BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 2 2012, pages 222–228
doi:10.1093/bioinformatics/btr653

Gene expression

Advance Access publication November 26, 2011

Modelling time course gene expression data with finite mixtures of linear additive models

Bettina Grün[1,*], Theresa Scharl[2] and Friedrich Leisch[3]

[1]Department of Applied Statistics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, [2]Austrian Centre of Industrial Biotechnology, Muthgasse 11, 1190 Wien and [3]Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences Vienna, Gregor Mendel Strasse 33, 1180 Wien, Austria

Associate Editor: Janet Kelso

ABSTRACT

Summary: A model class of finite mixtures of linear additive models is presented. The component-specific parameters in the regression models are estimated using regularized likelihood methods. The advantages of the regularization are that (i) the pre-specified maximum degrees of freedom for the splines is less crucial than for unregularized estimation and that (ii) for each component individually a suitable degree of freedom is selected in an automatic way. The performance is evaluated in a simulation study with artificial data as well as on a yeast cell cycle dataset of gene expression levels over time.

Availability: The latest release version of the R package flexmix is available from CRAN (http://cran.r-project.org/).

Contact: bettina.gruen@jku.at

Received on September 8, 2011; revised on November 21, 2011; accepted on November 22, 2011

1 INTRODUCTION

Time-course microarray experiments make it possible to look at the gene expression of thousands of genes at several time points simultaneously. Clustering of gene expression patterns is, in general, used to identify common temporal or spatial expression patterns. Cluster results contribute to the regulatory network of gene expression, i.e. suggest functional pathways and interaction between genes. In the literature, numerous methods for clustering time-course gene expression data have been proposed [see, for example, Androulakis *et al.* (2007)]. Besides traditional methods like hierarchical clustering or the classical k-means algorithm, model-based clustering is frequently used. Model-based clustering has the advantage to provide a framework to determine the number of clusters and the role of each variable in the clustering process (e.g. Maugis *et al.*, 2009).

Suitable models for time-course gene expression data need to be able to (i) distinguish between groups with different expression patterns and to (ii) determine smooth curves for the development over time. Finite mixtures of regression models, e.g. of linear models (LMs) or of linear mixed models, with splines as covariates were proposed for this purpose (Celeux *et al.*, 2005; Luan and Li, 2003; Ng *et al.*, 2006). The use of splines to determine the covariate matrices has the advantage that the functional relationship between the dependent and independent variables does not need to be specified *a priori* and arbitrary smooth functions can be fit. The disadvantage is that the flexibility of the covariate space needs to be fixed before model estimation, i.e. needs to be assumed to be known. If the covariate space is determined in a data-driven way, different mixture models where the flexibility of the spline functions is varied need to be compared using model selection techniques. Even when imposing the restriction that the same degree of flexibility applies to all components, a considerable number of different models needs to be estimated and compared. While Luan and Li (2003), Celeux *et al.* (2005) and Ng *et al.* (2006) discuss model selection with respect to the number of components and recommend the Bayesian information criterion (BIC) for this purpose, they do not address the issue of selecting appropriate degrees of freedom for the splines. In their applications, only models with an *a priori* fixed number of degrees of freedom are fitted and compared.

Linear additive models (LAMs) model the dependent variable as a sum of smooth functions of the covariates (Hastie and Tibshirani, 1990). These smooth functions can either be non-parametric local smoothers or also spline functions. In the case of spline functions, in general, penalized regression splines are used and the degree of flexibility for the smooth functions is determined by choosing an appropriate smoothing parameter. The smoothing parameter can either be selected using generalized cross-validation (GCV) or (restricted) maximum likelihood [(RE)ML] (Wood, 2006). In the latter case, the smoothing parameter is determined using the maximal marginal (restricted) likelihood integrated over the penalized coefficients, which after re-parameterization are assumed to follow a normal distribution with mean zero and variance indirectly proportional to the smoothing parameter. LAMs with regularized estimation therefore allow to estimate the degree of flexibility in a data-driven way. This ensures that a suitable and parsimonious model is selected. The exact maximum flexibility (i.e. the degrees of freedom of the splines) allowed is less important and the different models arising from changing this hyperparameter can be compared on a coarser grid reducing the number of models evaluated in the model selection step.

