix			
		8	22.09(.72)	4.114	2.361	2.361	2.361
	Girls	10	23.04(.64)	2.361	4.114	2.361	2.361
MN		12	23.99(.64)	2.361	2.361	4.114	2.361
		14	24.94(.72)	2.361	2.361	2.361	4.114
		8	23.58(.57)				
-2×	Boys	10	24.92(.51)				
log likelihood		12	26.28(.51)				
=278.13		14	27.63(.57)				
		8	22.07(.62)	3.066	2.117	2.117	2.117
	Girls	10	22.99(.58)	2.117	3.066	2.117	2.117
MT		12	23.91(.59)	2.117	2.117	3.066	2.117
		14	24.82(.65)	2.117	2.117	2.117	3.066
$\widehat{\nu} = 5.11$		8	23.57(.60)				
-2×	Boys	10	24.86(.57)				
log likelihood		12	26.16(.58)				
=268.28		14	27.45(.64)				
		8	22.09(.68)	3.515	2.507	2.507	2.507
	Girls	10	23.04(.64)	2.507	3.515	2.507	2.507
CN		12	23.99(.64)	2.507	2.507	3.515	2.507
$\widehat{\delta} = 0.063$		14	24.94(.68)	2.507	2.507	2.507	3.515
$\widehat{\lambda} = 0.100$		8	23.70(.57)				
-2×	Boys	10	24.99(.53)				
log likelihood		12	26.29(.53)				
=265.37		14	27.59(.57)				

5 Robust Factor Analysis

Factor analysis is a branch of multivariate analysis that is concerned with the internal relationships of a set of variables when these relationships can be taken to be linear, or approximately so. Initially, factor analysis was developed by psychometricians and in the early days approximate methods of estimation only were available, of which the most celebrated was the *centroid* or simple summation method. The *principal factor* and *minres* methods are more recent approximate methods. (see Harman, 1967 and his references) Efficient estimation procedures were based on the method of maximum likelihood (Lawley and Maxwell, 1963). Difficulties of a computational nature were experienced, however, and it was not until the advent of electronic computers and a new approach to the solution of the basic equations by Jöreskog(1967) that the maximum likelihood approach became a feasible proposition.

An alternative approach to calculating ML estimates was suggested by Dempster *et al.* (1977) and has been examined further by Rubin and Thayer (1982, 1983) and Bentler and Tanaka (1983). Its use depends on the fact that if we could observe the factor scores we could estimate the parameters by regression methods.

Bentler and Tanaka pointed out some problems on Rubin and Thayer's example and the EM algorithm for ML factor analysis. As concerns slowness of convergence of the EM algorithm, it does not seem to be so important in recent highly developed circumstance of computers. On the other hand, the problem of example of Rubin and Thayer seems to be due to insufficient implement of the algorithm. In particular it is important what kind of criterion for convergence is selected. Applying the EM algorithm we have to use strict criterion for convergence, which was also noted by Bentler and Tanaka, because small renewal of parameters is performed by one step of iteration when the model includes a large number of parameters.

The ordinary ML factor analysis is based on the assumption that the observations follow the multivariate normal distribution. It is well-known that the analysis under

the normality assumption is sensitive to outliers. In fact, in practical applications of factor analysis, we often meet the cases that the normality assumption is inappropriate because the data include some extreme observations. Rubin and Thayer (1982) mentioned that, after deriving the ML method for factor analysis under the multivariate normal assumption, " the entire issue of the sensitivity of results to the assumption of multivariate normality is important for the wise application of the technique in practice".

Also in this case we replace the multivariate normal distribution by scale mixtures of multivariate normal distributions in order to reduce the influence of outliers and then derive a robust method in ML factor analysis. Rubin and Thayer (1982) presented equations to implement ML factor analysis via the EM algorithm treating factor scores as missing data. In this case, besides the factor scores, we cannot observe the mixing variables. Therefore we have to construct new algorithm for the estimation by applying the EM algorithm.

5.1 Robust model

Suppose $\boldsymbol{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{Z}_i + \boldsymbol{e}_i \quad (i = 1, ..., n)$, where $\boldsymbol{Y}_i$ is an observed p-component vector, $\boldsymbol{Z}_i$ is an unobserved m-component vector of factor-scores and $\boldsymbol{e}_i$ is a vector of errors (or errors plus specific factors). $\boldsymbol{\alpha}$ is a vector of means and the $p \times m$ matrix $\boldsymbol{\beta}$ consists of factor loadings.

In this paper, we use scale mixtures of multivariate normal distributions instead of the normality assumption, considering following two typical backgrounds. In employing such heavier-tailed symmetric distributions as underlying distributions, we consider two practical possibilities as follows;

1: Case that a group is not homogeneous from the beginning, and the existence of partial subjects that have abnormal capability is supposed. Namely, both of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are assumed to follow scale mixtures of normal distributions.

2: Case that the latent ability itself of a group is homogeneous, but at the point of time when manifest response $\boldsymbol{Y}$ is observed, outliers mix. Originally, since specific factors are the result of the mixture of many factors including errors, as the assumption for the distribution of specific factors, to apply scale mixtures of normal distributions is more realistic rather than normal distributions.

Concrete statistical models based on the above two cases are as follows:

Model 1:

We assume that conditional on unobserved q_i, $\boldsymbol{e}_i$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}/q_i$ and $\boldsymbol{Z}_i$ is also normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}/q_i$, and that $\boldsymbol{e}_i$ and $\boldsymbol{Z}_i$ are mutually independent, where $\boldsymbol{\Psi}$ is a diagonal matrix and $\mathbf{I}$ is the unit matrix. q_i is a positive random variable with the probability (density) function $M(q_i)$.

Then conditional on q_i ,

$$\begin{pmatrix} \boldsymbol{Y}_i \\ \boldsymbol{Z}_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix}, \Sigma^{(1)}/q_i \right),$$

where

$$\boldsymbol{\Sigma}^{(1)} = \begin{pmatrix} \boldsymbol{\Sigma}^{(1)}_{YY} & \boldsymbol{\Sigma}^{(1)}_{YZ} \\ \boldsymbol{\Sigma}^{(1)}_{ZY} & \boldsymbol{\Sigma}^{(1)}_{ZZ} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Psi} & \boldsymbol{\beta} \\ \boldsymbol{\beta}' & \mathbf{I} \end{pmatrix}.$$

Model 2:

$\boldsymbol{Z}_i$ is, independently of q_i, normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}$. Conditional on q_i, $\boldsymbol{e}_i$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}/q_i$. Thus conditional on q_i,

$$\begin{pmatrix} \boldsymbol{Y}_i \\ \boldsymbol{Z}_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{\Sigma}^{(2)}_i \right),$$

where

$$\boldsymbol{\Sigma}^{(2)}_i = \begin{pmatrix} \boldsymbol{\Sigma}^{(2)}_{YY_i} & \boldsymbol{\Sigma}^{(2)}_{YZ_i} \\ \boldsymbol{\Sigma}^{(2)}_{ZY_i} & \boldsymbol{\Sigma}^{(2)}_{ZZ_i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Psi}/q_i & \boldsymbol{\beta} \\ \boldsymbol{\beta}' & \mathbf{I} \end{pmatrix}.$$

5.2 Estimation of the parameters

In this section, assuming the number of factors is known, we give the estimates of parameters by applying the EM-algorithm, treating q and $\boldsymbol{Z}$ as missing data, that iteratively maximizes the likelihood supposing q and $\boldsymbol{Z}$ were observed. First we consider the estimation under Model 1. The following lemma enables us to easily handle the log likelihood.

