Research Article

Statistics in Medicine

Received 26 October 2010, Accepted 21 July 2011 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4365

A mixture model with random-effects components for classifying sibling pairs

F. Martella,[a*†] J. K. Vermunt,[b] M. Beekman,[c,d] R. G. J. Westendorp,[e] P. E. Slagboom[c,d] and J. J. Houwing-Duistermaat[f]

In healthy aging research, typically multiple health outcomes are measured, representing health status. The aim of this paper was to develop a model-based clustering approach to identify homogeneous sibling pairs according to their health status. Model-based clustering approaches will be considered on the basis of linear mixed effect model for the mixture components. Class memberships of siblings within pairs are allowed to be correlated, and within a class the correlation between siblings is modeled using random sibling pair effects. We propose an expectation–maximization algorithm for maximum likelihood estimation. Model performance is evaluated via simulations in terms of estimating the correct parameters, degree of agreement, and the ability to detect the correct number of clusters. The performance of our model is compared with the performance of standard model-based clustering approaches. The methods are used to classify sibling pairs from the Leiden Longevity Study according to their health status. Our results suggest that homogeneous healthy sibling pairs are associated with a longer life span. Software is available for fitting the new models. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: model-based clustering; family data; random effects; healthy aging

1. Introduction

In healthy aging research, typically, data on subjects older than a specific age threshold are collected, and health is measured via several health-related variables such as blood pressure, cholesterol levels, and mini mental state examination. A formal definition for being healthy is lacking, and subjects are for example assigned a healthy status when they have a low blood pressure, low cholesterol levels, and a high score for the mental state examination questionnaire. Identification of genetic factors for healthy aging is hampered by the lack of a formal definition. Instead of using an arbitrary definition for health status, we consider in this paper a latent variable approach in which the latent variable represents health. To increase the probability for segregation of genetic factors, genetic studies on healthy aging use a family design of families with multiple 'old' subjects. For example, the Leiden Longevity Study ([1]) and the European study Genetics of Healthy Aging ([2]) collect data on nonagenarian sibling pairs. Thus, a latent variable approach that is applicable to genetic studies should be able to deal with the family structure. The aim of this paper was to develop a method to classify sibling pairs according to their health status using a model-based approach. The statistical challenge is to deal with the correlation within sibling pairs and to include the specific features of the outcome variable healthy aging.

Genetic linkage analysis is an often used approach to identify new chromosomal regions which harbor genes involved in the etiology of the traits. To enhance gene finding, it has been advocated to first identify

[a] *Dipartimento di Scienze Statistiche, Facoltá di Ingegneria dell'Informazione, Informatica e Statistica, Sapienza Universitá di Roma, Rome, Italy*
[b] *Department of Methodology and Statistic, Tilburg University, Tilburg, The Netherlands*
[c] *Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands*
[d] *Netherlands Consortium for Healthy Ageing, Leiden, The Netherlands*
[e] *Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands*
[f] *Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands*
**Correspondence to: F. Martella, Dipartimento di Scienze Statistiche Facoltá di Scienze Statistiche, Sapienza Universitá di Roma P.le Aldo Moro, 5 - I 00185 Rome, Italy.*
[†] *E-mail: francesca.martella@uniroma1.it*

subtypes of traits by using cluster methods based on measures of distance [3, 4]. Alternatively, model-based cluster methods can be used to determine subtypes of various disorders [5–7]. These methods have also been applied to family data, but the dependence between outcomes of family members is ignored by these methods. In contrast, Labbe *et al*. [8] proposed a model-based clustering approach for nuclear families, and recently, their model was extended to deal with larger pedigrees [9]. Their models assume that class memberships of relatives are independent, conditional on the class memberships of the parents. However, this assumption is violated when, within sibships, additional correlation is present because of shared environmental factors. In addition, their model can only be used when parental information is available. In human aging studies, information on the parents is typically not available.

In this paper, we propose to cluster data from sibling pairs by using mixtures of random effect models. Here, the random effects represent the shared genetic and environmental effects of siblings. Dependence among observations has been considered in model-based clustering approaches. A hierarchical mixture model with non-parametric random effects in a mixture model was proposed to capture the hierarchical structure of subjects within families [10] (which is itself a nonparametric random-effects model; see [11]). A mixture of linear mixed effect models (LMMs) was proposed to cluster repeated gene expression data [12]. This work was further developed by Ng *et al*. [13] for clustering correlated gene expression profiles obtained from various experimental designs. Our model is an extension of the mixtures of LMMs and allows for a more flexible structure of the component-specific covariance matrix of the random effects needed to appropriately model covariances within sibling pairs. The proposed model is implemented in the software LATENT GOLD program for latent variable modeling (Statistical Innovations Inc., Statistical Innovations, Belmont, MA USA, [14]). The proposed model basically assumes: (i) each sibling pair is a vector of repeated observations; (ii) the class membership of the two siblings is allowed to be correlated (not with a random effect but simply with an association parameter); (iii) after accounting for the correlation between the two siblings' class memberships, variables are still correlated between the two siblings of a pair. This correlation is modeled by sibling pair-specific random effects; (iv) the correlation between variables of sibling pairs is allowed to vary across outcome variables; (v) the correlation between various variables within a sibling and between two siblings is modeled. The difference between our proposed model for sibling pair data and the currently available models [12, 13] will be described in detail in the method section.

As illustration, we apply the method to data on multiple health outcomes observed for nonagenarian sibling pairs from the Leiden Longevity Study. Data on six health variables are available. To enhance gene finding for healthy aging, the goal is to identify concordant healthy sibling pairs and discordant healthy sibling pairs. The sibling pairs will be classified in three categories, namely sibling pairs in good shape (concordant healthy sib-pairs CH), sibling pairs in bad shape (concordant unhealthy sib-pairs CUH), and one sibling in good shape and the other in bad shape (discordant sib-pairs DH).

The outline of this paper is as follows. In Section 2.1, we introduce our linear mixed-effect model to classify sibling pairs and describe the expectation–maximization (EM) algorithm for estimation of the model parameters. In Section 2.2, we consider the specific situation related to the classification of sibling pairs in concordant healthy, concordant unhealthy, and discordant pairs. In Section 3, the performance of the model to classify sibling pairs according to their health status is studied via simulations. We compare the performance of our new model with standard model-based clustering. In Section 4, the results obtained from applying the new model to data from the Leiden Longevity Study are presented. In the last section, some concluding remarks are given, and outlines for potential future research directions are provided. In the Appendix, details on syntax language for running the empirical example with the LATENT GOLD software are provided.

