Model based clustering of longitudinal data: application to modeling disease course and gene expression trajectories

A. Ciampi[1], H. Campbell, A. Dyachenko, B. Rich, J. McCusker, M. G. Cole

Abstract

We consider the problem of clustering time dependent data. The model is a mixture of regressions, with variance-covariance matrices that are allowed to vary within the extended linear mixed model family. We discuss applications to biomedical data and analyze two longitudinal data sets: one on patients with delirium, and the other on mosquito gene expression following infection.

1 Introduction

Model based clustering (MBC) is increasingly popular as a method for studying multivariate data. Most of the applications rely on the approach proposed by Banfield and Raftery (1993) and implemented in the R-package MCLUST. In this approach data are modeled as a mixture of multivariate normal distributions; an economic parameterization of the variance-covariance matrices is achieved by considering the spectral decomposition of the matrices.

New challenges appear when analyzing longitudinal data with non-negligible correlations. We are interested in two broad areas of application: clustering disease trajectories in a clinical setting, and clustering longitudinal data in gene expression at several points in time. The spectral decomposition is of limited help when working with such data, since it does not address the special form that the variance covariance matrices may take. In addition, longitudinal data consist of measurements taken repeatedly on a number of observational units, with the typical feature, especially in clinical settings, that both the number of measurements and the time points may differ across individual units. Analysis of such data is usually performed using the extended linear mixed model (ELMM), see Pinheiro and Bates (2000). However, the ELMM usually assumes a Gaussian distribution for all random effects and error terms. This assumption has been relaxed to include mixtures of Gaussian distributions; see, for instance, Belin and Rubin (1995), Tango (1998), Trottier (1998), Verbeke and Molenberghs (2000), Luan and

[1]McGill University & St-Mary's Hospital, Montreal, E-mail: antonio.ciampi@mcgill.ca

Li (2003), Gaffney and Smyth (2003), Celeux et al. (2005), Heard et al. (2006), Ng et al. (2006), De la Cruz-Mesa (2008). To the best of our knowledge, however, there is no fully developed method for simultaneously estimating parameters of both the random effects and the error term.

While the primary methods presented in this paper have been developed elsewhere, their application is affected by several subjective choices, based on pragmatic considerations and on assumptions, only valid in specific contexts—these are the focus of the current paper. We study a general model for representing mixtures of longitudinal data that generalizes previous attempts: a mixture of ELLM (Section 2). We develop an EM approach to the estimation of its parameters (Section 3) and validate it through simulations (Section 4). In Section 5 is we apply the model to the study of disease course, analyzing data on delirium in an elderly population. Section 6 presents an application to longitudinal gene expression data. Section 7 concludes the paper with a brief discussion.

2 The model

Let $Y_i(t_{ij})$ be the observation of the ith individual at time t_{ij}, for $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, where n is the number of individuals and m_i is the number of time points at which the ith individual has been observed. The ELMM can be written as follows:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{1}$$

where X_i and Z_i are design matrices:

$$X_i = \begin{pmatrix} g_1(t_{i1}) & \ldots & g_p(t_{i1}) \\ \ldots & \ldots & \ldots \\ g_1(t_{im_i}) & \ldots & g_p(t_{im_i}) \end{pmatrix}, \qquad Z_i = \begin{pmatrix} h_1(t_{i1}) & \ldots & h_q(t_{i1}) \\ \ldots & \ldots & \ldots \\ h_1(t_{im_i}) & \ldots & h_q(t_{im_i}) \end{pmatrix},$$

and:

$$\beta = (\beta_1, ..., \beta_p)', \qquad b_i = (b_{i1}, ..., b_{iq})', \qquad \epsilon_i \sim N(0, \sigma^2\Lambda_i)$$

with b_i and ϵ_i assumed independent. Here, Y_i is independent of Y_j for $i \neq j$ and Λ_i is an $m_i \times m_i$ matrix that may depend on i through the time intervals t_{ij}, j=1,...,m_i but not otherwise. Typically, Λ_i is parameterized in terms of a relatively small number of variance parameters. Furthermore, the distribution of the random effects, b_i, is assumed to be $N(0, \Psi)$ where Ψ is a symmetric positive definite matrix which may depend on parameters to be estimated. Finally, the g_i's and the h_i's denote the elements of a basis in function space. In practice the columns of Z_i are often chosen as a subset of the columns of X_i. We have:

$$Y_i \mid b_i \sim N(X_i\beta + Z_i b_i, \sigma^2\Lambda_i)$$

and:

$$Y_i \sim N(X_i\beta, \Sigma_i), \quad \Sigma_i = (Z_i\Psi Z_i^T + \sigma^2\Lambda_i).$$

The random effects b_i may be considered as missing data, and maximum likelihood estimation is done by the EM algorithm (see Lindstrom and Bates (1988)).