In Section 2, the model is specified and methods for estimation and inference in a ML framework are discussed. Section 3 evaluates the performance of the mixtures of LAMs with regularized estimation using artificial data, which resembles time-course gene expression patterns. The number of noise genes as well as the variation within components are varied to assess the influence of these data characteristics on the performance. An application to the yeast cell

[*]To whom correspondence should be addressed.

cycle dataset from Spellman *et al.* (1998) is presented in Section 4. The article concludes with a summary and an outlook.

2 MODEL SPECIFICATION AND ESTIMATION

The mixture density h of a finite mixture model with K components is given for gene i by

$$h(Y_i|B_i,\Theta) = \sum_{k=1}^{K} \pi_k \left[\prod_{j=1}^{J_i} f(y_{ij}|\mu_k(u_{ij}),\theta_k) \right].$$

$Y_i = (y_{ij})_{j=1,\dots,J_i}$ is the response for gene i, u_{ij} are the p predictors (optionally also including an intercept) for gene i derived from a set of basis functions for the explanatory variable and its j-th observation. The predictors are summarized to give $U_i = (u_{ij})_{j=1,\dots,J_i}$. Θ denotes the vector of all parameters for the mixture density h. For the component weights π_k it holds that $\pi_k > 0$ for all k and $\sum_{k=1}^{K} \pi_k = 1$. The repeated observations for the same gene are assumed to be independent given the component membership. The component density function f is assumed to belong to the same parametric family for all components and differ only with respect to the mean parameter given by $\mu_k(u_{ij})$ and the dispersion parameter θ_k. In the following, we assume that f is the normal density and that the mean parameter depends on the covariates u_{ij} using spline-based smooth functions. The dispersion parameter θ_k is in the following denoted by σ_k^2.

This model specification covers mixtures of LMs as well as mixtures of LAMs. For both models, the components have the same parametric distributional form and hence the same number of formal parameters. However, the parameters are estimated differently. For LMs, the likelihood is directly maximized. LAMs are fitted using regularized estimation, i.e. the regression coefficients in the components are penalized to arrive at a compromise between model fit and smoothness of the fitted curve.

In the following, an EM-type algorithm is proposed for fitting mixtures of LAMs using regularized estimation. The Estimation–Maximization (EM) algorithm (Dempster *et al.*, 1977) is a general method for finding the ML estimate for general finite mixture models. It provides a common framework for ML estimation in a missing data setting. For finite mixtures, the missing information is the component membership of the genes. By augmenting the data with the missing information, the so-called complete data is derived and the corresponding complete-data log-likelihood is in general easier to maximize. The EM algorithm is an iterative method which (i) determines the expected complete-data log-likelihood given the observed data and the current parameter estimates in the E-step and (ii) maximizes the expected complete-data log-likelihood to derive new parameter estimates in the M-step. For mixtures of LAMs with regularized estimation, the proposed algorithm also consists of an E- and an M-step, but includes an additional step where the smoothing parameter is adapted. Since the smoothing parameter is a hyperparameter, this step is referred to as H-step and the resulting algorithm is called an HEM algorithm. The E-step involves determining the expected penalized complete-data log-likelihood, which requires to determine the *a posteriori* probabilities of the genes for each of the components. The M-step consists of maximizing separately for each component the component-specific penalized likelihoods of the observations weighted with their *a posteriori* probabilities for this component. Between E- and M-step, the currently optimal smoothing parameter for each component is determined using the component-specific likelihoods of the observations weighted with their current *a posteriori* probabilities within the random effects framework for the penalized parameters. Hence, the algorithm is identical to the case of finite mixtures of LMs fitted using unregularized likelihoods except that in the M-step some regression coefficients are estimated after imposing a penalty on them using a smoothing parameter and that an additional step for determining the smoothing parameter is required.