Lemma 4

$$| \boldsymbol{\Sigma}^{(1)} | = | \boldsymbol{\Psi} |,$$

$$| \boldsymbol{\Sigma}^{(2)} | = q^{-p} | \boldsymbol{\Psi} |,$$

$$\boldsymbol{\Sigma}^{(1)-1} = \begin{pmatrix} \boldsymbol{\Psi}^{-1} & -\boldsymbol{\Psi}^{-1}\boldsymbol{\beta} \\ -\boldsymbol{\beta}'\boldsymbol{\Psi}^{-1} & I_m + \boldsymbol{\beta}'\boldsymbol{\Psi}^{-1}\boldsymbol{\beta} \end{pmatrix},$$

$$\boldsymbol{\Sigma}^{(2)-1} = \begin{pmatrix} q\boldsymbol{\Psi}^{-1} & -q\boldsymbol{\Psi}^{-1}\boldsymbol{\beta} \\ -q\boldsymbol{\beta}'\boldsymbol{\Psi}^{-1} & I_m + q\boldsymbol{\beta}'\boldsymbol{\Psi}^{-1}\boldsymbol{\beta} \end{pmatrix}.$$

If $\boldsymbol{Z}$'s and q's are observed in addition $\boldsymbol{Y}$, the log likelihood ℓ is

$$\ell = \text{Const.} - \frac{n}{2}\log | \boldsymbol{\Psi} | - \frac{1}{2}\sum_{i=1} n\mathrm{tr}[q_i\{\boldsymbol{\Psi}^{-1}((\boldsymbol{Y}_i - \boldsymbol{\alpha})(\boldsymbol{Y}_i - \boldsymbol{\alpha})' - 2(\boldsymbol{Y}_i - \boldsymbol{\alpha})\boldsymbol{Z}_i'\boldsymbol{\beta}' + \boldsymbol{\beta}\boldsymbol{Z}_i\boldsymbol{Z}_i'\boldsymbol{\beta}')\}]$$

and the sufficient statistics are

$$\sum q_i, \quad \sum q_i\boldsymbol{Y}_i, \quad \sum q_i\boldsymbol{Z}_i, \quad \sum q_i\boldsymbol{Y}_i\boldsymbol{Y}_i', \quad \sum q_i\boldsymbol{Y}_i\boldsymbol{Z}_i'.$$

Let

$$\begin{pmatrix} \boldsymbol{S}_{YY} & \boldsymbol{S}_{YZ} \\ \boldsymbol{S}_{ZY} & \boldsymbol{S}_{ZZ} \end{pmatrix} = \begin{pmatrix} \sum q_i\boldsymbol{Y}_i\boldsymbol{Y}_i'/q_0 & \sum q_i\boldsymbol{Y}_i\boldsymbol{Z}_i'/q_0 \\ \sum q_i\boldsymbol{Z}_i\boldsymbol{Y}_i'/q_0 & \sum q_i\boldsymbol{Z}_i\boldsymbol{Z}_i'/q_0 \end{pmatrix},$$

$$\begin{pmatrix} \boldsymbol{C}_{YY} & \boldsymbol{C}_{YZ} \\ \boldsymbol{C}_{ZY} & \boldsymbol{C}_{ZZ} \end{pmatrix} = \begin{pmatrix} \boldsymbol{S}_{YY} - \bar{\boldsymbol{Y}}\bar{\boldsymbol{Y}}' & \boldsymbol{S}_{YZ} - \bar{\boldsymbol{Y}}\bar{\boldsymbol{Z}}' \\ \boldsymbol{S}_{ZY} - \bar{\boldsymbol{Z}}\bar{\boldsymbol{Y}}' & \boldsymbol{S}_{ZZ} - \bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}' \end{pmatrix},$$

where

$$\bar{\boldsymbol{Y}} = \sum q_i \boldsymbol{Y}_i / q_0 \ , \ \bar{\boldsymbol{Z}} = \sum q_i \boldsymbol{Z}_i / q_0 \ , \ q_0 = \sum q_i \ ,$$

then

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \bar{\boldsymbol{Y}} - \hat{\boldsymbol{\beta}} \bar{\boldsymbol{Z}}, \\ \hat{\boldsymbol{\beta}} &= \boldsymbol{C}_{YZ} \boldsymbol{C}_{ZZ}^{-1}, \\ \hat{\boldsymbol{\Psi}} &= \mathrm{diag}(\boldsymbol{C}_{YY} - \boldsymbol{C}_{YZ} \boldsymbol{C}_{ZZ}^{-1} \boldsymbol{C}_{ZY}) q_0 / n. \end{aligned} \tag{43}$$

We, however, cannot observe q's and $\boldsymbol{Z}$'s. Thus we must calculate the conditional expectations of above sufficient statistics given $\boldsymbol{Y}$'s.

[E-step]

We give the conditional expectations of the sufficient statistics given $\boldsymbol{Y}$'s as following; Frist we let

$$w_i = \mathrm{E}(q_i \mid \boldsymbol{Y}_i),$$

where the specific form of w_i depends on the model for the distribution of q_i (i.e. M(.), and see examples.)

We note that in this case we could regard w_i as the weight of $\boldsymbol{Y}_i$ in following procedure. Therefore, we could easily find the extreme observations by checking w_i (see the result of Mardia *et al.* (1979)'s data in Section 5.4).

$$\begin{aligned} \mathrm{E}(q_i \boldsymbol{Z}_i \mid \boldsymbol{Y}_i) &= \mathrm{E}\{q_i \mathrm{E}(\boldsymbol{Z}_i \mid q_i, \boldsymbol{Y}_i) \mid \boldsymbol{Y}_i\} \\ &= w_i \hat{\boldsymbol{Z}}_i, \end{aligned}$$

since the conditional expectation of $\boldsymbol{Z}$ given q and $\boldsymbol{Y}$ does not depend on q.

$$\begin{aligned} \mathrm{E}(q_i \boldsymbol{Z}_i \boldsymbol{Z}_i' \mid \boldsymbol{Y}_i) &= \mathrm{E}\{q_i \mathrm{E}(\boldsymbol{Z}_i \boldsymbol{Z}_i' \mid q_i, \boldsymbol{Y}_i) \mid \boldsymbol{Y}_i\} \\ &= \mathrm{E}\{q_i (\hat{\boldsymbol{Z}}_i \hat{\boldsymbol{Z}}_i{}' + \mathrm{Cov}(\boldsymbol{Z}_i \mid q_i, \boldsymbol{Y}_i)) \mid \boldsymbol{Y}_i\} \\ &= w_i \hat{\boldsymbol{Z}}_i \hat{\boldsymbol{Z}}_i{}' + \boldsymbol{\Sigma}_{ZZ}^{*} \end{aligned}$$

where

$$\hat{\boldsymbol{Z}}_i = \boldsymbol{\Sigma}_{ZY}^{(1)} \boldsymbol{\Sigma}_{YY}^{(1)-1} (\boldsymbol{Y}_i - \boldsymbol{\alpha}),$$
$$\boldsymbol{\Sigma}_{ZZ}^{*} = \boldsymbol{\Sigma}_{ZZ}^{(1)} - \boldsymbol{\Sigma}_{ZY}^{(1)} \boldsymbol{\Sigma}_{YY}^{(1)-1} \boldsymbol{\Sigma}_{YZ}^{(1)}.$$

[M-step]

We compute the update estimates with the equations (43) replaced by their conditional expectations from E-step.