Since the joint likelihood of $(y_i^T, b_i^T)^T$ is equal to that of $((y_i \mid b_i)^T, b_i^T)^T$, we have:

$$\begin{bmatrix} y_i \mid b_i \\ b_i \end{bmatrix} \sim N\left(\begin{pmatrix} X_i\beta + Z_i b_i \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2\Lambda_i & 0 \\ 0 & \Psi \end{bmatrix}\right) \tag{2}$$

Then, as is done for missing data, we can write the "complete data" log-likelihood (Trottier 1998):

$$\begin{aligned} l(\beta, \sigma^2, \Psi, \Lambda \mid y, b) = \\ -\frac{1}{2}\sum_{i=1}^{n}\Bigg(& m_i \log(2\pi) + m_i \log(\sigma^2) + \log(\mid (\Psi) \mid) + \log(\mid (\Lambda_i) \mid) \\ & + b_i^T(\Psi^{-1})b_i + \frac{(Y_i - X_i\beta - Z_i b_i)^T(\Lambda_i)^{-1}(Y_i - X_i\beta - Z_i b_i)}{\sigma^2}\Bigg). \end{aligned} \tag{3}$$

Under the assumption that the n individuals are sampled from K distinct component distributions, we can write

$$Y_i = \sum_{k=1}^{K} \alpha_k (X_i\beta_k + Z_i b_{i(k)} + \epsilon_{i(k)}) \tag{4}$$

where the α_k's are the mixing coefficients:

Under this model formulation, each component, k, is distinct, uniquely defined by β_k, Ψ_k, σ_k^2, and Λ_k. The log-likelihood of the mixture model can be defined as:

$$\sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} \alpha_k \exp\left\{l_k(\beta_k, \sigma_k^2, \Psi_k, \Lambda_k \mid y_i, b_{i(k)})\right\}\Big) \tag{5}$$

where, $l_k(\cdot|\cdot)$ is given in Equation (3). Direct maximization of the log-likelihood can be quite difficult due to the sum of terms inside the logarithm. However, we can again complete the data by considering the unobserved latent indicator variables $\delta_{i(k)}$, which is equal to 1 if observation i belongs to cluster k and 0 otherwise, and write the complete data log-likelihood (Celeux et al., 2005):

$$l = \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{\delta_{i(k)} \log(\alpha_k) + \delta_{i(k)} l_k(\beta_k, \sigma_k^2, \Psi_k, \Lambda_k \mid y_i, b_{i(k)})\Big\} \tag{6}$$

where l_k is as in equation (3). Thus, with this "double completion" of the data, maximum likelihood estimates of the parameter vector $\theta = (\alpha, \beta, \sigma^2, \Psi, \Lambda)$ can be obtained using an EM approach as described in the next section.

3 Model estimation and inference

3.1 EM algorithm

The EM algorithm consists of iterating until convergence between the following E- and M-steps. At iteration $q > 0$, the E-step consists of computing the expectation

of the 'complete' log-likelihood knowing the observed data and a current value for the parameters $\theta^{[q]} = (\alpha^{[q]}, \beta^{[q]}, \Psi^{[q]}, (\sigma^2)^{[q]}, \Lambda^{[q]})$, i.e.:

$$Q(\theta \mid \theta^{[q]}) = E\left[l(\theta \mid y, \delta, b) \mid y, \theta^{[q]}\right] = \sum_{i=1}^{n}\sum_{k=1}^{K}\left(\tau_{i(k)}^{[q]}\log(\alpha_k) + \tau_{i(k)}^{[q]}E\left[l_k(\beta_k, \sigma_k^2, \Psi_k, \Lambda_k \mid y_i, b_{i(k)}) \mid y, \theta^{[q]}\right]\right)$$

where the complete data log-likelihood l is given in Equation (6), and $\tau_{i(k)}^{[q]} = E[\delta_{i(k)}|y, \theta^{[q]}]$ are the so-called posterior probabilities of component membership, computed by:

$$\tau_{i(k)}^{[q]} = P(i \in C_k \mid y_i, \theta^{[q]}, \alpha^{[q]}) = \frac{\alpha_k^{[q]} g_{m_i}(y_i \mid \theta_k^{[q]})}{\sum_{l=1}^{K}\alpha_l^{[q]} g_{m_i}(y_i \mid \theta_l^{[q]})}$$

where C_k denotes the k-th cluster and $g_{m_i}(y_i|\theta_k^{[q]}) = \exp(l_k(\theta_k^{[q]}|y_i))$ denotes the density of the k-th mixture component. The M-step consists of setting $\theta^{[q+1]} = \arg\max_\theta Q(\theta \mid \theta^{[q]})$. Details are given in Appendix A. Suffice it to say here that the M-step uses the same numerical methods for the estimation of the parameters of the Λ_i matrix as in the R functions `lme` and `gls`; therefore, though our approach follows essentially Celeux et al. (2005), it also borrows from Pinheiro and Bates (2000).

3.2 Initial values

It is well known that the EM algorithm can be quite sensitive to the choice of starting values. A number of different strategies for choosing starting values have been proposed(McLachlan and Peel, 2000). Following Celeux et al.(2005), we perform a large number of short runs(10 iterations) of the EM from different k-means results. The starting values which initialized the best "short run" solution (i.e. the short-run solution to achieve the highest log-likelihood), are then selected as starting values. When the number of observations is not equal across individuals, k-means is performed on regression parameters obtained from linear regressions on each individual.