The weighted likelihood in the M-step is regularized by introducing a penalty term K based on the spline basis used to generate U and the roughness penalty desired by the analyst. For example, for B-splines the most common roughness penalty is the integral of the square of the second derivative. If δ are the regression coefficients for U, the penalty term is given by $\lambda\delta^\top K\delta$. For reparameterizing the model to a mixed model representation, the regression coefficients are decomposed into unpenalized and penalized regression coefficients for the covariates $X = (X_i)_{i=1,\dots,n}$ and $Z = (Z_i)_{i=1,\dots,n}$ of the n genes respectively. The requirements between the penalization matrix K and the covariate matrices X and Z for this decomposition are listed in Kneib (2006, chapter 5.1). For finite mixtures of LAMs with regularized estimation, the parameter vector Θ consists of $(\pi_k, \beta_k, b_k, \sigma_k^2)_{k=1,\dots,K}$ where β_k are the coefficients of the fixed effects and b_k the coefficients of the random effects which are used for the penalized covariates. The smoothing parameters are given by the hyperparameters $\Lambda = (\lambda_k)_{k=1,\dots,K}$. The penalized log-likelihood Δ_p is given by

$$\log\Delta_p(\Theta|Y,X,Z,\Lambda) =$$

$$\sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \pi_k \prod_{j=1}^{J_i} f(y_{ij}|x_{ij}^\top\beta_k + z_{ij}^\top b_k, \sigma_k^2)\right] - \frac{1}{2}\sum_{k=1}^{K} \lambda_k b_k^\top b_k$$

with $Y = (Y_i)_{i=1,\dots,n}$. The penalized log-likelihood corresponds to the joint log-likelihood of the observations and the random effects. In a mixture of random effects models, the marginal log-likelihood after integrating out the random effects would be considered.

The penalized complete-data log-likelihood given the data and the component membership assignments c is equal to

$$\log\Delta_{p,c}(\Theta|Y,X,Z,c,\Lambda) =$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\left[\log(\pi_k) + \sum_{j=1}^{J_i} \log f(y_{ij}|x_{ij}^\top\beta_k + z_{ij}^\top b_k, \sigma_k^2)\right]$$

$$-\frac{1}{2}\sum_{k=1}^{K} \lambda_k b_k^\top b_k.$$

c_{ik} equals one if gene i is assigned to component k and zero otherwise. Each gene is assigned exactly to one component. The penalized complete data log-likelihood is linear in the unobserved component membership assignments c and the expected component membership assignments are given by the *a posteriori* probabilities $\hat{c}$.

E-step: $\hat{c}_{ik} = c_{ik}(\hat{\Theta})$ denotes the *a posteriori* probability for gene i to be from component k given the current parameter estimates $\hat{\Theta}$. It is determined by

$$\hat{c}_{ik} \propto \hat{\pi}_k \left[\prod_{j=1}^{J_i} f(y_{ij}|\hat{\mu}_k(x_{ij}, z_{ij}), \hat{\sigma}_k^2)\right],$$

where $\hat{\mu}_k(x_{ij}, z_{ij}) = x_{ij}\hat{\beta}_k + z_{ij}\hat{b}_k$.

H-step: the hyperparameter Λ is determined in this step. For each of the component-specific models, the likelihoods of the observations weighted with the *a posteriori* probabilities are used. The smoothing parameter λ_k is estimated separately for each component by integrating out the regularized coefficients and maximizing the resulting likelihood with respect to λ_k. In the Appendix A, it is shown that this is equivalent to maximizing the likelihood of a multivariate normal distribution with a suitable variance–covariance matrix.