We would get the ML estimates applying repeatedly E-step and M-step until convergence.

Example 9

To consider the contaminated multivariate normal case, let

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

then

$$w_i = \frac{1 - \delta + \delta\lambda^{1+p/2} \exp\{(1-\lambda)d_i^2/2\}}{1 - \delta + \delta\lambda^{p/2} \exp\{(1-\lambda)d_i^2/2\}},$$

where

$$d_i^2 = (Y_i - \boldsymbol{\alpha})' \boldsymbol{\Sigma}_{YY}^{(1)-1} (Y_i - \boldsymbol{\alpha}).$$

Example 10

If $q_i \times \nu$ has the chi-squared distribution with ν degrees of freedom, the marginal distribution is the multivariate t distribution and

$$w_i = \frac{\nu + p}{\nu + d_i^2}.$$

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

We note that w_i is decreasing for d_i in above two examples.

For model 2, we also get the estimates in the same way. But the calculation of the conditional expectations in E-step is more complicated, because $\mathrm{E}(\boldsymbol{Z}_i \mid q_i, \boldsymbol{Y}_i)$ depends on q_i in this case. In the following example, we show the expectations which would be needed in E-step, in the contaminated multivariate normal case.

Example 11

$\mathrm{M}(q_i)$ is the same to that in example 1, that is, this example is the contaminated multivariate normal case.

$$E(q_i \mid \boldsymbol{Y}_i) = \frac{(1-\delta) \mid \boldsymbol{\Sigma} \mid^{-1/2} + \delta\lambda \mid \boldsymbol{\Sigma}^* \mid^{-1/2} D^2}{(1-\delta) \mid \boldsymbol{\Sigma} \mid^{-1/2} + \delta \mid \boldsymbol{\Sigma}^* \mid^{-1/2} D_i^2} = w,$$

$$E(q_i Z_i \mid \boldsymbol{Y}_i) = \boldsymbol{\beta}'(h_{1i}\boldsymbol{\Sigma}^{-1} + \lambda h_{\lambda i}\boldsymbol{\Sigma}^{*-1})(\boldsymbol{Y}_i - \boldsymbol{\alpha}),$$

$$\begin{aligned} E(q_i \boldsymbol{Z}_i \boldsymbol{Z}_i' \mid \boldsymbol{Y}_i) &= w_i I_p + \boldsymbol{\beta}'[h_{1i}\boldsymbol{\Sigma}^{-1}\{I_p + (\boldsymbol{Y}_i - \boldsymbol{\alpha})(\boldsymbol{Y}_i - \boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}\} \\ &\quad + \lambda h_{\lambda i}\boldsymbol{\Sigma}^{*-1}\{\mathrm{I}_p + (\boldsymbol{Y}_i - \boldsymbol{\alpha})(\boldsymbol{Y}_i - \boldsymbol{\alpha})'\boldsymbol{\Sigma}^{*-1}\}]\boldsymbol{\beta}, \end{aligned}$$

where

$$h_{1i} = \frac{(1-\delta) \mid \boldsymbol{\Sigma} \mid^{-1/2}}{(1-\delta) \mid \boldsymbol{\Sigma} \mid^{-1/2} + \delta \mid \boldsymbol{\Sigma}^* \mid^{-1/2}} D_i^2,$$

$$h_{\lambda i} = 1 - h_{1i} = \frac{\delta \mid \boldsymbol{\Sigma}^* \mid^{-1/2}}{(1-\delta) \mid \boldsymbol{\Sigma} \mid^{-1/2} + \delta \mid \boldsymbol{\Sigma}^* \mid^{-1/2}} D_i^2,$$

$$D_i^2 = \exp\{\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{\alpha})'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{*-1})(\boldsymbol{Y}_i - \boldsymbol{\alpha})\},$$

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}'\boldsymbol{\beta} + \boldsymbol{\Psi},$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{\beta}'\boldsymbol{\beta} + \boldsymbol{\Psi}/\lambda.$$

If the model of the distribution function of q_i includes some unknown or unspecified parameters, additional calculations given in Section 3.4 are needed.

5.3 Simulation study

This section compares the efficiency of each maximum likelihood estimators from the multivariate normal distribution model, multivariate t distribution model, and contaminated multivariate normal distribution model as response data $\boldsymbol{Y}_i$ and random vector of factor scores $\boldsymbol{Z}_i$. we conducted simulations to numerically compare robustness, including the influence of misspecification of underlying distributions.

When misspecification of underlying distributions in a family of scale mixtures of multivariate normal distributions happens, the sample variance covariance matrix is not a consistent estimator of variance covariance matrix, but sample correlation coefficient matrix is consistent estimator of correlation coefficient matrix. This fact suggests us that we had better distinguish between estimation based on covariance matrix and that based on correlation matrix.

5.3.1 Simulation plan

Our numerical model is based on a two factor model for the open/closed book data in Mardia *et al.* (1979). Here, we set the order p of response data as 5 and the number m of common factors as 2 and made the following settings for the factor loading matrix and the specific variance matrix:

$$\boldsymbol{\alpha} = \mathbf{0},$$

$$\boldsymbol{\beta}' = \begin{bmatrix} 0.63 & 0.70 & 0.89 & 0.78 & 0.73 \\ 0.37 & 0.31 & 0.05 & -0.20 & -0.20 \end{bmatrix},$$

$$\boldsymbol{\Psi} = \mathrm{diag}(0.46, 0.42, 0.20, 0.35, 0.43).$$

This numerical model is based on the maximum likelihood estimates made by Mardia *et al.*. During the generation of artificial data, we used the following four distributions for the factor scores $\boldsymbol{Z}_i$ and the error terms $\boldsymbol{e}_i$.

1) Multivariate normal distribution (MN)

2) Multivariate t distribution with 10 degrees of freedom (T10)

3) Multivariate t distribution with 4 degrees of freedom (T4)

4) Contaminated multivariate normal distribution (CN)

$$0.9 \times \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{\Sigma}\right) + 0.1 \times \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{\Sigma}/0.0767\right)$$

For the four sets of artificial data based on different assumptions of underlying distribution generated from the above models, we calculated the following four MLEs.

a) MLE on the assumption of multivariate normal distribution
(normal MLE)

b) MLE on the assumption of multivariate t distribution with 10 degrees of freedom
(t10-type MLE)

c) MLE on the assumption of multivariate t distribution with 4 degrees of freedom
(t4-type MLE)

d) MLE on the assumption of contaminated multivariate normal distribution
(contaminated MLE)

Thus, one experiment is enough to calculate a total of 16 estimates with regard to factor loadings $\boldsymbol{\beta}$ and specific variances $\boldsymbol{\Psi}$. For each of these estimates, we calculated the following estimation criteria.

The square root of the root mean squared error with regard to factor loadings:

$$\{\sum_{i=1}^{p}\sum_{j=1}^{m}(\widehat{\beta_{ij}} - \beta_{ij})^2/pm\}^{1/2},$$

where $\hat{\boldsymbol{\beta}}$ is rotated to satisfy that $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\beta}}$ is diagonal.

The square root of the root mean squared error with regard to specific variances:

$$\{\sum_{i=1}^{p}(\widehat{\Psi_i} - \psi_i)^2/p\}^{1/2}.$$

We made a simulation with different sample sizes 200 and 400. The simulation size was 200.