3.3 Standard Errors

The asymptotic covariance matrix of the maximum-likelihood estimates, $\hat{\theta}$, is equal to the inverse of the expected Information matrix, $\mathcal{I}(\theta)$, which can be approximated by following Louis(1982)'s decomposition of $\mathcal{I}(\hat{\theta})$:

$$\mathcal{I}(\hat{\theta}) = \mathrm{E}_\eta(B(y,\theta)) - \mathrm{E}_\eta(S(y,\theta)S^T(y,\theta)) + \mathrm{E}_\eta(S(y,\theta))\mathrm{E}_\eta(S^T(y,\theta)) \qquad (7)$$

where η represents the *missing data*, y, the *observed data* and:

$$B(y,\theta) = \frac{\partial^2 \log l(\theta \mid y)}{\partial\theta^2}, \quad S(y,\theta) = \frac{\partial \log l(\theta \mid y)}{\partial\theta} \qquad (8)$$

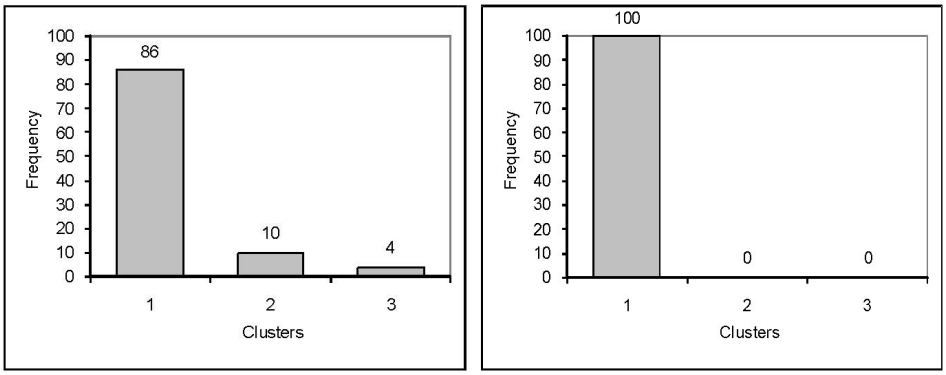


Figure 1: Histograms of number of clusters retrieved by the algorithm when 1 cluster is simulated using AIC(left panel) and BIC(right panel).

The standard errors of $\hat{\theta}$ are then given by the diagonal elements of $(\mathcal{I}^{-1}(\hat{\theta}))^{1/2}$. See Appendix B for details.

3.4 Assessing the Number of Clusters

Assessing the "correct" number of components or clusters in finite mixture models is a fundamental and challenging question. The minimum AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) rules are popular choices and both are presented in our analysis. The performance of the minimum AIC and BIC rules is investigated in simulation studies presented in the next Section.

4 Evaluation through simulation

We performed a limited simulation study. We varied K from 1 to 7. For every fixed K, we generated 100 samples from K multivariate normal distributions with exchangeable variance-covariance matrices and expectations linear in time; we then applied our method to estimate the parameters and chose the number of clusters using both the AIC and the BIC. This was repeated 100 times. The detailed forms of the distributions were chosen so as to mimic the results of the example described below. We give in Figures 1, 2 and 3, the results for $K = 1$, 3 and 6 in the form of frequencies of number of retrieved class within the 100 repetitions. As it can be seen, both criteria perform reasonably well, with a tendency towards more conservative choices for the BIC and more liberal ones for the AIC. Though we have not carried out systematic explorations beyond those reported here, our experience suggests that the behaviour of the AIC and BIC is essentially the same in many situations.

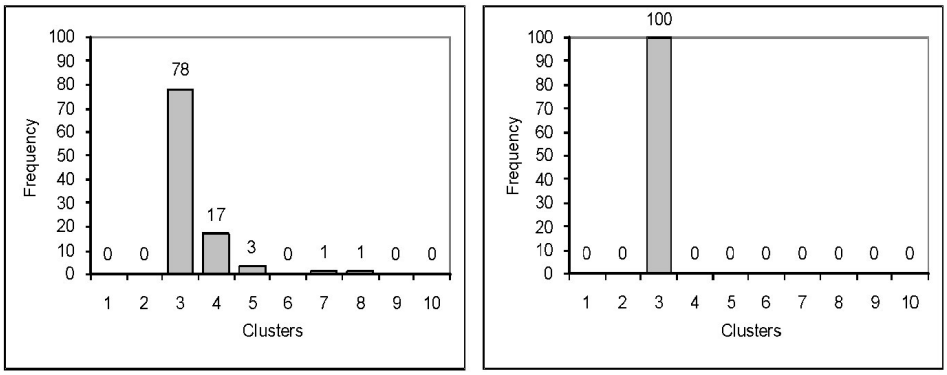


Figure 2: Histograms of number of clusters retrieved by the algorithm when 3 clusters are simulated using AIC(left panel) and BIC(right panel).