M-step: the component sizes are determined separately from the component distribution-specific parameters. The determination of the component sizes are the same as for the normal EM algorithm. They are determined for each k by

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{c}_{ik}.$$

For the smoothing parameter λ_k estimated in the H-step, the component distribution specific parameters β_k, b_k and σ_k^2 are determined by maximizing separately for each k the regularized likelihood for each component using the weighted observations.

$$\sum_{i=1}^{n} \hat{c}_{ik} \left[\sum_{j=1}^{J_i} \log f(y_{ij}|x_{ij}\beta_k + z_{ij}b_k, \sigma_k^2) \right] - \frac{1}{2}\lambda_k b_k^\top b_k$$

Conditional on the estimates of λ_k and σ_k^2, the predicted values of the random effects b_k and the maximum likelihood estimates of the parameters β_k minimize the penalized sum of squares (Wood, 2006, p. 300) and hence, maximize the component specific density. The predicted values of the random effects b_k are determined by

$$\begin{aligned}\hat{b}_k &= (\hat{Z}_k^\top \hat{Z}_k + \hat{\sigma}_k^2 \hat{\lambda}_k I)^{-1} \hat{Z}_k^\top (\hat{Y}_k - \hat{X}_k \hat{\beta}_k) \\ &= (Z^\top \mathrm{diag}(\hat{c}_k) Z + \hat{\sigma}_k^2 \hat{\lambda}_k I)^{-1} Z^\top \mathrm{diag}(\hat{c}_k)(Y - X\hat{\beta}_k),\end{aligned}$$

where $\hat{Y}_k = (\sqrt{\hat{c}_{ik}} y_{ij})_{ij}$ and $\hat{X}_k, \hat{Z}_k$ and $\hat{c}_k$ are analogously defined.

Proposition 1. *The penalized log-likelihood* $\log \Delta_p(\Theta|Y,X,Z,\Lambda)$ *is increased or identical in each step conditional on* Λ.

Proof. Assume that Θ' is the new and Θ is the old parameter estimate. In the M-step, the following function is maximized with respect to Θ'.

$$\begin{aligned}Q(\Theta'|\Theta,\Lambda) &= \mathbb{E}\left[\log \Delta_{p,c}(\Theta'|Y,X,Z,c,\Lambda)|\Theta\right] \\ &= \log \Delta_p(\Theta'|Y,X,Z,\Lambda) + H(\Theta'|\Theta,\Lambda)\end{aligned}$$

where

$$H(\Theta'|\Theta,\Lambda) = \mathbb{E}\left[\log(c_{ik}(\Theta'))|\Theta\right] = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}(\Theta) \log\left(c_{ik}(\Theta')\right).$$

The Gibbs' inequality implies that

$$H(\Theta'|\Theta,\Lambda) \le H(\Theta|\Theta,\Lambda) \quad \text{for all } \Theta, \Theta'.$$

Because of the maximization in the M-step it holds that

$$Q(\Theta'|\Theta,\Lambda) \ge Q(\Theta|\Theta,\Lambda).$$

Hence it follows

$$\log \Delta_p(\Theta'|Y,X,Z,\Lambda) \ge \log \Delta_p(\Theta|Y,X,Z,\Lambda).$$

The HEM algorithm is deterministic given a specific initialization, e.g. by providing *a posteriori* probabilities for the observations to start with an M-step. After modification of the EM algorithm, the likelihood is not maximized any more, but the penalized likelihood is maximized conditional on the estimates of the smoothing parameters. The likelihood is not increased in each step, because stronger regularization might also induce a decrease. However, no change of the likelihood during the HEM algorithm indicates that a fixed point has been reached. The algorithm is stopped if either a maximum number of iterations has been performed or if the relative change of the log-likelihood is smaller than a pre-specified threshold. The relative change is determined using $|L_q - L_{q-1}|/(|L_q| + 0.1)$ where L_q denotes the log-likelihood at the q-th iteration. In the following, the maximum number of iterations is always set to 5000.

Especially for mixtures where the component distribution is a normal distribution, problems might occur during the EM as well as the HEM algorithm. In this case, the mixture likelihood is unbounded because infinite values emerge if the variance of one of the components is zero. To avoid estimation problems in the M-step due to very small components, components where the size is smaller than a pre-specified proportion are omitted during the HEM algorithm. The *a posteriori* probabilities are then re-calculated for the remaining components. In the following, this threshold is always set to 0.005.