2. Linear mixed-effect model for clustering of sibling pairs

We consider the clustering of n families $\mathbf{y}_i$ $(i = 1, \ldots n)$, where we let $\mathbf{y}_i = (\mathbf{y}'_{i1}, \mathbf{y}'_{i2})'$ represent two siblings in the ith family, and $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijH})'$ $(j = 1, 2)$ contains H outcomes measured on the jth sibling in the ith family. The observed $2H$-dimensional vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are assumed to follow a discrete mixture of Gaussian distributions (components) with unknown proportions $\pi_1, \ldots, \pi_K (\sum_{k=1}^{K} \pi_k = 1)$ as follows:

$$f(\mathbf{y}_i|\boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k N(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ represent the component-specific mean vectors and covariance matrices, respectively ($k = 1, \ldots, K$). To take into account the dependence between siblings, we add the assumption that siblings within the same family belong to the same mixture component. In this way, we are able to model the covariance structure between sibling pairs from the same family. To model the correlation between siblings in the mixture model (1), it is assumed that, conditional on each kth component, $\mathbf{y}_i$ has the following structure:

$$\mathbf{y}_i = \boldsymbol{\beta}_k + \mathbf{X}\mathbf{u}_{ik} + \mathbf{e}_i, \tag{2}$$

where $\boldsymbol{\beta}_k$ represents the $2H$-dimensional component-specific fixed effect vector modeling the conditional mean of $\mathbf{y}_i$ in the kth component. $\boldsymbol{\beta}_k$ has the following form:

$$\boldsymbol{\beta}_k = \left(\boldsymbol{\beta}_k^{1'}, \boldsymbol{\beta}_k^{2'}\right)' = \left(\beta_{1k}^1, \beta_{2k}^1, \ldots, \beta_{Hk}^1, \beta_{1k}^2, \beta_{2k}^2, \ldots, \beta_{Hk}^2\right)',$$

with $\boldsymbol{\beta}_k^j$ representing the component-specific fixed effect vector on the jth sibling and β_{hk}^j the component-specific fixed effect for the jth sibling corresponding to the hth variable ($j = 1, 2$; $h = 1, \ldots, H$). The term $\mathbf{u}_{ik}$ appearing in Equation (2) represents the component-specific H-dimensional vector of random effects which is used to induce dependence between sibling pairs in the same family, whereas $\mathbf{X}$ is a $2H \times H$ known design matrix associated to random effects, where

$$\mathbf{X} = \begin{bmatrix} \mathbf{I}_H \\ \mathbf{I}_H \end{bmatrix}$$

where $\mathbf{I}_H$ is a $H \times H$ identity matrix. The $2H$-dimensional component-specific error vector $\mathbf{e}_i = (\tilde{\mathbf{e}}_i', \tilde{\mathbf{e}}_i')'$, where $\tilde{\mathbf{e}}_i$ is the H-dimensional error vector refereed to each sibling in the ith family. The error vector $\mathbf{e}_i$ and the random effect $\mathbf{u}_{ik}$ are assumed to be mutually independent. The distributions of $\mathbf{e}_i$ and $\mathbf{u}_{ik}$ are taken to be multivariate normal $\mathrm{N}(\mathbf{0}, \tilde{\mathbf{F}}_k)$ and $\mathrm{N}(\mathbf{0}, \mathbf{S}_k)$, respectively. $\tilde{\mathbf{F}}_k$ is a block diagonal matrix where the off-block elements are equal to null matrices indicating that within a cluster the residuals of two siblings are independent. Formally, we assume that

$$\tilde{\mathbf{F}}_k = \begin{pmatrix} \mathbf{F}_k & \mathbf{0}_H \\ \mathbf{0}_H & \mathbf{F}_k \end{pmatrix} = \begin{pmatrix} f_{11k} & f_{12k} & \ldots & f_{1Hk} & 0 & 0 & \ldots & 0 \\ f_{12k} & f_{22k} & \ldots & f_{2Hk} & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ f_{1Hk} & f_{2Hk} & \ldots & f_{HHk} & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & f_{11k} & f_{12k} & \ldots & f_{1Hk} \\ 0 & 0 & \ldots & 0 & f_{12k} & f_{22k} & \ldots & f_{2Hk} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & f_{1Hk} & f_{2Hk} & \ldots & f_{HHk} \end{pmatrix}$$

where $\mathbf{F}_k$ is equal for both sibling pairs and represents the variability due to random measurement error in the kth cluster, whereas $\mathbf{0}_H$ represents a $H \times H$ null matrix. $\mathbf{S}_k$ is defined as:

$$\mathbf{S}_k = \begin{pmatrix} s_{11k} & s_{12k} & \ldots & s_{1Hk} \\ s_{12k} & s_{22k} & \ldots. & s_{2Hk} \\ \ldots & \ldots & \ldots & \ldots \\ s_{1Hk} & s_{2Hk} & \ldots. & s_{HHk} \end{pmatrix}$$

and represents the variability due to the random effects in the kth cluster. Thus, conditional on belonging to the kth component, the resulting mean vector and covariance matrix are respectively $\boldsymbol{\mu}_k = \boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_k = \mathbf{X}\mathbf{S}_k\mathbf{X}' + \tilde{\mathbf{F}}_k$, where

$$\boldsymbol{\Sigma}_k = \begin{pmatrix} f_{11k}+s_{11k} & f_{12k}+s_{12k} & \ldots & f_{1Hk}+s_{1Hk} & s_{11k} & s_{12k} & \ldots & s_{1Hk} \\ f_{12k}+s_{12k} & f_{22k}+s_{22k} & \ldots & f_{2Hk}+s_{2Hk} & s_{12k} & s_{22k} & \ldots & s_{2Hk} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ f_{1Hk}+s_{1Hk} & \ldots & \ldots & f_{HHk}+s_{HHk} & s_{1Hk} & \ldots & \ldots & s_{HHk} \\ s_{11k} & s_{12k} & \ldots & s_{1Hk} & f_{11k}+s_{11k} & f_{12k}+s_{12k} & \ldots & f_{1Hk}+s_{1Hk} \\ s_{12k} & s_{22k} & \ldots & s_{2Hk} & f_{12k}+s_{12k} & f_{22k}+s_{22k} & \ldots & f_{2Hk}+s_{2Hk} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ s_{1Hk} & \ldots & \ldots & s_{HHk} & f_{1Hk}+s_{1Hk} & \ldots. & \ldots & f_{HHk}+s_{HHk} \end{pmatrix}. \tag{3}$$

The component-specific covariance between the hth outcome of two siblings of the same pair is equal to s_{hhk}, whereas the component-specific variance of the hth outcome is given by $f_{hhk} + s_{hhk}$ ($h = 1, \ldots, H$). In formula (3), the upper-left and lower-right submatrices and the upper-right and lower-left submatrices contain the covariances within a sibling and between the two siblings of a pair belonging to the kth cluster, respectively. Note that the proposed model assumes that the measurements of siblings from different families are independent conditional on cluster membership.