5.3.2 Result and discussion

Tables 14 and 15 show RMSEs, summarizing all tables related to multivariate normal distribution, contaminated-type multivariate normal distribution and multivariate t distribution. These four distribution patterns are arranged in the order of small to large kurtosis. Table 13 shows multivariate kurtosis (Mardia, 1970) of these four distribution patterns and these distributions can be arranged in order of the multivariate kurtosis as follows;

$$MN < T10 < CN < T4.$$

Table 13 : Multivariate kurtosis

Distribution	Normal	T(df 10)	Contam.	T(df 4)
Multivariate kurtosis	99	128	383	∞

In the direction of the line in Table 14 and Table 15, we provided distribution forms of random values used in the generation of artificial data. In the direction of the row, we provided distribution forms assumed in the creation of the maximum likelihood method. The diagonal cell in the tables, therefore, shows the efficiency of each maximum likelihood method under the right assumption of underlying distribution patterns. The non-diagonal cell, on the other hand, shows the efficiency of each maximum likelihood method under wrong assumptions of underlying distributions. The figures in parentheses show the relative ratios for each line on the basis that the figure for the diagonal cell is 100. The figure in parentheses for the final line (Mean) indicates the averages of all these relative rations from all rows. The smaller the relative ratio is, the more robust a specific distribution pattern is with regard to erroneous regulations.

First we discuss about the RMSEs of estimates of factor loadings in a comparison of multivariate normal distribution with contaminated multivariate normal distribution in Table 15. With regard to artificial data that follows the multivariate normal model, the efficiency of contaminated-type MLE almost corresponds to that of normal-type MLE. On the other hand, with regard to artificial data that follows the contaminated-type multivariate normal model, the efficiency of the same contaminated-type MLE far exceeds that of normal-type MLE. That is, when underlying distribution shifts from normal to contaminated-type distributions, normal MLE loses its efficiency. Contaminated-type MLE, on the other hand, proves robust with regard to the distribution slippage. This tendency is similar in estimating specific variances. But an increasing difference in robustness between the two MLEs in response to a rise in sample size is even larger than that when estimating factor loadings.

Next we compare RMSEs of estimates of factor loadings with regard to multivariate normal distribution and multivariate t distribution. With regard to artificial data that follows the multivariate normal distribution model, we can say that the efficiency of normal MLE differs little from that of multivariate t-type MLE with 4 and 10 degrees of freedom. But, when the sample size is large as 200, the efficiency of multivariate t-type MLE with 4 degrees of freedom is slightly lower than of the other two multivariate t distribution model with 10 degrees of freedom, the efficiency of the multivariate t-type MLE with 10 and 4 degrees of freedom is fairly high, but the efficiency of normal MLE is comparatively lower. With regard to artificial data that follows the multivariate t distribution model with 4 degrees of freedom, the efficiency of the multivariate t-type MLE with 10 degrees of freedom is lower, but the efficiency of normal MLE is even lower. On the average, MLE robustness with regard to the slippage of assumptions of underlying distribution is the highest in the case of multivariate t-type MLE with 4 degrees of freedom, followed by multivariate t-type MLE with 10 degrees of freedom, and then by multivariate normal MLE. Normal MLE is thus least robust. That is, the robustness of the resulting maximum likelihood estimator increases in direct proportion

to the multivariate kurtosis of the distribution pattern assumed in the creation of the maximum likelihood method. This tendency increases as the sample size rises. The same tendency is seen in the estimation of specific variances. The increase in difference of MLE robustness according to an increase in sample size is larger than that in the estimation of factor loadings.

Result of comparison of the contaminated multivariate normal distribution with multivariate t distributions is different from the cases including the multivariate normal distribution. As in the earlier case, the RMSE value corresponding to the upper triangular cell is smaller than the RMSE of the lower triangular cell with the diagonal cell as the borderline. That is, we can say that the maximum likelihood method assuming a longer tailed than generated data distribution affects the robustness stemming from erroneous regulations of underlying distribution in the maximum likelihood method, less than the maximum likelihood method assuming a distribution pattern with a shorter tailed. But, unlike the earlier example including multivariate normal distribution, the gap between contaminated-type distribution and multivariate t distribution is not so wide.

Generally speaking, MLE under normal distribution is less robust than MLE that assumes a heavier-tailed distribution.

Table 14 : Root Mean Squared Error ($\times 1000$)

Standardized Data

Factor loadings

$n = 200$

	Assumption of distribution			
Data	Normal	T(df 10)	Contam.	T(df 4)
Normal	64(100)	65(102)	66(103)	70(109)
T(df 10)	73(106)	69(100)	75(109)	71(103)
Contam.	116(178)	66(102)	65(100)	66(102)
T(df 4)	99(152)	65(101)	74(113)	65(100)
Mean	(134)	(101)	(106)	(103)

$n = 400$

	Assumption of distribution			
Data	Normal	T(df 10)	Contam.	T(df 4)
Normal	46(100)	49(106)	48(102)	51(109)
T(df 10)	56(119)	52(100)	54(103)	52(100)
Contam.	88(248)	50(110)	46(100)	49(107)
T(df 4)	76(206)	53(108)	59(120)	49(100)
Mean	(139)	(106)	(106)	(104)

Table 14 : (Continued)

Specific variances

$n = 200$

	Assumption of distribution			
Data	Normal	T(df 10)	Contam.	T(df 4)
Normal	72(100)	74(103)	72(101)	78(108)
T(df 10)	79(107)	74(100)	81(109)	76(103)
Contam.	115(153)	75(101)	75(100)	76(101)
T(df 4)	108(142)	78(103)	87(114)	76(100)
Mean	(126)	(102)	(106)	(103)

$n = 400$

	Assumption of distribution			
Data	Normal	T(df 10)	Contam.	T(df 4)
Normal	52(100)	55(107)	54(102)	57(109)
T(df 10)	62(106)	58(100)	60(104)	58(100)
Contam.	93(176)	57(109)	52(100)	54(103)
T(df 4)	79(150)	57(109)	64(122)	53(100)
Mean	(133)	(106)	(107)	(103)

Table 15 : Root Mean Squared Error ($\times 1000$)

Factor loadings

$n = 200$

	Assumption of distribution			
Data	Normal	T(df 10)	Contam.	T(df 4)
Normal	74(100)	80(108)	75(101)	92(124)
T(df 10)	118(155)	76(100)	87(114)	79(104)
Contam.	333(444)	95(127)	75(100)	80(106)
T(df 4)	263(346)	105(139)	100(131)	76(100)
Mean	(261)	(119)	(111)	(109)

$n = 400$

	Assumption of distribution			
Data	Normal	T(df 10)	Contam.	T(df 4)
Normal	54(100)	63(117)	54(100)	78(144)
T(df 10)	97(170)	57(100)	65(114)	64(112)
Contam.	447(843)	72(136)	53(100)	55(104)
T(df 4)	256(419)	93(152)	87(143)	61(100)
Mean	(383)	(126)	(114)	(115)

Table 15 : (Continued)

Specific variances

$n = 200$

	Assumption of distribution			
Data	Normal	T(df 10)	Contam.	T(df 4)
Normal	69(100)	84(122)	70(101)	106(154)
T(df 10)	123(171)	72(100)	81(113)	88(122)
Contam.	450(643)	94(134)	70(100)	76(109)
T(df 4)	342(489)	112(160)	106(151)	70(100)
Mean	(351)	(129)	(116)	(121)