For finite mixture models, multimodality of the likelihood is generally observed and the initialization strategy is crucial to determine a good estimate. Different initialization strategies were proposed for the EM algorithm for finite mixture models. For an overview and a comparison of different methods for mixtures of LMs and linear mixed models, see Scharl *et al.* (2010) in the context of time-course gene expression data analysis. Scharl *et al.* (2010) recommend the strategy of several short runs of the EM algorithm with a liberal convergence criterion followed by a long run of the EM algorithm initialized in the best solution from the short runs, where convergence is determined with a strict criterion. This procedure was originally proposed for finite mixtures of multivariate Gaussian distributions in Biernacki *et al.* (2003).

Model selection needs to address the determination of (i) the number of components and (ii) the maximum number of degrees of freedom for the components. In both cases, the BIC criterion is proposed. The BIC is the general recommendation to determine the number of components for model-based clustering (Fraley and Raftery, 2002) and mixtures of regression models (Celeux *et al.*, 2005; Luan and Li, 2003; Ng *et al.*, 2006). Other model selection criteria such as the Akaike information criterion (AIC) are also suggested. Information criteria have the advantage that they are easily derived from the fitted model because essentially only the log-likelihood needs to be evaluated as well as the effective degrees of freedom determined. In addition, they select a suitable model according to a compromise between model fit and model complexity.

3 SIMULATION STUDY

The performance of finite mixtures of LAMs with regularized estimation is first evaluated on artificial datasets. These datasets are designed to resemble time-course gene expression patterns. The datasets are generated in the same way as in Scharl *et al.* (2010). The number of components is 15 plus an additional noise component of genes, the number of time points is 16 and the component sizes vary between 10 and 100 yielding a total of 630 genes. The difficulty of the problem is varied with respect to three parameters and for each of these parameters three different levels are used. The standard deviation (SD) of genes around the component centers is varied with values 0.1, 0.3 and 0.5, the number of noise genes is varied with values 100, 500 and 1000 and the SD of the noise genes is varied with values 0, 1 and 2. No individual differences between the genes within the same component are specified. For each experimental setting, 50 different datasets were generated. The finite mixtures of LAMs are estimated in a regularized way with a penalized regression spline as smooth for time. Thin plate regression splines are used and the maximum number of degrees of freedom is 15 with an additional intercept.

The models are estimated using (i) initialization in the true classification and (ii) 10 short runs randomly initialized followed by a long run of the HEM algorithm initialized in the best solution of the short runs. The short runs are terminated if the change in the relative log-likelihoods is smaller than 0.01 and the long run if the change is smaller than 10^{-6}. The long run is initialized in the best solution of the short runs with respect to the log-likelihood. The number of components are set to 16 for random initialization.

The performance of the mixtures of LAMs with regularized estimation is evaluated by determining the Rand index corrected for chance (Hubert and Arabie, 1985) when comparing the true classification to the induced classification by the fitted model. The cluster performance of the model fitted with initialization in the true classification is shown in Figure 1. The resulting classification is nearly perfect if the SD of the genes around the component centers as well as the SD of the noise genes is small regardless of the number of noise genes. Increasing the SD of the genes around the component centers while keeping the SD of the noise genes small leads to still