As mentioned in the introduction, our model extends the models used for clustering of gene expression data [12, 13]. The mixture of mixed effect models proposed by Celeux *et al*. [12] is given by Equations (1) and (2), with restrictions $\mathbf{S}_k = s_k\mathbf{I}$ and $\mathbf{F}_k = f_k\mathbf{I}$, that is, unequal, isotropic, error and random component-specific matrices which lead to $\boldsymbol{\Sigma}_k = s_k\mathbf{X}\mathbf{X}' + f_k\mathbf{I}$. The model of Ng *et al*. [13] is slightly more general. Their model allows the variances within a sibling to vary across the various variables, that is, $\mathbf{F}_k = (f_{1k}, \ldots, f_{Hk})'\mathbf{I}$. Both models are not suited for sibling pair data because they assume that the covariances within a sibling are zero and that the covariances between the two siblings in a pair do not vary across outcome variables.

2.1. *Maximum likelihood estimation*

The maximum likelihood estimates of model parameters can be derived for instance through the EM algorithm [15–17], where missing data are of two types: the indicator variables, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$, for $i = 1, \ldots, n$ coming from multinomial distributions with priors π_k, and the random effects $\mathbf{u}_{ik}$ for each family in the kth cluster. Thus, the EM algorithm consists of maximizing iteratively the expectation of the following complete log-likelihood conditional on the observed data and the current value of parameter estimates:

$$
\begin{aligned}
\log L_C(\boldsymbol{\phi}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left[f(\mathbf{y}_i, \mathbf{u}_{ik}|\boldsymbol{\theta}_k)\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left[N(\boldsymbol{\beta}_k + \mathbf{X}\mathbf{u}_{ik}, \tilde{\mathbf{F}}_k)\right] + \sum_{k=1}^{K} z_{ik}\log\left[N(\mathbf{0}, \mathbf{S}_k)\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}[-\frac{1}{2}(2H + H)\log(2\pi) \\
&\quad - \frac{1}{2}\log(|\tilde{\mathbf{F}}_k|) - \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\beta}_k - \mathbf{X}\mathbf{u}_{ik})'\tilde{\mathbf{F}}_k^{-1}(\mathbf{y}_i - \boldsymbol{\beta}_k - \mathbf{X}\mathbf{u}_{ik}) \\
&\quad - \frac{1}{2}\log(|\mathbf{S}_k|) - \frac{1}{2}\mathbf{u}_{ik}'\mathbf{S}_k^{-1}\mathbf{u}_{ik}] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}[-\frac{1}{2}(2H + H)\log(2\pi) \\
&\quad - \frac{1}{2}\log(|\mathbf{F}_k|) - \frac{1}{2}(\mathbf{y}_{i1} - \boldsymbol{\beta}_k^1 - \mathbf{u}_{ik})'\mathbf{F}_k^{-1}(\mathbf{y}_{i1} - \boldsymbol{\beta}_k^1 - \mathbf{u}_{ik}) \\
&\quad - \frac{1}{2}\log(|\mathbf{F}_k|) - \frac{1}{2}(\mathbf{y}_{i2} - \boldsymbol{\beta}_k^2 - \mathbf{u}_{ik})'\mathbf{F}_k^{-1}(\mathbf{y}_{i2} - \boldsymbol{\beta}_k^2 - \mathbf{u}_{ik}) \\
&\quad - \frac{1}{2}\log(|\mathbf{S}_k|) - \frac{1}{2}\mathbf{u}_{ik}'\mathbf{S}_k^{-1}\mathbf{u}_{ik}]
\end{aligned}
$$

where $\boldsymbol{\phi}$ represents the unknown parameter vector and $\boldsymbol{\theta}_k$ contains the unknown elements of $\boldsymbol{\beta}_k$, $\mathbf{S}_k$, and $\tilde{\mathbf{F}}_k$ ($k = 1, .., K$).

In the mth iteration of the E-step, we compute the conditional expectation of the log-likelihood function for complete data. As usual, $\log L_C(\boldsymbol{\phi})$ is linear in the component labels and taking the expectation implies that the component labels z_{ik} are replaced by their expected values, $w_{ik}^{(m)} = Pr(z_{ik} = 1|\mathbf{y}; \boldsymbol{\phi}^{(m)})$, which represent the posterior probabilities that the ith sibling pair belongs to the kth cluster, conditional on the observed data and the current parameter estimates. Therefore, the E-step is reduced to the computation of

$$w_{ik}^{(m)} = \frac{\hat{\pi}_k^{(m)} N(\hat{\boldsymbol{\beta}}_k^{(m)}, \mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(m)} N(\hat{\boldsymbol{\beta}}_k^{(m)}, \mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})}.$$

In the $(m+1)$th iteration M-step, model parameters are estimated by maximizing the expected log-likelihood for complete data with respect to π_k, $\boldsymbol{\beta}_k$ $\mathbf{S}_k$, and $\tilde{\mathbf{F}}_k$, respectively. Using the conditional expectation of the sufficient statistics $\mathbf{u}'_{ik}\mathbf{u}_{ik}$ and $(\mathbf{y}_{ik} - \mathbf{X}\mathbf{u}_{ik})$, we obtain the following estimates for $k = 1, \ldots, K$:

$$\hat{\pi}_k^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(m)}}{n}$$

$$\hat{\mathbf{S}}_k^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(m)} \hat{\mathbf{S}}_k^{(m)}\mathbf{X}'(\mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})^{-1}(\mathbf{y}_{ik} - \hat{\boldsymbol{\beta}}_k^{(m)})(\mathbf{y}_{ik} - \hat{\boldsymbol{\beta}}_k^{(m)})'(\mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})^{-1}\mathbf{X}\hat{\mathbf{S}}_k^{(m)}}{\sum_{i=1}^{n} w_{ik}^{(m)}}$$
$$+ \left[\mathbf{I_H} - \hat{\mathbf{S}}_k^{(m)}\mathbf{X}'(\mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})^{-1}\mathbf{X}\right]\hat{\mathbf{S}}_k^{(m)}$$

$$\hat{\mathbf{F}}_k^{(m+1)} = \frac{\frac{1}{2}\sum_{i=1}^{n} w_{ik}^{(m)}\left[(\mathbf{y}_{i1} - \hat{\boldsymbol{\beta}}_k^{1(m)})(\mathbf{y}_{i1} - \hat{\boldsymbol{\beta}}_k^{1(m)})' + (\mathbf{y}_{i2} - \hat{\boldsymbol{\beta}}_k^{2(m)})(\mathbf{y}_{i2} - \hat{\boldsymbol{\beta}}_k^{2(m)})'\right]}{\sum_{i=1}^{n} w_{ik}}$$
$$- \frac{\sum_{i=1}^{n} w_{ik}\left[\hat{\mathbf{S}}_k^{(m)}\mathbf{X}'(\mathbf{X}\hat{\mathbf{S}}^{(m)}{}_k\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}_k^{(m)})\right]\left[(\mathbf{y}_{i1} - \hat{\boldsymbol{\beta}}_k^{1(m)})' + (\mathbf{y}_{i2} - \hat{\boldsymbol{\beta}}_k^{2(m)})'\right]}{\sum_{i=1}^{n} w_{ik}}$$
$$+ \frac{\sum_{i=1}^{n} w_{ik}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}'(\mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\beta}}_k^{(m)})(\mathbf{y}_i - \hat{\boldsymbol{\beta}}_k^{(m)})'(\mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})^{-1}\mathbf{X}\hat{\mathbf{S}}_k^{(m)}}{\sum_{i=1}^{n} w_{ik}}$$
$$+ \hat{\mathbf{S}}_k^{(m)} - \hat{\mathbf{S}}_k^{(m)}\mathbf{X}'(\mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})^{-1}\mathbf{X}\hat{\mathbf{S}}_k^{(m)}$$