$n = 400$

	Assumption of distribution			
Data	Normal	T(df 10)	Contam.	T(df 4)
Normal	51(100)	71(139)	53(104)	74(145)
T(df 10)	111(206)	54(100)	65(120)	73(135)
Contam.	447(843)	77(145)	53(100)	60(113)
T(df 4)	361(592)	107(175)	91(149)	61(100)
Mean	(435)	(140)	(118)	(123)

Table 16 : Variances and MSE of estimates of each parameters ($\times 100$)

Standardized Data		
Multivariate Normal Distribution		
Variances of estimates		
Factor loadings		Specific variances
20	76	68
14	83	51
6	54	8
12	38	24
13	47	22
MSE of estimates (100×Variance/MSE)		
Factor loadings		Specific variances
21(97)	77(99)	72(95)
14(99)	84(99)	54(95)
7(97)	56(97)	8(99)
12(98)	41(94)	24(99)
23(99)	50(94)	22(99)

Multivariate t Distribution (df 10)		
Variances of estimates		
Factor loadings		Specific variances
31	128	96
23	134	76
13	93	12
23	61	30
22	70	30
MSE of estimates (100×Variance/MSE)		
Factor loadings		Specific variances
32(96)	129(99)	103(93)
23(97)	137(99)	83(92)
13(96)	99(95)	12(99)
23(98)	66(93)	31(98)
22(99)	74(94)	30(99)

Table 16 : (Continued)

Contaminated Normal Distribution		
Variances of estimates		
Factor loadings		Specific variances
94	525	192
51	500	125
31	247	29
55	217	81
58	212	81
MSE of estimates (100×Variance/MSE)		
Factor loadings		Specific variances
96(98)	553(95)	217(88)
51(99)	531(94)	135(92)
33(95)	262(93)	30(98)
56(98)	227(96)	87(93)
59(97)	215(99)	82(99)

Multivariate t Distribution(df 4)		
Variances of estimates		
Factor loadings		Specific variances
73	393	164
51	357	115
26	218	23
50	191	661
58	190	725
MSE of estimates (100×Variance/MSE)		
Factor loadings		Specific variances
75(97)	417(94)	177(93)
52(99)	369(98)	129(89)
28(93)	229(95)	23(99)
51(99)	194(99)	687(96)
59(99)	197(97)	740(98)

Table 17 : Variances and MSE of estimates of each parameters ($\times 100$)

Multivariate Normal Distribution		
Variances of estimates		
Factor loadings		Specific variances
30	82	62
27	74	49
18	43	7
25	37	21
24	44	19
MSE of estimates ($100 \times$ Variance/MSE)		
Factor loadings		Specific variances
31(99)	82(99)	64(96)
27(99)	74(99)	51(96)
18(97)	44(98)	8(99)
25(98)	39(96)	21(99)
25(99)	45(96)	19(97)

Multivariate t Distribution (df 10)		
Variances of estimates		
Factor loadings		Specific variances
61	180	158
52	195	106
40	114	16
56	87	49
48	89	46
MSE of estimates ($100 \times$ Variance/MSE)		
Factor loadings		Specific variances
131(46)	187(96)	219(72)
127(41)	197(99)	163(65)
131(30)	122(94)	364(43)
124(46)	109(80)	105(47)
105(46)	108(83)	140(33)

Table 17 : (Continued)

Contaminated Normal Distribution		
Variances of estimates		
Factor loadings		Specific variances
324	575	522
188	855	243
176	350	858
251	379	292
257	310	155
MSE of estimates (100×Variance/MSE)		
Factor loadings		Specific variances
1372(23)	760(76)	1334(39)
1166(16)	894(95)	1591(15)
1576(11)	375(93)	459(19)
1371(18)	616(61)	1183(25)
1085(24)	589(53)	1615(10)

Multivariate t Distribution(df 4)		
Variances of estimates		
Factor loadings		Specific variances
197	508	264
132	422	210
149	166	79
140	207	175
161	240	184
MSE of estimates (100×Variance/MSE)		
Factor loadings		Specific variances
917(21)	691(73)	1200(22)
855(15)	480(88)	1331(16)
1263(12)	173(96)	401(20)
1032(14)	302(69)	910(19)
909(18)	343(70)	1285(14)

The influence of misspecification of underlying distributions in the non-standardized data cases is much more than that in the standardized data cases. The most important factor of decrease of efficiency is bias caused by misspecifications of underlying distributions. Table 16 and Table 17 show variances and MSEs of estimates of each parameters under the assumption of multivariate normal distributions, based on standardized data, and based on non-standardized data, respectively. The estimates in the latter case include serious bias.

5.4 Application to real data

This section applies the factor analysis method (which we propose in this paper) to the open/closed book data (Mardia *et al.*, 1979, Table 1.2.1) which has been used frequently as an example of multivariate analysis, and describes the advantages of the method in analysis of actual data.

Table 18 : Open/closed Book Data

Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
77	82	67	67	81
63	78	80	70	81
75	73	71	66	81
55	72	63	70	68
63	63	65	70	63
53	61	72	64	73
51	67	65	65	68
59	70	68	62	56
62	60	58	62	70
64	72	60	62	45
52	64	60	63	54
55	67	59	62	44
50	50	64	55	63
65	63	58	56	37
31	55	60	57	73
60	64	56	54	40
44	69	53	53	53
42	69	61	55	45
62	46	61	57	45
31	49	62	63	62
44	61	52	62	46
49	41	61	49	64
12	58	61	63	67
49	53	49	62	47
54	49	56	47	53
54	53	46	59	44
44	56	55	61	36
18	44	50	57	81

Table 18 : (Continued)

Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
46	52	65	50	35
32	45	49	57	64
30	69	50	52	45
46	49	53	59	37
40	27	54	61	61
31	42	48	54	68
36	59	51	45	51
56	40	56	54	35
46	56	57	49	32
45	42	55	56	40
42	60	54	49	33
40	63	53	54	25
23	55	59	53	44
48	48	49	51	37
41	63	49	46	34
46	52	53	41	40
46	61	46	38	41
40	57	51	52	31
49	49	45	48	39
22	58	53	56	41
35	60	47	54	33
48	56	49	42	32
31	57	50	54	34
17	53	57	43	51
49	57	47	39	26
59	50	47	15	46
37	56	49	28	45
40	43	48	21	61
35	35	41	51	50
38	44	54	47	24

Table 18 : (Continued)

Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
43	43	38	34	49
39	46	46	32	43
62	44	36	22	42
48	38	41	44	33
34	42	50	47	29
18	51	40	56	30
35	36	46	48	29
59	53	37	22	19
41	41	43	30	33
31	52	37	27	40
17	51	52	35	31
34	30	50	47	36
46	40	47	29	17
10	46	36	47	39
46	37	45	15	30
30	34	43	46	18
13	51	50	25	31
49	50	38	23	9
18	32	31	45	40
8	42	48	26	40
23	38	36	48	15
30	24	43	33	25
3	9	51	47	40
7	51	43	17	22
15	40	43	23	18
15	38	39	28	17
5	30	44	36	18
12	30	32	35	21
5	26	15	20	20
0	40	21	9	14

Table 18 lists the open/closed book data used here. This is a set of data obtained from five tests (Mechanics, Vectors, Algebra, Analysis and Statistics) made on 88 subjects. In algebraic, Analytic and Statistical tests, the subjects were allowed to

refer to their textbooks (open book examination). In Mechanical and Vectors tests, the subjects kept their textbooks closed (closed book examination). Mardia *et al.* (1979) calculated maximum likelihood estimates under the assumption of multivariate normality while using the algorithm derived by Jöreskog (1967). Mardia *et al.* concluded that, as far as factors were concerned, two-factor model fits the data well and interprets the two factors as the first factor that shows general capabilities and the second factor that emphasizes the capabilities of closed book examinations in comparison to those of open book examinations. They also calculated maximum likelihood estimates under one-factor model. Table 19 shows the maximum likelihood estimates under both one-factor and two-factor models and the results of tests of goodness of fit. We can not reject both normal factor analysis models, but two-factor model is selected through *AIC*.