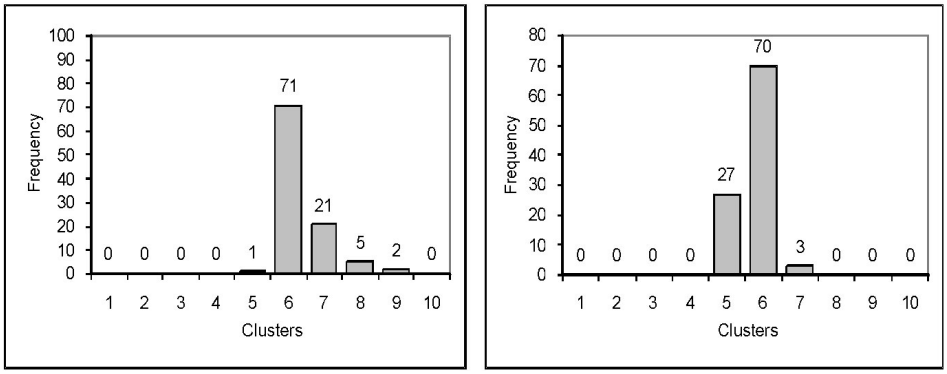


Figure 3: Histograms of number of clusters retrieved by the algorithm when 6 clusters are simulated using AIC(left panel) and BIC(right panel).

5 Disease Trajectory Data: Delirium

Delirium is a condition often encountered in hospitalized elderly populations. The Delirium Index (DI), is a validated measure of delirium severity developed at St. Mary's Hospital, (McCusker et al. 2004). DI scores range from 0 to 21 and higher scores indicated more severe delirium. We used data from 229 St. Mary's patients hospitalized between 1996 and 1999. Patients were evaluated with the DI at enrolment, and several times during the following 15 days. Measurement times were unequally spaced and differed across individuals. In order to account for correlation among repeated measurements, four different models were fit:

Independence

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2 Id)$$

Component	α	β	σ^2	Ψ	ϕ
1 – Steady Course	0.153	11.692, -0.015	3.912	3.428	0.438
2 – Fluctuating	0.224	9.956, 0.0078	23.171	2.059	0.387
3 – Worsening	0.204	5.742, 0.032	1.340	2.602	0.001
4 – Recovery	0.139	4.451, -0.343	2.912	0.550	0.615
5 – Fluctuating Recovery	0.279	8.569, -0.316	6.105	2.750	0.215

Table 1: Parameter Estimates for 5 cluster AR(1) and random intercept solution. Except for the slopes of components 1 and 2, all parameters are significantly different from zero at the 0.05 level according to likelihood ratio tests.

Random Intercept

$$y_i = \beta_0 + \beta_1 x_i + b_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2 Id)$$

Autoregressive

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2 R_i), \qquad R_i = AR(1)(\phi)$$

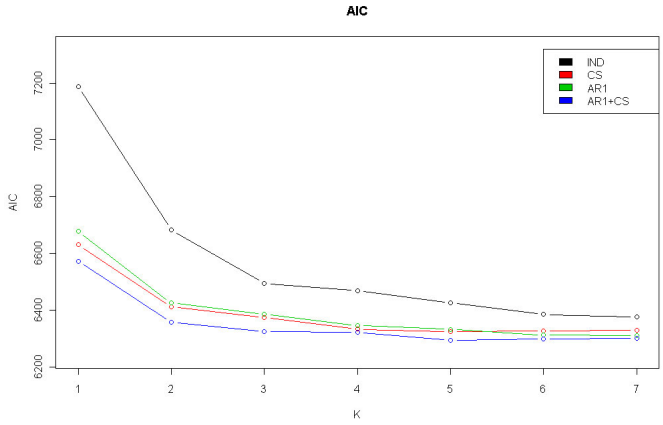
Random intercept and autoregressive

$$y_i = \beta_0 + \beta_1 x_i + b_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2 R_i), \qquad R_i = AR(1)(\phi)$$

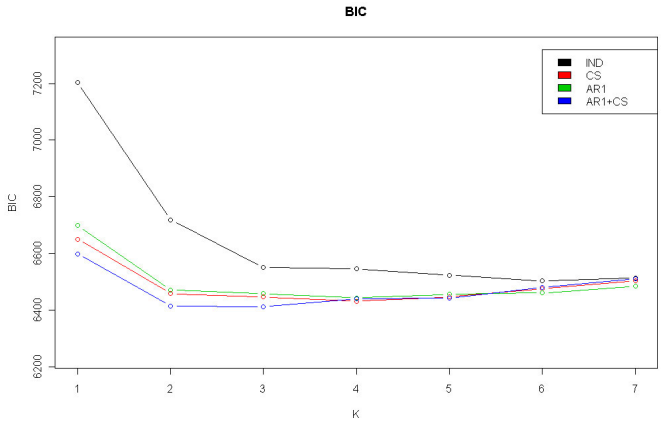
The DI curves have a great variety of shapes and by fitting a mixture model to the longitudinal data, it is our goal to reduce these shapes to a few "typical" ones which may be interpreted as distinct courses of the illness. Figure 4 shows AIC and BIC values for the 28 different models fit. A 5-component model with both AR(1) correlation and random intercept is selected for further investigation as it provides good clinical interpretation and is the best model according to the AIC. The mixture model has log-likelihood of -3118.480, AIC = 6284.961 and BIC = 6407.968. Parameter estimates appear in Table 1.