$$\hat{\boldsymbol{\beta}}_k^{j(m+1)} = \frac{\sum_{i=1}^{n} w_{ik}[(\hat{\mathbf{S}}_k^{(m)} + \hat{\mathbf{F}}_k^{(m)})^{-}(\hat{\mathbf{S}}_k^{(m)} + \hat{\mathbf{F}}_k^{(m)})^{-1}(\mathbf{y}_{ijk} - \hat{\boldsymbol{\beta}}_k^{j(m)})]}{\sum_{i=1}^{n} w_{ik}^{(m)}} + \hat{\boldsymbol{\beta}}_k^{j(m)}$$

Furthermore, a flexible estimator of $\mathbf{u}_{ik}$ may be based on the posterior mean of $\mathbf{u}_{ik}$ conditional on the observed data:

$$\hat{\mathbf{u}}_{ik}^{(m+1)} = w_{ik}^{(m)}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}'(\mathbf{X}\hat{\mathbf{S}}_k^{(m)}\mathbf{X}' + \hat{\tilde{\mathbf{F}}}_k^{(m)})^{-1}(\mathbf{y}_{ik} - \hat{\boldsymbol{\beta}}_k^{(m)}). \tag{4}$$

When running the EM algorithm, the Woodbury identity matrix [18] can be used for the inversion given in Equation (4), that is, the component-specific covariance matrix can be written as follows:

$$(\mathbf{X}\mathbf{S}_k\mathbf{X}' + \tilde{\mathbf{F}}_k)^{-1} = \tilde{\mathbf{F}}_k^{-1} - \tilde{\mathbf{F}}_k^{-1}\mathbf{X}(\mathbf{S}_k^{-1} + \mathbf{X}'\tilde{\mathbf{F}}_k^{-1}\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{F}}_k.$$

Especially for high dimensional data, this formula saves computation time because the left-hand side involves an inversion of a $2H \times 2H$ matrix, whereas for the right-hand side of the formula, only $H \times H$ matrices have to be inverted:

$$\tilde{\mathbf{F}}_k^{-1} = \begin{pmatrix} \mathbf{F}_k & \mathbf{0}_H \\ \mathbf{0}_H & \mathbf{F}_k \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{F}_k^{-1} & \mathbf{0}_H \\ \mathbf{0}_H & \mathbf{F}_k^{-1} \end{pmatrix}.$$

The EM algorithm to obtain the maximum likelihood parameter estimates given the number of clusters K is implemented in the LATENT GOLD software 4.5. Typically, the number of clusters is unknown and has to be chosen using approaches available in the standard finite mixture model literature (see McLachlan and Peel [17]). For example, the number of clusters may be chosen by analyzing the number of modes (mainly based on intuition), by applying likelihood-based approaches, and by using Bayesian and penalized likelihood methods (AIC (Akaike information criterion) [19], BIC (Bayesian information criterion) [20], AWE (Approximate weight of evidence) [21] and so on). In this paper, we consider a specific problem where the number of clusters is known.

2.2. Specification of the linear mixed effect model for discovering clusters of sibling pairs on the basis of two health outcomes

As mentioned in the introduction, our interest is to discover three categories of sibling pairs: CH, CUH, and DH sibling pairs. We consider in detail the case of two health outcomes. Without loss of generality, it is assumed that healthy subjects have low values for the two measured health outcomes, y_{ij1} and y_{ij2}. Here, i is the family index, and j denotes the sibling within the ith family. Siblings with at least one high value for one of the outcome variables are also considered to be unhealthy. In order to capture the structural information of these data, we fix the number of cluster K equal to four. In particular, we assume that the first cluster represents CH sibling pairs, the second and third clusters DH sibling pairs (first sibling is healthy and second one is unhealthy and vice versa), and the fourth cluster CUH sibling pairs. On the basis of these assumptions, the component-specific mean vectors in (2) are set as follows: $\boldsymbol{\beta}^{1'}_1 = \boldsymbol{\beta}^{2'}_1 = \boldsymbol{\beta}^{2'}_2 = \boldsymbol{\beta}^{1'}_3 < \boldsymbol{\beta}^{1'}_2 = \boldsymbol{\beta}^{2'}_3 = \boldsymbol{\beta}^{1'}_4 = \boldsymbol{\beta}^{2'}_4$, that is, the means of y_{ij1}'s and y_{ij2}'s for healthy siblings are equal and lower than the means of y_{ij1}'s and y_{ij2}'s for unhealthy siblings, which are equal as well. Moreover, we have:

- $\mathbf{u}_{ik} = (u_{ik1}, u_{ik2})'$ with $\mathbf{u}_{ik} \sim \mathrm{N}(\mathbf{0}, \mathbf{S}_k)$ and $\mathbf{S}_k = \begin{pmatrix} s_{11k} & s_{12k} \\ s_{12k} & s_{22k} \end{pmatrix}$,
- $\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$,
- $\mathbf{e}_{ik}$ with $\mathbf{e}_{ik} \sim \mathrm{N}(\mathbf{0}, \tilde{\mathbf{F}}_k)$ and $\tilde{\mathbf{F}}_k = \begin{pmatrix} f_{11k} & f_{12k} & 0 & 0 \\ f_{12k} & f_{22k} & 0 & 0 \\ 0 & 0 & f_{11k} & f_{12k} \\ 0 & 0 & f_{12k} & f_{22k} \end{pmatrix}$,

with $\tilde{\mathbf{F}}_2 = \tilde{\mathbf{F}}_3$, $\pi_2 = \pi_3$ and $\mathbf{S}_2 = \mathbf{S}_3$. The last three constraints come from the exchangeability of classes 2 and 3 because of the exchangeability of the two siblings. Details to fit this model using the LATENT GOLD software 4.5 are given in the Appendix.

A straightforward extension of this model is to consider three instead of two classes at subject level by also including an intermediate class. This model yields six classes at sibling pair level, namely three concordant classes and three discordant classes. To fit this model, K=9 has to be used. The ordering of the classes can be modeled using the following restrictions for the mean parameters: $\boldsymbol{\beta}^{1'}_1 = \boldsymbol{\beta}^{2'}_1 = \boldsymbol{\beta}^{1'}_4 = \boldsymbol{\beta}^{2'}_7 = \boldsymbol{\beta}^{1'}_5 = \boldsymbol{\beta}^{2'}_8 < \boldsymbol{\beta}^{1'}_2 = \boldsymbol{\beta}^{2'}_2 = \boldsymbol{\beta}^{2'}_4 = \boldsymbol{\beta}^{1'}_7 = \boldsymbol{\beta}^{1'}_9 = \boldsymbol{\beta}^{1'}_6 < \boldsymbol{\beta}^{1'}_3 = \boldsymbol{\beta}^{2'}_3 = \boldsymbol{\beta}^{2'}_5 = \boldsymbol{\beta}^{2'}_6 = \boldsymbol{\beta}^{1'}_8 = \boldsymbol{\beta}^{2'}_9$.