Table 19 : Maximum likelihood estimates

One-Factor Model		Two-Factor Model		
Factor loadings	Specific variances	Factor loadings		Specific variances
.599	.641	.628	.373	.466
.667	.555	.695	.312	.419
.917	.159	.899	-.050	.189
.772	.403	.780	-.201	.352
.724	.476	.727	-.200	.431
Chi-Square(DF=5)	8.651	Chi-Square(DF=1)		0.075

In order to describe the advantages of the newly proposed robust factor analysis method, we add two quasi-outliers to the original open/closed book data, and analyze the data using proposed method. Two quasi-outliers are as follows:

$$\text{No.89}: \{0,\ 82,\ 15,\ 70,\ 9\}$$
$$\text{No.90}: \{77,\ 9,\ 80,\ 9,\ 81\}.$$

These data are generated by combination of maximum and minimum values of each test.

5.4.1 Model selection

Table 20 shows summary of model fits when we fitted the ML factor analysis model using the multivariate normal, multivariate t and contaminated multivariate normal distributions with unknown mixing parameters. Such parameters were estimated by the method given in Section 3.4 simultaneously with regression and variance-covariance parameters. According to AIC, one factor model with the contaminated multivariate normal distribution is selected.

The results of ML estimation undel each model are given in Table 21. We have an improper solution in two factor model with the multivariate normal distribution, but the other two factor models have proper solutions. This robust method is available for protection of improper solutions due to outliers.

Table 20 : Summary of model fits

Distribution	Number of Factors	-2×log likelihood	Number of Parameters	AIC
MN	1	3580.7	15	3610.7
MT	1	3532.5	16	3564.5
MT	2	3528.8	21	3570.8
CN	1	3520.7	17	3554.7
CN	2	3514.5	22	3558.5

Table 21 : Maximum likelihood estimates

Multivariate Normal Distribution				
One-factor Model		Two-factor Model		
Factor loadings	Specific variances	Factor loadings		Specific variances
.628	.466			
.695	.419	*		
.899	.189			
.780	.352			
.727	.431			
Multivariate t Distribution				
One-factor Model		Two-factor Model		
$\hat{\nu} = 7.21$		$\hat{\nu} = 7.43$		
.628	.606	.629	.237	.548
.649	.579	.665	.314	.459
.896	.197	.882	-.030	.221
.744	.447	.746	-.030	.442
.730	.467	.748	-.272	.366
Contaminated Multivariate Normal Distribution				
One-factor Model		Two-factor Model		
$\hat{\delta} = 0.034, \hat{\lambda} = 0.085$		$\hat{\delta} = 0.025, \hat{\lambda} = 0.065$		
.609	.630	.642	.418	.413
.673	.547	.678	.240	.483
.916	.162	.896	-.055	.195
.757	.428	.766	-.191	.377
.726	.473	.730	-.202	.427

* : Improper Solutions

5.4.2 Outliers

The preceding section indicated that, from the viewpoint of *AIC*, the factor analysis model with contaminated normal distributions or multivariate t distributions better fits the data than the factor analysis model with conventional multivariate normal distributions. This fact, however, dose not justify the idea that contaminated multivariate normal or multivariate t distributions are more desirable than multivariate normal distribution as a population distribution model for latent factor scores and error terms. A more justifiable idea would be that this fact indicates that this set of data includes some values deviating from the majority of data, that is, outliers. This requires a method for detecting these outliers from the set of data: the convergent value of w_i obtained in E-step can be used as effective statistics to detect outliers.

Table 22 : Factor scores (multivariate normal model)

	Factor scores			Factor scores			Factor scores	
No.	1st	2nd	No.	1st	2nd	No.	1st	2nd
1	2.18	.27	31	.27	.15	61	-.64	1.32
2	2.46	-.55	32	.24	-.14	62	-.58	.30
3	2.13	-.10	33	-.05	-1.41	63	-.33	-.13
4	1.49	-.30	34	-.14	-1.00	64	-.67	-.26
5	1.47	-.33	35	.17	.05	65	-.58	-.21
6	1.58	-.81	36	.26	-.08	66	-.60	1.88
7	1.37	-.46	37	.41	.29	67	-.67	.52
8	1.55	.02	38	.16	-.39	68	-.82	.69
9	1.08	-.20	39	.32	.36	69	-.40	-.06
10	1.27	.57	40	.31	.44	70	-.51	-.60
11	1.02	-.12	41	.28	-.63	71	-.57	.81
12	1.01	.25	42	.01	.16	72	-1.06	-.43
13	.87	-.62	43	.14	.60	73	-.76	.90
14	.93	.67	44	.14	.33	74	-.89	-.08
15	.63	-1.06	45	.01	.81	75	-.63	.13
16	.80	.62	46	.13	.26	76	-.80	1.67
17	.63	.25	47	-.12	.35	77	-1.44	-.50
18	.90	.11	48	.09	-.45	78	-.88	-.38
19	.74	-.12	49	-.00	.24	79	-1.18	.03
20	.59	-1.24	50	.05	.70	80	-1.16	-.17
21	.48	-.05	51	.01	-.02	81	-1.26	-2.04
22	.51	-.71	52	.03	-.66	82	-1.11	.50
23	.50	-1.50	53	-.06	.98	83	-1.19	.29
24	.28	-.11	54	-.17	1.27	84	-1.34	.22
25	.44	-.00	55	-.14	.56	85	-1.35	-.57
26	.18	.20	56	-.36	.18	86	-1.72	-.15
27	.44	-.07	57	-.64	-.56	87	-2.72	.35
28	-.09	-1.59	58	-.11	.01	88	-2.42	.87
29	.68	-.08	59	-.67	.37			
30	-.03	-.91	60	-.40	.32			

w_i is the conditional expectation $E(q_i \mid Y_i)$. If the q_i follows the one-point distribution which constantly takes 1, the estimation method described in Section 5.2 serves to produce the maximum likelihood estimate under multivariate normal distribution. That is, the nearer the value w is to 1, the better that data fits the factor analysis model in multivariate normality. On the contrary, the nearer the value w_i is to 0, the more that data is likely to be out of the multivariate normal-type factor analysis model. This is easily understandable from the fact that the estimation algorithm proposed in this paper is equivalent to the iteratively reweighted least square algorithm at the moment the factor score of each data is observed and that w acts as a weight imposed on each data.

Table 23 : Conditional Expectations of q (One-factor CN model)

	1	2	3	4	5	6	7	8	9	0
	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.95	1.00	1.00
30	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
40	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	0.98	1.00	0.99	1.00	1.00	1.00	1.00
60	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
70	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
80	0.23	1.00	1.00	1.00	1.00	1.00	0.97	0.99	0.09	0.09

Table 23 shows convergent values of w_i obtained under the one-factor contaminated normal model. The values w_i of a total of 90 subjects of data are arranged. The tables indicates that the 89st and 90th subject have values w_i of 0.1 or less, and that the 81st subject also has relatively small w. A close look at the original data of the 81st subject indicates that the scores in closed-book-type mechanical and vector examinations are very low (the total of the scores of these two subjects is the lowest of all 88 subjects),

while the open book examination of the other three subjects indicated approximate the average scores. This is evident from the factor score related to the two extracted factors (Table 22). The factor score of the second factor of the 81st subject is as low as -2.27.