5.1 Interpretation

The **steady course** component represents a course that is quite stable: the slope is negligible (and non-significant at the 0.05 level) and the variance parameter reasonably small. The **fluctuating course** component is similar to the first, but the variance is large, suggesting that patients fluctuate around a stable state. The component named **worsening** has a positive slope, indicating a DI that increases in time, hence a worsening of the delirium. The remaining two components have negative slopes and are therefore named **recovery**: however, in one of them we have a high variance and therefore we qualify the recovery as fluctuating. From a general point of view, differences in the variance parameters may seem uninteresting. However in delirium studies it is very important to identify components with large fluctuations since fluctuating severity is considered a fundamental characteristics of 'true' delirium.



(a) AIC



(b) BIC

Figure 4: AIC and BIC for different models across K

5.2 Remarks

A few comments are in order. Firstly, the four models considered in this analysis do not exhaust the possibilities of the ELMM. For example we could have fitted a model with random intercept and slope; unfortunately further exploration was limited by computational power. However, we have chosen the four models that are most currently used in biostatistical practice when analyzing longitudinal data: they reflect simple and intuitive hypotheses as to how correlation might arise.

Secondly, it should be noted that, as the AIC and BIC curves show, the selection of the number of clusters depends on the model. While this may be seen as a limitation, it is by no means an uncommon occurrence: for example this depen-

dence is commonly observed when working with mixtures of multivariate normal distributions and allowing hypotheses other than homoscedasticity, e.g. using the mclust R package (Banfield and Raftery, 1993).

Thirdly, the determination of the number of classes remains a fairly subjective exercise. Indeed neither the AIC nor the BIC provide absolutely objective criteria, as is demonstrated by the numerous alternatives proposed in the litterature. In this work we have not attempted to develop new approaches, but have limited ourselves to the most popular ones.

Fourthly, though we have shown by limited simulations that the BIC works better than the AIC in our context, we have actually preferred to retain the 5-cluster solution corresponding to the minimum AIC rather than the 2-cluster solution which minimizes the BIC. This illustrates both the limits and the advantages of using a certain degree of subjectivity. Indeed, the BIC of the 5-cluster solution is not very different from the minimum BIC. On the other hand, a five cluster solution was proposed in a previous work on delirium by Sylvestre et al. (2006), who applied an exploratory approach combining principal component analysis with k-means clustering. The interpetation of our 5-clusters is very similar to the interpretation of the five clusters found by these authors, yet it has the advantage of being model based.

Finally, though we have not studied robustness systematically, we have found that the point estimates of the fixed effect coefficients are fairly stable regardless of whether or not we include the random effect and/or the AR(1) term. This is encouraging, but further explorations are desirable.

6 Time-course gene expression data

Microarray analysis is a valuable tool in molecular biology, as it permits to assess the expression levels of a large number of genes simultaneously. In view of the complexity of biological networks, it is useful to study gene expression not only at a specific point in time, as in early microarray experiments, but also longitudinally. Expression time profiles can indeed be very useful to find co-regulated and functionally related groups of genes. We analysed a set of longitudinal gene expression data already studied by Heard et al. (2002). The data consists of 2771 gene expression time profiles (each with 6 non-equally spaced observations) from mosquitoes which have been infected with a bacterial agent. Visualization of the raw data is not very informative (Figure 5).

Heard et al. (2002) proposed a Bayesian model-based hierarchical clustering algorithm to cluster genes having similar expression profiles that led to a 17-cluster solution. From this, interpretable graphs were obtained. Their solution assumed data to be uncorrelated, so that in practice their model is a mixture of ordinary regressions. In contrast, we used the wealth of submodels within the ELMM to find a non-trivial correlation structure that fits the data. To model the trajectories, we used a flexible family of basis functions called the truncated power spline basis, as in Heard et al. (2002):

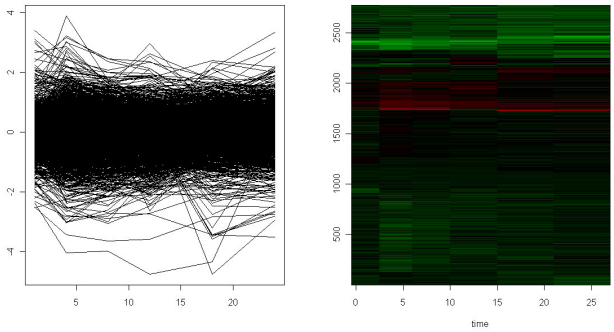


Figure 5: and heat map of the *Salmonella typhi* data presented in Heard et al. (2002)

$$g_1(t_{is}) = 1 \qquad g_j(t_{is}) = (t_{is} - t_{i,j-1})_+, s = 1, ..., n, j = 2, ..., m_i$$

where $(\cdot)_+$ is the positive part function.