3. Simulations

To study the performance of the model, we carried out a simulation study. We simulated 100 data sets of $n = 500$ sibling pairs. For each sibling, we simulated two health parameters ($H = 2$). The number of clusters K is equal to 4. Table I gives the model parameters of the simulation. These parameters are close to the parameter estimates obtained from our real data (shown in the next section). The largest cluster includes CH sibling pairs, two clusters represent DH, and the last cluster represents the CUH sibling pairs. Note that we use smaller correlations for outcomes of siblings from DH pairs compared with correlations for outcomes of siblings from concordant pairs, and the correlations between outcomes of siblings from CH are assumed to be larger than the corresponding correlations from CUH. The model parameters were estimated by using LATENT GOLD software (computation took about 2 s per data set). We evaluated the performance of the proposed model in terms of:

1. the ability to correctly estimate the model parameters;
2. the degree of agreement between the true and the estimated partition membership by using three agreement indices: Modified Rand Index, Jaccard Index, and Hubert Index [22]. In case of perfect agreement between the true partition and the estimated one, the values of these three indices are equal to 1;
3. detecting the correct number of clusters.

Table I. Simulation setting and parameter estimates corresponding to our model fitted to one of the replicates.

	Parameter setting	Parameter estimation
Healthy		
β_1	1.00	1.00
β_2	0.60	0.60
Unhealthy		
β_3	3.00	3.00
β_4	2.00	2.00
Concordant healthy		
π_1	0.60	0.5909
$\mathbf{F}_1$	$\begin{pmatrix} 0.05 & 0.02 \\ 0.02 & 0.03 \end{pmatrix}$	$\begin{pmatrix} 0.05 & 0.02 \\ 0.02 & 0.03 \end{pmatrix}$
$\mathbf{S}_1$	$\begin{pmatrix} 0.40 & 0.05 \\ 0.05 & 0.50 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.05 \\ 0.05 & 0.50 \end{pmatrix}$
Discordant		
$\pi_2\ (=\pi_3)$	0.14	0.1382
$\mathbf{F}_2\ (=\mathbf{F}_3)$	$\begin{pmatrix} 0.06 & 0.05 \\ 0.05 & 0.09 \end{pmatrix}$	$\begin{pmatrix} 0.06 & 0.05 \\ 0.05 & 0.09 \end{pmatrix}$
$\mathbf{S}_2\ (=\mathbf{S}_3)$	$\begin{pmatrix} 0.08 & 0.01 \\ 0.01 & 0.10 \end{pmatrix}$	$\begin{pmatrix} 0.08 & 0.01 \\ 0.01 & 0.10 \end{pmatrix}$
Concordant unhealthy		
π_4	0.13	0.1327
$\mathbf{F}_4$	$\begin{pmatrix} 0.05 & 0.04 \\ 0.04 & 0.06 \end{pmatrix}$	$\begin{pmatrix} 0.05 & 0.04 \\ 0.04 & 0.06 \end{pmatrix}$
$\mathbf{S}_4$	$\begin{pmatrix} 0.20 & 0.07 \\ 0.07 & 0.20 \end{pmatrix}$	$\begin{pmatrix} 0.20 & 0.07 \\ 0.07 & 0.20 \end{pmatrix}$

The third column of Table I shows the estimated model parameters of a data set (to avoid local maxima, we use various starting points). These estimates agree with the model parameters used for the simulation. In Table II, the degree of agreement indices are given. The given values are averages over the 100 replicates. Because all indices are above 0.95, our model performs well in recovering the true partitions.

To study the performance of our model in terms of identifying the correct number of clusters, we also fitted models with three classes at subject level namely healthy, unhealthy, and an intermediate class. This model corresponds to nine classes for the sibling pairs: the first, the second, and the third clusters represent CH, CM, and CUH sibling pairs, and the fourth and the seventh, the fifth and the eighth, and the sixth and the ninth clusters represent DMH, DHUH, and DMUH, respectively. On the basis of these assumptions, the component-specific mean vectors in (2) were set as follows: $\boldsymbol{\beta}^{1'}_1 = \boldsymbol{\beta}^{2'}_1 = \boldsymbol{\beta}^{1'}_4 = \boldsymbol{\beta}^{2'}_7 = \boldsymbol{\beta}^{1'}_5 = \boldsymbol{\beta}^{2'}_8 < \boldsymbol{\beta}^{1'}_2 = \boldsymbol{\beta}^{2'}_2 = \boldsymbol{\beta}^{2'}_4 = \boldsymbol{\beta}^{1'}_7 = \boldsymbol{\beta}^{1'}_9 = \boldsymbol{\beta}^{1'}_6 < \boldsymbol{\beta}^{1'}_3 = \boldsymbol{\beta}^{2'}_3 = \boldsymbol{\beta}^{2'}_5 = \boldsymbol{\beta}^{2'}_6 = \boldsymbol{\beta}^{1'}_8 = \boldsymbol{\beta}^{2'}_9$, that is, the means of y_{ij1}'s and y_{ij2}'s for healthy siblings are equal and lower than the means of y_{ij1}'s and y_{ij2}'s for moderate siblings, which are equal and lower than the means of y_{ij1}'s and y_{ij2}'s for unhealthy siblings which are equal as well. Our new model with nine clusters gave a higher BIC value (10,811) than the correct model with four clusters (3081). Hence, a model with four clusters should be preferred. The cluster proportions were close to the truth, namely $\pi_1 = 0.60$, $\pi_3 = \pi_5 = \pi_8 = 0.13$, $\pi_2 = \pi_4 = \pi_6 = \pi_7 = \pi_9 \approx 0.00$. Thus, the underlying data structure of four sibling pair classes was recovered.

To illustrate the improvement in performance of the proposed model relative to standard model-based clustering techniques, the Mixture Modelling Software for Matlab (MIXMOD, [23]) was used to analyze the replicates. Here, the data from each pair were treated as a single vector. The variables were allowed

Table II. Measures of degree of agreement: mean value over 100 replicates.