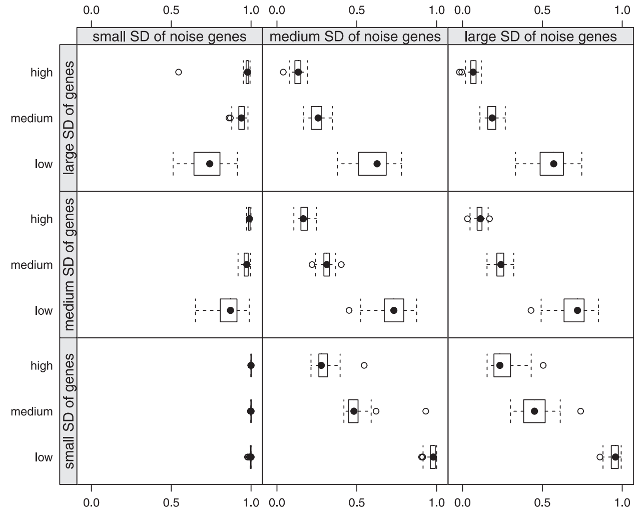

Fig. 1. Adjusted Rand index for initialization in the true classification for finite mixtures of LAMs with regularized estimation with the number of noise genes on the *y*-axis.

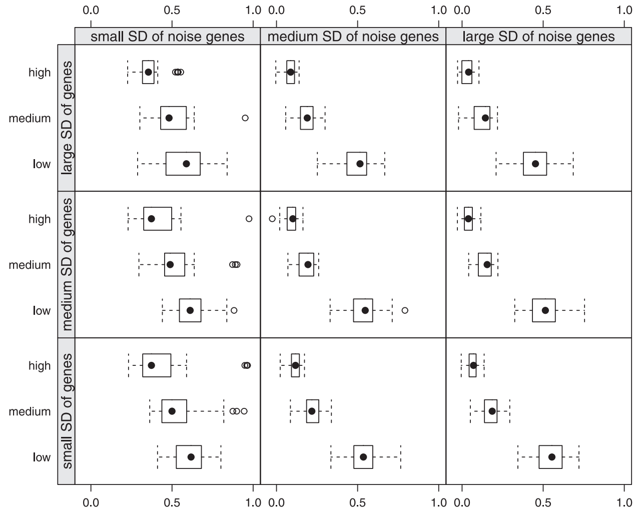

Fig. 2. Adjusted Rand index for the best models detected using short runs followed by a long run for finite mixtures of LAMs with regularized estimation with the number of noise genes on the *y*-axis.

good cluster results, even though the performance is clearly worse if the number of noise genes is only low. This indicates that forming a separate component for the noise genes is more difficult if there are less noise genes. If the SD of noise genes is at least medium (i.e. differs stronger from the SD of the other genes), the performance is best if the number of noise genes is low and the SD of genes is also only small.

The Rand indices adjusted for chance derived by comparing the true classification to the classification induced by the models fitted using the random initialization strategy are given in Figure 2. The cluster performance is clearly worse than for the models fitted with initialization in the true solution. The performance is better for lower number of noise genes regardless of the values of the other parameters. Small SD of noise genes also improves the cluster performance, whereas the SD of the other genes hardly seems to make any difference.

The adjusted Rand index when comparing the true classification to the induced classification by the fitted model is considerably lower for the models fitted with the random initialization strategy than for the models fitted with initialization in the true classification for all datasets. Two explanations are possible: (i) the HEM algorithm was not able to detect the global optimum and (ii) the optimal solution according to the likelihood criterion differs from the true underlying model. The log-likelihood values of the models fitted using initialization in the true classification and using the random initialization strategy are compared to investigate the reason. The comparison indicates that if the SD of the genes around the component centers is only small, the initialization in the true classification leads to considerably better results. For medium and large SD of the genes, the results are reversed: the random initialization strategy leads, in general, to higher likelihood values with a stronger difference for medium than for large SD of the genes around component centers.

During the HEM algorithm, components that are <0.005 are dropped. For initialization in the true solution, the number of components retained are 16 components in 68% of the cases and 15 components (one dropped during the EM algorithm) in 15% of the cases for the fitted models. However, for the random initialization strategy the median number of components converged to is 12. For the different parameter settings, only small differences in the number of components retained are observed.

The results of the simulation study indicate that finite mixtures of LAMs with regularized estimation give very good results in the optimal situation where the true classification is known and this performance deteriorates slightly for more difficult classification problems. Even in the situation where the classification is not known, the random initialization strategy gives reasonable results.