We fit three different correlation structures: an independence model, an autoregressive model as well as a model with random intercept and autoregressive structure. We varied K from 1 to 26. According to the BIC, see Figure 6, the best model is a 14-cluster model with both random intercept and autoregressive structure. A heatmap of the data is presented in Figure 7. Comparing the fitted values(right) across clusters gives a visual measure of between cluster heterogeneity. Comparing the fitted values(right) to the raw observations(left) gives a sense of within cluster homogeneity. The clustering is shown in more detail in Figure 8.

7 Discussion

We have presented a straightforward method for modeling heterogeneity in longitudinal data. We have proposed a mixture of regressions with components in the Extended Linear Mixed Model (ELMM) (Pinheiro and Bates, 2000). The ELMM consists of a random effect portion (LMM) extended by the addition of an error term with correlation matrix defined up to a small number of parameters to be estimated from data. We have limited ourselves to an autoregressive error term. Our approach to parameter estimation is based on the EM algorithm of Celeux et al. (2005) for mixtures of LMM, augmented by numerical methods which are essentially those used in the `lme` and `gls` R functions of Pinheiro and Bates (2000). The theoretical novelty of this approach is modest: it permits, on the one hand, to deal with correlated errors of a type that is important in applied research, and, on the other, suggests further extension to a catalog of possible error correlation structures such as those contained in the `lme` and `gls` R functions. Although other authors have considered mixtures of regressions for longitudinal data, no one has

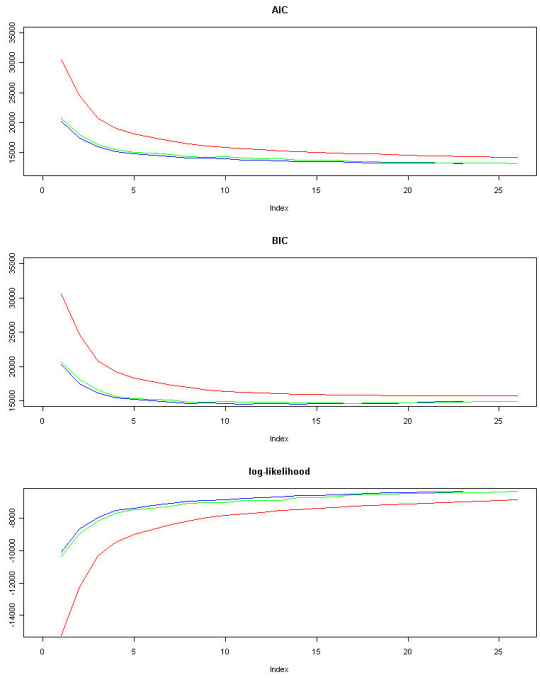


Figure 6: log-likelihood, AIC and BIC plots for different models(independence(red), autoregressive(green) and autoregressive with random intercept(blue)) fit for $K = 1,...26$

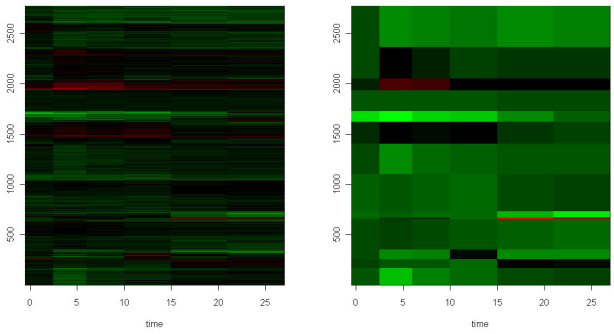


Figure 7: Clustered gene expression profiles form the *Salmonella typhi* data

yet achieved the generality that can be achieved with the ELMM. An appropriate modeling of the correlation structure of longitudinal data is important: not only does it provide useful insight into the dynamical process under study, but it also leads to fewer clusters, hence to a more economical model of the data. We

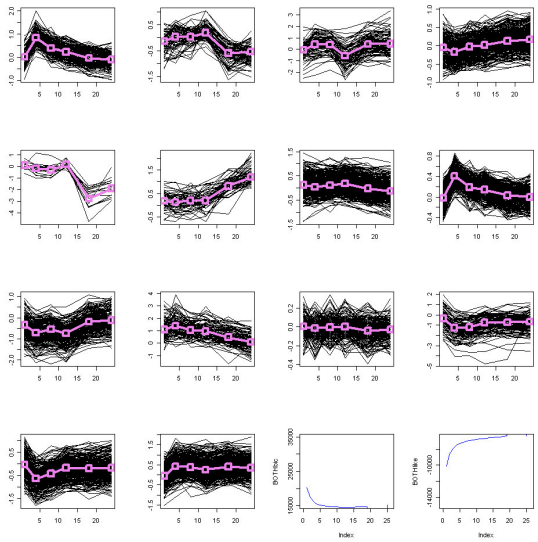


Figure 8: 14 cluster solution suggested by autoregressive AR(1) model with random intercept model. BIC = 14528.19

have also shown that mixtures of regressions offer an important tool in classifying course of diseases from clinical data and longitudinal gene expression data, providing easy-to-interpret analyses.