Model	Modified Rand index	Jaccard Index	Hubert Index
The proposed model	0.98	0.97	0.98
Standard model-based clustering	0.95	0.94	0.95

to be correlated, but the specific structure of the sibling pair data represented by formula (3) is ignored, that is, the covariance structure uses more parameters. We first fitted all 28 MIXMOD models with four clusters. In 54 of the 100 simulations, the best model was Gaussian_pk_Lk_D_Ak_D, whereas for the remaining 46 replicates, the best model was Gaussian_pk_Lk_Ck. Table III shows for each model the BIC value of the replicate with the best model fit. Our proposed model gives a smaller BIC value (3081) than the best-fitting standard mixed model (BIC value of 11,220). Table IV presents the model parameter estimates corresponding to the replicate and the best fitting model (BIC $= 11,220$). Our model appears to outperform the standard models in terms of ability to correctly estimate model parameters. Moreover, the three agreement indices averaged over the 100 simulations are lower than the indices obtained using our model (second line Table III). Note also that our model is more parsimonious with respect to the standard model-based clustering. For example, in this simulation study the number of parameters used in the covariance matrices is reduced from $2H(2H+1)/2+2H(K-1)$ to $(K-1)H(H+1)$.

Finally, the standard 28 MIXMOD models were fitted using nine clusters instead of correct four clusters. The model which fitted best to the data had a BIC value of 11,462, which is similar to the BIC value for models with four clusters (11,220). The obtained prior probabilities for class memberships are $\pi_1 = 0.14$, $\pi_2 = 0.16$, $\pi_3 = 0.13$, $\pi_4 = 0.08$, $\pi_5 = 0.10$, $\pi_6 = 0.12$, $\pi_7 = 0.20$, $\pi_8 = 0.02$, and $\pi_9 = 0.05$, which do not agree with the true four classes. Thus, for our simulations, standard models are not able to pinpoint the correct number of clusters underlying the outcomes.

4. Real data

As application, we are interested in clustering using six health outcomes measured in 427 families from the Leiden Longevity Study. The aim of the analysis was to classify sibling pairs in CH, CUH, and DH pairs using these health outcomes. Families were recruited between 2002 and 2006. Schoenmaker

Table III. BIC values for each MIXMOD model with $K = 4$ obtained in replicates with the best fit for that model.

Model	BIC
Gaussian *pLI*	17,341.129476
Gaussian *pLkI*	16,701.368851
Gaussian *pLB*	16,336.983629
Gaussian *pLkB*	16,624.959196
Gaussian *pLBk*	17,804.373944
Gaussian *pLkBk*	16,498.108325
Gaussian *pLC*	17,462.176722
Gaussian *pLkC*	16,410.484218
Gaussian *pLDAkD*	17,662.605223
Gaussian *pLkDAkD*	15,903.007364
Gaussian *pLDkADk*	17,285.972361
Gaussian *pLkDkADk*	16,194.564931
Gaussian *pLCk*	13,328.303004
Gaussian *pLkCk*	12,581.307822
Gaussian *pkLI*	12,586.753144
Gaussian *pkLkI*	12,710.137657
Gaussian *pkLB*	12,555.373083
Gaussian *pkLkB*	12,359.955583
Gaussian *pkLBk*	12,085.169249
Gaussian *pkLkBk*	12,244.604392
Gaussian *pkLC*	13,437.305213
Gaussian *pkLkC*	12,492.306760
Gaussian *pkLDAkD*	12,242.787730
Gaussian *pkLkDAkD*	11,220.407120
Gaussian *pkLDkADk*	11,622.510648
Gaussian *pkLkDkADk*	11,553.403224
Gaussian *pkLCk*	11,341.642045
Gaussian *pkLkCk*	11,274.373955

Table IV. Parameter estimates for model Gaussian $pkLkDAkD$ for the replicate showing the best fit.

	CH	CUH	DH1	DH2
$\boldsymbol{\mu}_k$	[0.58,0.96,0.60,0.99]’	[1.94,2.93,1.95,3.01]’	[0.63, 0.98, 2.05, 2.97]’	[1.99, 2.95, 0.60, 1.08]’
$\boldsymbol{\Sigma}_k$	$\begin{pmatrix} 0.52 & 0.12 & 0.45 & 0.01 \\ - & 0.62 & 0.01 & 0.61 \\ - & - & 0.51 & 0.13 \\ - & - & - & 0.67 \end{pmatrix}$	$\begin{pmatrix} 0.24 & 0.04 & 0.19 & 0.00 \\ - & 0.27 & 0.00 & 0.21 \\ - & - & 0.25 & 0.05 \\ - & - & - & 0.30 \end{pmatrix}$	$\begin{pmatrix} 0.12 & 0.47 & 0.06 & 0.01 \\ - & 0.17 & 0.00 & 0.08 \\ - & - & 0.14 & 0.06 \\ - & - & - & 0.16 \end{pmatrix}$	$\begin{pmatrix} 0.13 & 0.67 & 0.07 & 0.02 \\ - & 0.21 & 0.01 & 0.10 \\ - & - & 0.15 & 0.07 \\ - & - & - & 0.21 \end{pmatrix}$
π_k	0.56	0.14	0.17	0.13

et al. [24] describe the design in detail. Briefly, families participating in the Leiden Longevity Study have at least two siblings meeting four inclusion criteria: (1) men are aged 89 years or above, and women are aged 91 years or above; (2) subjects have at least one living brother or one living sister who fulfills the first criterion and is willing to participate; (3) the nonagenarian sibship has an identical mother and father; (4) the parents of the nonagenarian sibship are Dutch and Caucasian. Note that for selection, different age cutoff points for men and women were used because of differences in life expectancies. There were no selection criteria on health or demographic characteristics. Blood samples were taken at baseline. The six health variables used in the study were mini mental state examination (MMSE), thyroxine (T4), triiodothyronine (T3), glucose levels, low-density lipoprotein (LZ), and high-density lipoprotein (HZ) particle sizes [25–27]. To obtain approximately normally distributed data, we transformed MMSE and glucose. For MMSE, we used square root of 30 minus MMSE (tMMSE) as proposed by [28]. For glucose, we used a log transformation (Lglucose). Table V shows the descriptives of the variables. For the most (transformed) variables, low values mean healthy, unless for LZ and HZ (where it is vice versa). To have an idea about the ability of each health variable to discriminate between two classes (healthy and unhealthy), we fitted a standard k-means algorithm [29] with two clusters, ignoring the family membership. Table V also shows the obtained cluster means, and the last column presents the relative differences in the means of the two clusters. In detail, tMMSE seems to be promising in discrimination with a relative difference of 0.79. In addition, Lglucose, T3, and T4 show moderate differences in mean between the two clusters, namely a relative difference of 0.25, 0.25, and 0.26, respectively. However, fitting the proposed model by using all of the four health outcomes (namely, tMMSE, Lglucose, T3, and T4), where three are not well discriminating, seems to obscure the impact of tMMSE; in fact, the obtained clusters were close to each other (results not presented here). Therefore, we decided to focus mainly on the effect of tMMSE together with just one of the other outcomes which have the same discriminant power. We then applied our proposed model to classify sibling pairs by using two health parameters, namely, tMMSE and either Lglucose, T3, or T4. Because of missing data, we used data on 603 individuals forming 354 sibling pairs. When the sibship size was larger than two, siblings were in multiple pairs. For Lglucose in combination with tMMSE, we found various local maxima, each of which gives a solution in which the difference in Lglucose between healthy and unhealthy was rather small. Also, the combination of tMMSE and T3 did not result in well interpretable solutions. Here, the problem seems to be that T3 decreases with age. For the combination of tMMSE and T4, our model gave the best results. Table VI shows the descriptions of the three clusters. We obtained the following parameter estimates: a mean of 1.94 for tMMSE and 15.53 for T4 in healthy individuals and a mean of 2.65 for tMMSE and 17.2 for T4 in unhealthy individuals. The proportions of CH, CUH, and DH siblings were 0.63, 0.09, and 0.28. The correlations between tMMSE of a sibling pair within CH, CUH, and DH were 0.20, 0, and 0.06,

Table V. The descriptives of the outcomes and the k-means results of the 427 sibships of the Leiden Longevity Study.