Comparing Table 19 and Table 21, we can conclude that the results under non-normal distribution assumptions are less influenced by two additional quasi-outliers than those under normal assumptions. In particular, the two-factor normal model is a Heywood case, namely, some unique factor has negative variance.

5.5 Discussion and conclusion

We conducted a simulation to prove that the ML factor analysis method under the assumption of multivariate t distribution or multivariate contaminated distribution is more robust with regard to erroneous regulations of these underlying distribution patterns than the conventional ML factor methods under the assumption of multivariate normal distribution for factor scores and error terms. According to this simulation, we found that the ML factor analysis method under the assumption of large kurtosis dose not lose much of its efficiency of estimation if its kurtosis is smaller than that of assumed distribution, even though the true distribution of the data is different from what is assumed. On the other hand, if the kurtosis of the true distribution followed by the data is higher than that of the assumed distribution of the ML factor analysis method, (that is, in the case of the ingress of extreme values or outliers), much of efficiency of the maximum likelihood estimation is lost. This indicates that the ML factor analysis method under multivariate normal distribution with small kurtosis is less robust with regard to the decay of distribution assumptions than multivariate t distribution and multivariate contaminated normal distribution. On the other hand, the ML factor analysis method under the assumption of distribution with the largest possible kurtosis is robust in the sense that its efficiency of estimation remains constant

regardless of the misspecification of distribution assumptions.

When we apply the normal ML factor analysis, we are sometimes faced to general problems, for example nonconvergence, improper solutions (Heywood case), factor rotations and identifiability. In our method we can avoid the problem concerned with improper solution (because we use the EM algorithm to obtain the ML estimates), but in such case the EM algorithm can not stop (see Dempster *et al.* 1977). The rest problems are essential issues for the ML factor analysis. We also have to pay attention to these problems when applying robust factor analysis.

6 References

[1] Aitkin, M. and Wilson, G.T. (1980) Mixture models, outliers, and the EM algorithm, *Technometrics*, **22**, 325-331.

[2] Andersen, P.K. (1982) Testing goodness-of-fit of Cox's regression model, *Biometrics*, **38**, 67-77.

[3] Anderson, T.W. (1987) Multivariate linear relations, *Proc. 2nd Inter. Tampere Con. Statist.* (Pukkila and Puntanen eds.) 9-36.

[4] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.M. and Tukey, J.W.(1972) *Robust Estimates of Location; Survey and Advances*, Princeton University Press.

[5] Andrews, D.F. and Mallows, C.L.(1974) Scale mixtures of normal distributions, *J. Roy. Statist. Soc.* **B 36**, 99-102.

[6] Barnett V. and Lewis, T. (1978) *Outliers in statistical data*, Wiley.

[7] Bentler, P.M. and Tanaka, J.S.(1983) Problems with EM algorithms for ML factor analysis, *Psychometrika* **48**, 247-251.

[8] Breslow, N. (1981) Odds ratio estimators when the data are sparse, *Biometrika,* **68**, 73-84.

[9] Browne, M.W. and Shapiro, A. (1988) Robustness of normal theory methods in the analysis of linear latent variate models, *British J. of Math. Statist. Psychol.* **41**, 193-208.

[10] Clayton, D.G. (1974) Some odds ratio statistics for the analysis of ordered categorical data, *Biometrika,* **61**, 525-531.

[11] Clayton, D.G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika,* **65**, 141-151.

[12] Cornish, E.A. (1954) The multivariate t-distribution associated with a set of normal standard deviates, *Australian journal of physics,* **7**, 531-542.

[13] Cox, D.R. (1972) Regression model and life tables (with discussion), *J. Roy. Statist. Soc.* **B 34**, 187-202.

[14] Cox, D.R. (1975) Partial likelihood, *Biometrika,* **62**, 269-279.

[15] 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.* **B 39**, 1-38.

[16] Dempster, A.P., Laird, N.M. and Rubin, D.B.(1980) Iteratively reweighted least squares for linear regression when errors are normal/independent distributed, *Multivariate Analysis*-V,(Krishnaiah,P.R. ed.), 35-57.

[17] Durbin, J. (1973) *Distribution theory for tests based on the sample distribution function*, CBMS-NSF **9**.

[18] Elston, R.C. and Grizzle (1962) Estimation of time-response curves and their confidence bands, *Biometrics*, **18**, 148-159.

[19] Emmette, (1949) Factor analysis by Lawley's method of maximum likelihood, *Brit. J. Psychol., Statist. Sect.*, **2**, 90-97.

[20] Harman, H.H. (1967) *Modern Factor Analysis*, University of Chicago Press.

[21] Hauck, W.W. (1988) The asymptotic relative efficiency of the Mantel-Haenszel estimator in the increasing-number-of-strata case, *Biometrics,* **44**, 379-384.

[22] Hauck, W.W., Anderson, S. and Leahy, F.J. (1987) Finite-sample properties of some old and some new estimators of a common odds ratio from multiple 2 × 2 tables, *Journal of the American statistical association*, **77**, 145-152.

[23] Jeffreys, H. (1961) *Theory of Probability*, Third ed., Oxford University Press.

[24] Jennrich, R.I. and Schluchter, M.D.(1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805-820.

[25] Jöreskog, K.G. (1967) Some contribution to maximum-likelihood factor analysis. *Psycometrika*, **32**, 443-482.

[26] Kalbfleisch, J.K. and Prentice, R.L. (1973) Marginal likelihoods based on Cox's regression and life model, *Biometrika,* **60**, 267-278.

[27] Kalbfleisch, J.K. and Sprott, D.A. (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion), *J. Roy. Statist. Soc.* **B 32**, 175-208

[28] Kano, Y, Berkane, M. and Bentler, P.M. (1990) Pseudo maximum likelihood estimation and elliptical distribution theory in a misspecified model, *UCLA Statistical Series* # 51.

[29] Kariya, T. and Sinha B.K. (1989) *Robustness of statistical tests*, Academic Press.

[30] Laird, N.M. and Ware, J.H. (1982) Random-effects models for longitudinal data, *Biometrics*, **38**, 963-974.

[31] Lange, K.L., Little R.J.A. and Taylor, J.M.G.(1989) Robust statistical modeling using the t distribution. *Journal of the American statistical association*, **84**, 881-896.

[32] Lawley, D.N. and Maxwell, A.E. (1963) *Factor analysis as a statistical method.* Butterworth.

[33] Little, R.J.A.(1988) Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics* **37**, 23-38.

[34] Louis, T.A.(1982) Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc.* **B 44**, 226-233.

[35] Mardia, K.V.(1970) Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519-530.

[36] Mardia, K.V., Kent, J.T. and Bibby, M.(1979) *Multivariate Analysis*, Academic Press.

[37] Pendergast, J.D. and Broffitt J.D. (1985) Robust estimation in growth curve models, *Com. Statist. - Theor. Meth.*, **18**, 1919-1939.

[38] Plackett, R.L. (1965) A class of bivariate distribution, *Journal of the American statistical association*, **60**, 516-522.

[39] Potthoff, R.F. and Roy S.N.(1964) A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51**, 313-326.

