Classification of repeated observations based on the growth curve model

Kenichi Satoh, Keiko Otani and Megu Ohtaki

Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, JAPAN

e-mail: ksatoh@hiroshima-u.ac.jp

SUMMARY. A classification method for repeated observations is proposed. The method consists of two procedures. First, individual response profiles are fitted by the polynomial regression model. Next, the fitted regression coefficients are classified based on the normal mixture model. The procedures can be embedded in one model, i.e., the growth curve model with extended random effects. It helps us to explore and describe unobserved background factors.

KEY WORDS: classification; repeated observations; growth curve model; heterogeneity; random effects; normal mixture distribution.
1. Introduction

Repeated observation data often involve more than two groups which are specified by some explanatory variables such as sex or sort of treatments. Several useful statistical methods, which are called “growth curve analysis”, have been proposed for such data (Wishart, 1938; Rao, 1965; Laird, Lange and Stram et al. 1987). Fujikoshi (1993, 1994) reviewed recent theoretical developments of the growth curve model. The model can deal with two types of explanatory variable. One is the so-called “between-individual variable” having a constant value during the observation period, such as sex, treatment group and blood type, and the other is the “within-individual variable” having a nonconstant value, such as body weight. If between-individual variables are the same, the individual response profiles are expected to show a common growth curve. Hirotsu (1983, 1991) pointed out that use of the ordinary growth curve analysis may be invalid unless individual response profiles are homogeneous in each group. In order to deal with the heterogeneity, Tango (1989) proposed a statistical method using a normal mixture model for analyzing balanced repeated observations and evaluated the efficacy of a drug by using those mixture ratios. Skene and White (1992) also discussed classification by using the normal mixture model; they assumed that mean vectors had ANOVA-type linear structure. Pauler and Laird (2000) suggested nonlinear mean functions in the mean vectors of the normal mixture model. All of these methods assume that the response observation vectors are distributed
with a normal mixture distribution. Therefore, they cannot be applied to unbalanced data, as it is required that the dimension of the observation vector or observation times have to be the same for all individuals. Assuming that the adjusted regression coefficient vectors rather than response observations have a normal mixture distribution, we propose a classification method based on growth curve data. Our procedure can be regarded as an extension of the growth curve model with random effects (Vonesh and Carter, 1987; Ohtaki, 1994), which is applicable to unbalanced data. We illustrate its good performance using data previously analyzed by Potthoff and Roy (1965).

2. Statistical model

Our growth curve model with normal mixture distribution is described in the following three stages. Stages 1 and 2 are the same as those due to Laird and Ware (1982). Stage 3 is an extension.

[Stage 1.] For $i$th individual with repeated observations $y_i = (y_{i1}, \cdots, y_{ip_i})'$, assume the following linear model:

$$y_{i} = X_{i} \beta_{i} + e_{i}, \quad e_{i} \sim i.i.d. \ N_p(0, \sigma^2 I_p), \quad (i = 1, \cdots, n),$$

(2.1)

where $p_i$ is the number of repetitions, $X_i$ is a known within-individual design matrix which may include a polynomials of up to $q - 1$ degree, $e_i$ is an observation error vector, $\sigma^2$ is an unknown variance parameter, and $\beta_i$ is
a regression coefficient vector. Then calculate the unbiased estimate of $\beta_i$ which is specified as

$$\hat{\beta}_i = (X_i'X_i)^{-1}X_i'y_i,$$  \hspace{1cm} (2.2)

for each individual ($i = 1, \ldots, n$).

[Stage 2.] The regression coefficient vectors in (2.1) may be written as

$$\beta_i = \Theta \quad a_i + m_i,$$  \hspace{1cm} (2.3)

where $a_i$ is a known between-individual design vector, $\Theta$ is an unknown parameter matrix, and $m_i$ is a random effects vector which can be considered as an adjusted regression coefficient vector in the sense that the between-individual variables are removed from its coefficient vector. Then the dimensions of the regression coefficient vectors are always equal even if the vectors of observation times are different among individuals, i.e., within-individual design matrices are unbalanced. It therefore enables us to classify response profiles with an unbalanced design.