Outcome	Size	Mean	SD	1st cluster mean	2nd cluster mean	r. difference
HZ	867	9.35	0.56	8.90	9.80	0.10
Lglucose	912	0.80	0.11	0.75	0.95	0.25
LZ	867	21.45	0.78	22.00	20.70	0.06
tMMSE	739	2.18	1.00	1.64	3.36	0.79
T3	903	4.01	0.67	4.80	3.80	0.25
T4	903	16.01	2.73	14.50	18.70	0.26

Table VI. Clustering results by using tMMSE and T4 outcomes.

	CH	DH	CUH
$\hat{\pi}_k$	0.63	0.09	0.28
Age mean (mode)	93.18 (91.90)	93.06 (91.20)	94.06 (91.00)
$\hat{\beta}_{tMMSE,k}$	1.935		2.653
$\hat{\beta}_{T4,k}$	15.534		17.214
corr_{tMMSE}	0.20	0	0.06
corr_{T4}	0.30	0.19	0.00

respectively, and between T4 of a sibling pair within CH, CUH, and DH were 0.30, 0.19, and 0. Moreover, the correlations between tMMSE and T4 within an individual of the CUH and DH clusters were negative, whereas the correlation between these health variables was positive in the CH cluster (results are not shown). Finally, we studied the relationship between membership of CH and two variables measuring mortality, namely, a family history score measuring excess survival of the parental generation and follow-up of the nonagenarian siblings (mean 3.3 years, maximum follow-up 7.5 years). We defined CH membership as having a posterior probability to belong to CH of larger than 0.5. It appeared that sibling pairs belonging to CH cluster have parents who lived longer ($p = 0.06$, T-test, one-sided p-value), and siblings who belong to the CH cluster had a significant longer follow-up time ($p = 0.005$, logrank test). These results suggest that CH membership is associated with a longer life span.

5. Conclusions

Model-based clustering is a tool for joint analysis of multiple and diverse variables. It can be used to identify clusters of subjects with similar profiles for these variables. We extended current tools for single subjects to cluster families by adding random family effects in the cluster-specific regression model and allowing class membership of the siblings to be associated. Our proposal can be seen as an extension of mixtures of LMMs proposed by [12] and [13] in the context of microarray data. Estimates of the parameters can be found by using the EM algorithm. The models can be fitted by the software LATENT GOLD. As illustration, we fitted the proposed model to identify healthy sibling pairs of the Leiden Longevity Study on the basis of two health outcomes. Via a simulation study, we showed that LATENT GOLD is able to recover the model parameters. We also showed that the proposed model outperformed the standard model-based clustering approach in terms of degree of agreement, identification of the correct number of clusters, and computational simplicity.

The formulae for the EM algorithm given for sibling pairs can be easily extended to general sibship sizes. In the future, we plan to extend our models by considering more advanced covariance structures. For example, the correlation between sibling pairs can be modeled as a function of sharing of marker alleles identical by descent at a position of the genome (genetic linkage analysis). Another straightforward extension of our interest is to include common risk factors (e.g., demographic and socio-economic characteristic of family) in the model for the prior cluster membership probabilities as in standard mixture models. In addition, because sibling pairs are often followed over time and repeated measures per sibling are available, it makes sense to expand the proposed model to deal with such a situation.

Concerning the data example, we were interested in classification of sibling pairs in CH, DH, and CUH classes. From the six health outcomes, two health outcomes appeared to discriminate between classes. We showed that obtained classification results were realistic. In particular, within the CH class, both siblings have beneficial values for both health outcomes. In this cluster, siblings have parents who reached older ages and also show excess survival themselves. On the other hand, within DH and CUH clusters, sibling pairs may be present who have beneficial values for only one of the two variables or for another unmeasured health variable, such as grip strength.

The number of variables which are measured within epidemiological studies is increasing rapidly. Single-variable analyses does not provide insight in the joint distribution of the variables. For further study about biological mechanisms for healthy aging using expensive techniques such as whole genome sequencing, the most interesting subset based on the joint distribution has to be identified. Model-based clustering is a tool that can provide the best subset. For example, pairs with largest probabilities to be concordant healthy and pairs with the largest probabilities to be discordant can be chosen.

To conclude, model-based clustering is a promising tool to obtain more insight in underlying biological data structures when many outcomes are measured at different scales.

6. Appendix

The model discussed in this article can be defined using LATENT GOLD software 4.5, which is a latent class program introduced by [14] that finds the ML estimates through the EM and Newton–Raphson algorithms (first EM and Newton–Raphson when close to the maximum). More specifically, we need to define a series of regression equations for the latent and the response variables, and the settings for the (residual) variances and covariances. Before defining the regression equations, we specify the technical and output options and the id variables, as well as the names and scale types of the latent, dependent

variables that play a role in the model. The model definition for the specification of the LMM for discovering clusters of sibling pairs on the basis of two health outcomes, described in Section 2.2 and fitted on real data in Section 4, consists of the following syntax language:

```
variables
 dependent y11 continuous, y12 continuous, y21 continuous, y22
 continuous; latent cluster1 2, cluster2 2, u1 continuous, u2
 continuous; equations (s1) u1 | cluster1 cluster2;
 (s2) u2 | cluster1 cluster2;
 (s12) u1 <-> u2 | cluster1 cluster2;
 cluster1 <- (k) 1;
 cluster2 <- (k) 1;
 (l) cluster1 <-> cluster2;
 y11 <- (a) 1 | cluster1 + (1) u1;
 y12 <- (b) 1 | cluster1 + (1) u2;
 y21 <- (a) 1 | cluster2 + (1) u1;
 y22 <- (b) 1 | cluster2 + (1) u2;
 (f1) y11 | cluster1 cluster2;
 (f2) y12 | cluster1 cluster2;
 (f1) y21 | cluster1 cluster2;
 (f2) y22 | cluster1 cluster2;
 (f12) y12 <-> y11 | cluster1 cluster2;
 (f12) y22 <-> y21 | cluster1 cluster2;
 s1[3] = s1[2];
 s2[3] = s2[2];
 s12[3] = s12[2];
 f1[3] = f1[2];
 f2[3] = f2[2];
 f12[3] = f12[2];
```

In detail, Cluster 1 represents two classes for sibling 1, and Cluster 2 represents two classes for sibling 2. The joint model has 2×2 or four classes. The first three lines of equations define the (co)variances in the $\mathbf{S}_k$'s which are allowed to differ across the four classes. The next three lines define the logit models for the class proportions for siblings 1 and 2 and their association; by using the same parameter label 'k' for the intercept '1', we get $\pi_2 = \pi_3$. The next four lines define the regression models for y's. By using the same labels 'a' and 'b', the parameters are equated for the two siblings, yielding the structure of beta described in the text. The next six lines define the residual (co)variances. The last six lines define the restrictions $\mathbf{S}_2 = \mathbf{S}_3$ and $\mathbf{F}_2 = \mathbf{F}_3$.