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

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

[42] Rubin, D.B. and Thayer, D.T.(1982). EM-algorithm for ML factor analysis, *Psychometrika* **47**, 69-75.

[43] Rubin, D.B. and Thayer, D.T.(1983). More on EM for ML factor analysis, *Psychometrika* **48**, 253-57.

[44] Schumacher, M. (1984) Two-sample tests of Cramér-von Mises and Kolmogorov-Snirnov type for randomly censored data, *International Statistical Review,* **52**, 263-281.

[45] Sutradhar, B.C. and Ali, M.M. (1986) Estimation of the parameters of a regression model with a multivariate t error variable, *Communications in statistics theory and methods,* **15**, 429-450.

[46] Tsiatis, A.A. (1981) A large sample study of Cox's regression model, *Annals of statistics*, **9**, 93-108.

[47] Tyler, D.E. (1983) Robustness and efficiency properties of scatter matrices, *Biometrika*, **70**, 411-420.

[48] Ware, J.H. (1983) Growth curves *Entry in Encyclopedia of the Statistical Sciences*, **4** , (Kotz, S., Johnson, N.L. and Read, C.B., eds.), Wiley.

[49] Ware, J.H. (1985) Linear models for the analysis of longitudinal studies, *The American statistician,* **39**, 95-101

[50] Watanabe, M. and Yamaguchi, K.(1989) ML factor analysis under scale mixtures of normal distributions. *Proceedings of the sixth Korea and Japan joint conference of statistics* 162-166.

[51] Wei, L.J. (1984) Testing goodness of fit for proportional hazard model with censored observations. *Journal of the American statistical association*, **79**, 649-652.

[52] Wong, W.H. (1986) Theory of partial likelihood. *Annals of statistics*, **14**, 88-123.

[53] Yamaguchi, K.(1986) A model for association in bivariate survival data based on proportional hazard, *Proc. 4th Korea-Japan joint conference of statistics,* 171-174.

[54] Yamaguchi, K.(1988) A Cramér-von Mises type test of the proportional hazards assumption, *Engineering sciences reports, Kyushu univ.*, **10**, 67-70.

[55] Yamaguchi, K.(1988) Finite sample properties of some estimators of a common odds ratio by a simulation study, *Proc. 5th Korea-Japan joint conference of statistics,* 74-76.

[56] Yamaguchi, K.(1989) Linear models with heavy-tailed error distributions. *The third Japan-China simposium on statistics* 288-292.

[57] Yamaguchi, K.(1990a) Analysis of repeated measures data with outliers, *Bulletin of Informatics and Cybernetics*, **24**, 71-80.

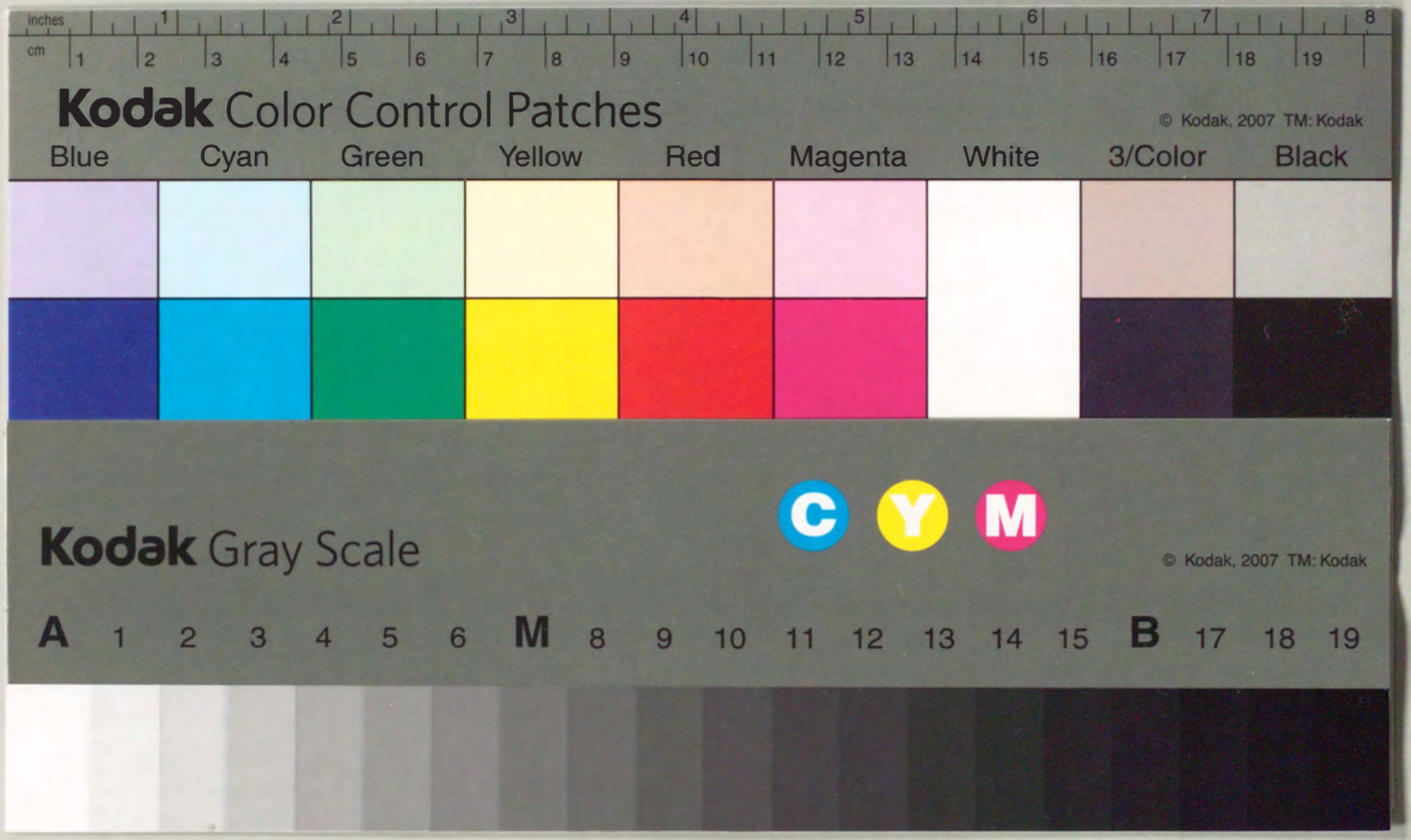
[58] Yamaguchi, K.(1990b) Analysis of repeated measures data with multivariate t or contaminated multivariate normal errors (in Japanese), *Bulletin of the Computational Statistics of Japan*, **3**, 11-22.

[59] Yamaguchi, K.(1990c) Generalized EM algprithm for models with contaminated normal error terms, *Statistical Methods and Data Analysis*, (NIki, N. Ed.), Scientist Inc., 107-114.

[60] Yamaguchi, K. and Watanabe, M. (1990a) ML factor analysis for scale mixtures of normal distributions and its robustness (in Japanese), *Keizaironshu*, Kansai Univ. **39** 161-182

[61] Yamaguchi, K. and Watanabe, M. (1990b) A simulation study for the influence of misspecified models in ML factor analysis, *Statistical Methods and Data Analysis*, (NIki, N. Ed.), Scientist Inc., 81-87.

[62] Zellner, A. (1976) Bayesian and non-Bayesian analysis of the regression model with multivariate student-t error terms, *Journal of the American statistical association*, **71**, 400-405.


inches
cm
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
Kodak Gray Scale
C
Y
M
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19