[Stage 3.] Assume that the random effects vector is distributed with $K$-component normal mixture distribution which is given by

$$m_i \overset{i.id.}{\sim} \sum_{j=1}^{K} \xi^{(j)}N_q(\mu^{(j)}, \Sigma^{(j)}),$$  \hspace{1cm} (2.4)

where $m_i$’s and $e_i$’s are independent, $\xi^{(j)}$ are the unknown mixture ratios which sum to unity, and $\mu^{(j)}$ and $\Sigma^{(j)}$ are the unknown local mean vector and covariance matrix for the $j$th component. Without loss generality,
the expected value of the random effects vector is assumed to be zero, i.e.,
\( E(\mathbf{m}_i) = \mathbf{0} \), by specifying the appropriate between-individual design vector.
Those components may express some unobserved background factors. Then
the \( j \)th regression coefficient vector can be described as
\[
\beta_i^{(j)} = \Theta \mathbf{a}_i + \mu^{(j)}.
\] (2.5)
This regression coefficient vector may make up for effects that can not be
explained by the between-individual variables in hand.

From Stages 1-3 it follows that the repeated observation vector may be
expressed as
\[
y_i = X_i \Theta \mathbf{a}_i + X_i \mathbf{m}_i + e_i,
\] (2.6)
with the expectation given by \( X_i \Theta \mathbf{a}_i \). Therefore, the response profiles due
to the \( j \)th local mean vector are given by \( X_i \Theta \mathbf{a}_i + X_i \beta_i^{(j)} \).

3. Estimation and Classification

In this section we discuss a method for estimating unknown parameters
of the normal mixture model. First we assume the case where the coefficient
parameter matrix \( \Theta \) is known. Let \( z_i = y_i - X_i \Theta \mathbf{a}_i \), so it is also a known
vector. From Stage 2, it is expressed as
\[
z_i = X_i \mathbf{m}_i + e_i.
\] (3.1)
Multiplying (3.1) by \((X_i'X_i)^{-1}X_i'\), we obtain

\[
\mathbf{\hat{m}}_i = \mathbf{m}_i + \mathbf{\delta}_i,
\]

(3.2)

where

\[
\mathbf{\hat{m}}_i = (X_i'X_i)^{-1}X_i'\mathbf{z}_i \text{and} \mathbf{\delta}_i = (X_i'X_i)^{-1}X_i'\mathbf{e}_i.
\]

(3.3)

Because \(X_i'X_i = O(p_i)\) under a mild regularity condition, the mean vector and covariance matrix of \(\mathbf{\delta}_i\) are given by

\[
\mathbb{E}(\mathbf{\delta}_i) = \mathbf{0} \text{ and } \text{Var}(\mathbf{\delta}_i) = \sigma^2(X_i'X_i)^{-1} = O(p_i^{-1}),
\]

(3.4)

respectively. Thus, we can estimate the normal mixture distribution by using \(\mathbf{\hat{m}}_i\) instead of \(\mathbf{m}_i\) when the \(\mathbf{m}_i\)'s are almost identically distributed, i.e., the \(X_i\)'s are almost equal or the numbers of repetitions \(p_i\)'s are large enough. It is convenient to write \(\mathbf{\hat{m}}_i\) in (3.3) as \(\mathbf{\hat{\beta}}_i - \Theta\mathbf{a}_i\).

Second, we consider the case where \(\Theta\) is an unknown parameter matrix. Recall that Vonesh and Carter (1987) proposed a restricted maximum likelihood estimator \(\hat{\Theta}\) under the condition that the random effects are distributed with a normal distribution, i.e., a single-component normal mixture distribution. Therefore we apply the estimator \(\hat{\Theta}\) instead of a known parameter matrix \(\Theta\) to the above discussion.