Adding covariates in the priors would imply having one extra line in variables. Just to give an idea, let us assume that we have three predictors which may have different values for the two siblings, then the extra line will be:

```
independent x11, x12, x13, x21, x22, x23;
```

Moreover, the equations for cluster1 and cluster2 have to be replaced by

```
cluster1 <- (k0) 1 + (k1) x11 + (k2) x12 + (k3) x13;
cluster2 <- (k0) 1 + (k1) x21 + (k2) x22 + (k3) x23;
```

Acknowledgements

This work was supported by grants from Innovation Oriented research Program IOP on Genomics (SenterNovem) and Netherlands Organization for Scientific Research (NWO).

References

1. Houwing-Duistermaat JJ, Callegaro A, Beekman M, Westendorp RG, Slagboom PE, Van Houwelingen JC. Weighted statistics for aggregation and linkage analysis of human longevity in selected families: The Leiden Longevity Study. *Statistics in Medicine* 2009; **28**:140–151.

2. Franceschi C, Bezrukov V, Blanch H, Bolund L, Christensen K, de Benedictis G, Deiana L, Gonos E, Hervonen A, Yang H, Jeune B, Kirkwood TB, Kristensen P, Leon A, Pelicci PG, Peltonen L, Poulain M, Rea IM, Remacle J, Robine JM, Schreiber S, Sikora E, Slagboom PE, Spazzafumo L, Stazi MA, Toussaint O, Vaupel JW. Genetics of healthy aging in Europe: the EU-integrated project GEHA (GEnetics of Healthy Aging). *Annals of the New York Academy of Sciences* 2007; **1100**:21–45.
3. Hallmayer JF, Jablensky A, Michie P, Woodbury M, Salmon B, Combrinck J, Wichmann H, Rock D, D'Ercole M, Howell S, Dragovic M, Kent A. Linkage analysis of candidate regions using a composite neurocognitive phenotype correlated with schizophrenia. *Molecular Psychiatry* 2003; **8**(5):511–523.
4. Fanous AH, Neale MC, Webb BT, Straub RE, O'Neill FA, Walsh D, Riley BP, Kendler KS. Novel linkage to chromosome 20p using latent classes of psychotic illness in 270 Irish high density families. *Biological Psychiatry* 2008; **64**:121–127.
5. Kendler KS, Karkowski LM, Walsh D. The structure of psychosis: latent class analysis of probands from the Roscommon Family Study. *Archives of General Psychiatry* 1998; **55**:492–499.
6. Sullivan PF, Kessler RC, Kendler KS. Latent class analysis of lifetime depressive symptoms in the national comorbidity survey. *American Journal of Psychiatry* 1998; **155**:1398–1406.
7. Neuman RJ, Todd RD, Heath AC, Reich W, Hudziak JJ, Bucholz KK, Madden PA, Begleiter H, Porjesz B, Kuperman S, Hesselbrock V, Reich T. Evaluation of ADHD typology in three contrasting samples: a latent class approach. *Journal of the American Academy of Child and Adolescent Psychiatry* 1999; **38**:25–33.
8. Labbe A, Bureau A, Merette C. Integration of genetic familial dependence structure in latent class model. *International Journal of Biostatistics* 2009; **5**(1), Article 6. DOI: 10.2202/1557-4679.1126.
9. Tayeb A, Labbe A, Bureau A, Mrette A. Solving genetic heterogeneity in extended families by identifying sub-types of complex diseases. *Computational Statistics* 2011. DOI: 10.1007/s00180-010-0224-2.
10. Vermunt JK. Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research* 2008; **17**:33–51.
11. Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 1999; **55**:117–128.
12. Celeux G, Martin O, Lavergne C. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling* 2005; **5**(3):243–267.
13. Ng S, McLachlan GJ, Wang K, Ben-Tovim Jones L, Ng S-W. A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* 2006; **22**(14):1745–1752.
14. Vermunt JK, Magidson J. LATENT GOLD *User's Manual*. Statistical Innovations Inc: Boston, 2000; 185.
15. Dempster AP, Laird NM, Rubin DA. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 1977; **39**(1):1–38.
16. McLachlan GJ, Krishnan T. *The EM algorithm and extensions.* Wiley series in probability and statistics. John Wiley & Sons, 1997.
17. McLachlan GJ, Peel D. *Finite Mixture Models*. Wiley: New York, 2000.
18. Woodbury Max A. *Inverting Modified Matrices*, Statistical Research Group, Mem. Rept. 42. Princeton University: Princeton, NJ, 1950.
19. Akaike H. On entropy maximization principle. In *Applications of Statistics*, Krishnaiah PR (ed.). North-Holland: Amsterdam, 1997; 27–41.
20. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**(2):461–464.
21. Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 1993; **49**:803–821.
22. Milligan GW, Cooper MC. A study of comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research* 1986; **21**:441–458.
23. Biernacki C, Celeux G, Govaert G, Langrognet F. Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics & Data Analysis* 2006; **51**(2):587–600.
24. Schoenmaker M, de Craen AJM, de Meijer PHM, Beekman M, Blauw GJ, Slagboom PE, Westendorp RGJ. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European Journal of Human Genetics* 2006; **14**:79–84. DOI: 10.1038/sj.ejhg.5201508.
25. Heijmans BT, Beekman M, Houwing-Duistermaat JJ, Cobain MR, Powell J, Blauw GJ, van der Ouderaa F, Westendorp RG, Slagboom PE. Lipoprotein particle profiles mark familial and sporadic human longevity. *PLoS Medicine* 2006; **3**:e495.
26. Vaarhorst AA, Beekman M, Suchiman EH, van Heemst D, Houwing-Duistermaat JJ, Westendorp RG, Slagboom PE, Heijmans BT. Lipid metabolism in long-lived families: the Leiden Longevity Study. *Age (Dordr)* 2010; **33**(2):219–227.
27. Rozing MP, Houwing-Duistermaat JJ, Slagboom PE, Beekman M, Frlich M, de Craen AJ, Westendorp RG, van Heemst D. Familial longevity is associated with decreased thyroid function. *Journal of Clinical Endocrinology & Metabolism* 2010; **95**(11):4979–4984.
28. Lima CP, Gadda HJ. Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer Methods and Programs in Biomedicine* 2005; **78**(2):165–173.
29. MacQueen JB. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press: Berkeley, 1967; 281–297.