Further research will aim to speed up the EM algorithm, which will also allow us to study even richer correlation structures. Amelioration of computing efficiency will allow us to carry out more extensive simulations and to study the robustness of key features of our models, e.g. fixed effect parameter estimates and selection of the number of clusters. We plan also to revisit the Bayesian approach of Heard et al. (2002).

References

[1] Banfield J.D., Raftery A.E. (1993). "Model based Gaussian and non Gaussian clustering." *Biometrics*, **49**: 803–821.

[2] Belin, T.R, Rubin, D.B. (1995). "The analysis of repeated-measures data on schizophrenic reaction times using mixture models." *Statistics in Medicine* **14**: 747–768.

[3] Celeux, G., Lavergne, C., Martin, O. (2005). "Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments". *Statist. Model.* **5**: 243–267.

[4] De la Cruz-Mesía, R., Quintana, F.A., Marshall, G. (2008) "Model Based Clustering for Longitudinal Data." *Computational Statistics and Data Analysis* **52**: 1441–1457.

[5] Gaffney, S.J., Smyth, P. (2003). "Curve clustering with random effects regression mixtures." In: Bishop, C.M., Frey, B.J. (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, KeyWest, FL.

[6] Heard, Holmes, Stephens (2006). "A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves." *Journal of the American Statistical Association* **101**(473): 18–29.

[7] Jennrich and Schluchter (1986) *Unbalanced Repeated-Measures Models with Structured Covariance Matrices*

[8] Luan Y., Li H. (2003). "Clustering of time course gene expression data using a mixed-effect model with B-splines." *Bioinformatics*, **19**, 474–482.

[9] McCusker J, Cole M, Dendukuri N, Belzile E. (2004). "The Delirium Index, a Measure of the Severity of Delirium: New Findings on Reliability, Validity, and Responsiveness". *Journal of the American Geriatric Society*, **52**:1744-1749.

[10] Ng, S. K., McLachlan, G. J., Wang, K., Ben-Tovim Jones, L., Ng, S. W. (2006) "A Mixture model with random-effects components for clustering correlated gene-expression profiles." *Bioinformatics* **22**: 1745–1752.

[11] Pinheiro J.C., Bates D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

[12] Tango T. (1998). "A mixture model to classify individual profiles of repeated measurements." In: *Data Science, Classification and Related Topics* (eds by C. Hayashi, et al.), Springer-Verlag, Tokyo, 247–254.

[13] Sylvestre, M. P., McCusker, L., Cole, M., Regeasse, A. Belzile, E., and Abrahamowicz, M. (2006). "Classification of patterns of delirium severity scores over time in an elderly population." *International Psychogeriatrics*, **18**: 667–680.

A Details on the M-step of the EM algorithm

We have for the q+1 th step:

$$\alpha_k^{[q+1]} = \sum_{i=1}^{n} \frac{\tau_{i(k)}^{[q]}}{n}$$

$$\beta_k^{[q+1]} = \left[\left(\sum_{i=1}^{n} \tau_{i(k)}^{[q]} \left((y_i - Z_i b_{i(k)}^{[q]}) \right)^T ((\Lambda_{ik}^{[q]})^{-1})(-X_i) \right) \left(\sum_{i=1}^{n} \tau_{i(k)}^{[q]} (X_i)^T ((\Lambda_{ik}^{[q]})^{-1})(-X_i) \right)^{-1} \right]^T$$

$$(\sigma_k^2)^{[q+1]} = \frac{1}{\sum_{i=1}^{n} \tau_{i(k)^{[q]}} m_i} \sum_{i=1}^{n} \tau_{i(k)^{[q]}} \left[E(e_{ik}^T \Lambda_{ik}^{-1} e_{ik} \mid y_i, \theta^{[q-1]}) \right]$$

where

$$E(e_{ik}^T \Lambda_{ik}^{-1} e_{ik} \mid y_i, \theta^{[q-1]}) = \sigma_k^{4[q-1]} (y_i - X_i \beta_k^{[q-1]})^T \Sigma_{ik}^{-1[q-1]} \Lambda_{ik}^{[q-1]} \Sigma_{ik}^{-1[q-1]} (y_i - X_i \beta_k^{[q-1]}) + m_i \sigma_k^{2[q-1]} - \sigma_k^{4[q-1]} tr(\Sigma_{ik}^{-1[q-1]} \Lambda_{ik}^{[q-1]})$$

and

$$\Psi_k^{[q+1]} = \frac{1}{\sum_{i=1}^{n} \tau_{i(k)}^{[q]} m_i} \sum_{i=1}^{n} \tau_{i(k)}^{[q]} E(b_{i(k)} b_{i(k)}^T \mid y_i, \theta^{[q-1]})$$

where

$$E(b_{i(k)} b_{i(k)}^T \mid y_i, \theta^{[q-1]}) = E(b_{i(k)} \mid y_i, \theta^{[q-1]}) E(b_{i(k)} \mid y_i, \theta^{[q-1]})^T + m_i \Psi_k^{[q-1]} - \Psi_k^{[q-1]} Z_i^T \Sigma_{ik}^{-1[q-1]} Z_i \Psi_k^{[q-1]}$$

and

$$E(b_{i(k)} \mid y_i, \theta^{[q-1]}) = \Psi_k^{[q-1]} Z_i^T \Sigma_{ik}^{-1[q-1]} (y_i - X_i \beta_k^{[q-1]})$$

Finally, consider the positive-definite matrices Λ_{ik} (there are nK of these). There are different ways that such matrices may be parametrized, depending on assumptions regarding the intra-individual covariance structure (Pinheiro and Bates, 2000). Let ϕ_k denote the set of parameters used in the parametrization of $\{\Lambda_{ik}\}_{i=1,\ldots,n}$. To estimate these parameters at iteration $[q]$, we use numerical maximization methods (e.g the `R` function `nlminb()`).