Classification by using the normal mixture distribution has been discussed (Wolf, 1970; Nakamura, 1995; Tango, 1989). First, the posterior probability of \(\mathbf{\hat{m}}_i\) due to the \(j\)th component \((j = 1, \ldots, K)\) is calculated under the fitted distribution. Next each individual response profile is classified.
4. Example

We analyzed data on distance from the center of the pituitary to the pterygomaxillary fissure which were published by Potthoff and Roy (1965). Figure 1 shows the individual response profiles by sex. It can be seen that the distance increases monotonously with age except for a few cases. The distances in boys are larger than those in girls in most cases. In this analysis we removed the data for two boys (ID 9 and 13) which are well-known as outliers, c.f., Chapter 2 of Imaizumi and Inaba (1997).

Next, we compared our model with the two-component normal mixture distribution and a linear regression function, to the model proposed by Vonesh and Carter (1987). Although the data have one between-individual variable “sex”, it is not used in our model and we let $a_i = 1$ for all $i$. As stated in Section 3, an unknown parameter matrix $\Theta$ was estimated; the estimate was $\hat{\Theta} = (22.182, 0.596)'$. By using $\hat{m}_i$ in (3.3), the normal mixture model

to the component which has the maximum probability. Although it is easy to incorporate a normal mixture distribution into a statistical model, the $K$-means method may be preferred in some cases. Selecting the number $K$ of components is another problem; Konishi and Nakamura (1998) discussed using some information-based criteria, such as $AIC$. 
was estimated and its local mean vectors were given by \( \hat{\mu}^{(1)} = (0.311, 0.220)' \) due to \textit{Class} 1 and \( \hat{\mu}^{(2)} = (-0.421, -0.224)' \) due to \textit{Class} 2.

- Insert Fig. 2 Here –
- Insert Fig. 3 Here –
- Insert Table 1 Here –

Figure 2 shows the scatter plot of \( \hat{m}_i \)’s and estimated local mean vectors. The classes 1 and 2 correspond closely to the groups of boys and girls, respectively. From the equation (2.5), estimated regression coefficient vectors due to the mean vectors were given by \( \hat{\beta}^{(1)} = (22.493, 0.816)' \) and \( \hat{\beta}^{(2)} = (21.762, 0.372)' \). The corresponding estimated response profiles from our model are presented in Figure 3. For comparison, estimated local mean response profiles from Vonesh and Carter model (see Table 1) are added.

- Insert Fig. 4 Here –

Finally, Figure 4 shows the result of classification of the response profiles. Thus, we conclude from the example that our classification method is useful for exploring unobserved background factors.
REFERENCES


Figure 1. Individual response profiles by sex.

Figure 2. Individual adjusted regression coefficients and classification results based on the normal mixture distribution.

Figure 3. Estimated response profiles due to local mean vectors (Class 1 and Class 2) and those of Vonesh and Carter model (Boy and Girl).

Figure 4. Classification of individual response profiles based on the growth curve model with an extended random effect.
Table 1. Estimated parameter matrix $\hat{\Theta}$ on the growth curve model proposed by Vonesh and Carter (1987), and the $P$-values obtained under the hypothesis that the parameter was zero.

<table>
<thead>
<tr>
<th>Name</th>
<th>Coeff.</th>
<th>S.D.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept for Boy</td>
<td>22.946</td>
<td>0.532</td>
<td>0.014</td>
</tr>
<tr>
<td>Gradient for Boy</td>
<td>0.687</td>
<td>0.066</td>
<td>0.061</td>
</tr>
<tr>
<td>Intercept for Girl</td>
<td>21.209</td>
<td>0.601</td>
<td>0.018</td>
</tr>
<tr>
<td>Gradient for Girl</td>
<td>0.479</td>
<td>0.075</td>
<td>0.099</td>
</tr>
</tbody>
</table>