B Details on computing the standard error of $\hat{\theta}$

Jennrich and Schluchter(1986) provide equations for the required score vector statistics and Hessian matrix in the homogeneous model. The required first and second derivatives for the mixture model are presented bellow. Derivatives with respect to the parameters that define Λ, (ϕ), must be calculated by numerical methods.

$$\frac{\partial l(\theta \mid Y)}{\partial \alpha_k} = \sum_{i=1}^{n} \frac{\delta_{i(k)}}{\alpha_k} - \frac{\delta_{i(K)}}{\alpha_K}, \quad {\scriptstyle k=1,\ldots,K-1} \tag{9}$$

$$\frac{\partial l(\theta \mid Y)}{\partial \beta_k} = \sum_{i=1}^{n} \delta_{i(k)} (X_i^T)(\Sigma_{ik}^{-1}) (y_i - X_i \beta_k), \quad {\scriptstyle k=1,\ldots,K} \tag{10}$$

$$\frac{\partial l(\theta \mid Y)}{\partial \sigma_k^2} = \sum_{i=1}^{n} \frac{\delta_{i(k)}}{2} \operatorname{tr}(\Sigma_{ik}^{-1}(y_i - X_i\beta_k)(y_i - X_i\beta_k)\Sigma_{ik}^{-1}\Lambda^{-1}), \quad {\scriptstyle k=1,\ldots,K} \tag{11}$$

$$\frac{\partial l(\theta \mid Y)}{\partial \Psi_k} = \sum_{i=1}^{n} \frac{\delta_{i(k)}}{2} \operatorname{tr}(\Sigma_{ik}^{-1}(y_i - X_i\beta_k)(y_i - X_i\beta_k)\Sigma_{ik}^{-1}Z_iZ_i^T), \quad {\scriptstyle k=1,\ldots,K} \tag{12}$$

$$\frac{\partial^2 l(\theta \mid Y)}{\partial \alpha_k^2} = \sum_{i=1}^{n} \frac{-\delta_{i(k)}}{\alpha_k^2} - \frac{\delta_{i(K)}}{\alpha_K^2}, \quad {\scriptstyle k=1,\ldots,K-1} \tag{13}$$

$$\frac{\partial^2 l(\theta \mid Y)}{\partial \beta_k^2} = \sum_{i=1}^{n} -\delta_{i(k)}(X_i^T\Sigma_{ik}^{-1}X_i), \quad {\scriptstyle k=1,\ldots,K} \tag{14}$$

$$\frac{\partial^2 l(\theta \mid Y)}{\partial (\sigma_k^2)^2} = \sum_{i=1}^{n} \frac{-\delta_{i(k)}}{2} \operatorname{tr}(\Sigma_{ik}^{-1}\Lambda_{ik}\Sigma_{ik}^{-1}(2(y_i - X_i\beta_k)(y_i - X_i\beta_k)^T - \Sigma_{ik})\Sigma_{ik}^{-1}\Lambda_{ik}), \quad {\scriptstyle k=1,\ldots,K} \tag{15}$$

$$\frac{\partial^2 l(\theta \mid Y)}{\partial \sigma_k^2 \partial \beta_{jk}} = \sum_{i=1}^{n} -\delta_{i(k)}X_{jk}(\Sigma_{ik}^{-1}\Lambda_{ik}\Sigma_{ik}^{-1}(y_i - X_i\beta_k)), \quad {\scriptstyle k=1,\ldots,K \quad j=1,\ldots,p} \tag{16}$$

$$\frac{\partial^2 l(\theta \mid Y)}{\partial \Psi_k^2} = \sum_{i=1}^{n} \frac{-\delta_{i(k)}}{2} \operatorname{tr}(\Sigma_{ik}^{-1}Z_iZ_i^T\Sigma_{ik}^{-1}(2(y_i - X_i\beta_k)(y_i - X_i\beta_k)^T - \Sigma_{ik})\Sigma_{ik}^{-1}Z_iZ_i^T), \quad {\scriptstyle k=1,\ldots,K} \tag{17}$$

$$\frac{\partial^2 l(\theta \mid Y)}{\partial \Psi_k \partial \beta_{jk}} = \sum_{i=1}^{n} -\delta_{i(k)}X_{jk}(\Sigma_{ik}^{-1}Z_iZ_i^T\Sigma_{ik}^{-1}(y_i - X_i\beta_k)), \quad {\scriptstyle k=1,\ldots,K \quad j=1,\ldots,p} \tag{18}$$