4 APPLICATION TO YEAST CELL CYCLE DATA

In the following, the performance of mixtures of LMs with unregularized estimation and of mixtures of LAMs with regularized estimation is compared using the yeast cell cycle dataset from Spellman *et al.* (1998). Spellman *et al.* (1998) measured the genome-wide mRNA levels for 6178 yeast ORFs simultaneously over approximately two cell cycle periods. The yeast cells were sampled at 7 min intervals for 119 min leading to a total of 18 time points after synchronization. Among these genes, 800 were classified as cell cycle-regulated genes by Spellman *et al.* (1998). In the following, we use the subset of genes that were classified as cell cycle-regulated genes and where the alpha factor arrest is available for all 18 time points. This leads to a final dataset of 613 genes. Almost the same dataset was used by Luan and Li (2003) to fit finite mixtures of mixed-effects models with B-splines. The time-course gene expression data split according to the grouping into five different cell cycle phases (M/G_1, G_1, S, S/G_1 and G_2/M) proposed by Spellman *et al.* (1998) is given in the left panel in Figure 3. The levels over time for each gene are joint and are given by the black lines. The mean values for each time point are determined in each group and indicated by the thicker light gray lines.

The following models are fitted and compared: (i) finite mixtures of LMs are estimated unregularized with an intercept and cubic

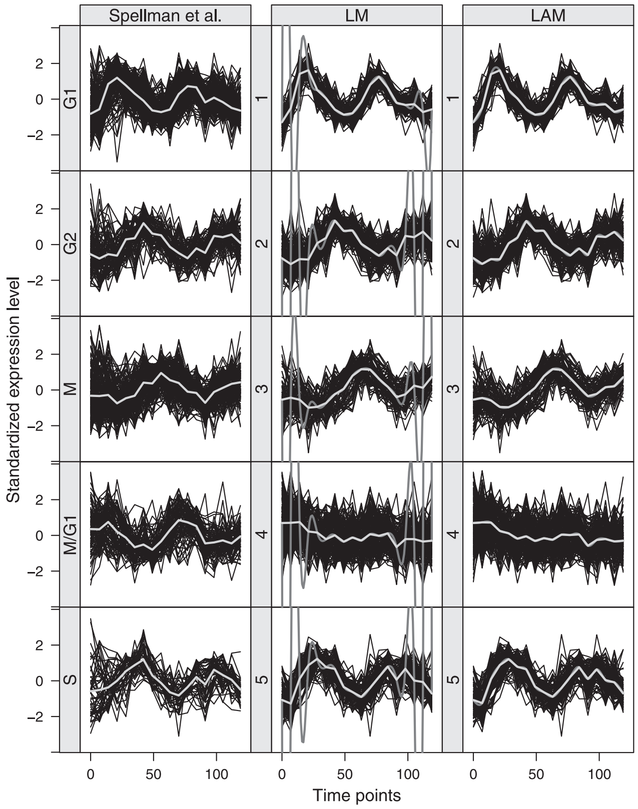

Fig. 3. Yeast cell cycle data using the Spellman *et al.* classification and the classification from fitted finite mixtures of LMs with unregularized estimation and LAMs with regularized estimation initialized in the Spellman *et al.* classification.

B-splines for the covariate time. The degrees of freedom are varied from 5 to 17. (ii) Finite mixtures of LAMs are estimated with a penalized regression spline as smooth for time. Thin plate regression splines are used in addition to an intercept and the maximum number of degrees of freedom is varied from 5 to 17 (by counting up in increments of 4). The decision to use a less fine grid for the LAMs with regularized estimation is based on the assumption that the degrees of freedom specified are less important in this case, because they only constitute an upper bound. As long as the number of degrees of freedom is sufficiently large, similar models can be expected to be chosen (Wood, 2006, p. 161).

In this comparison, finite mixtures of LAMs with regularized estimation have the advantage that the number of degrees of freedom for the splines are selected separately for each component. Hence, different degrees of freedom are possible for the components, while for finite mixtures of LMs with unregularized estimation the selected number of degrees of freedom is the same for all components. This drawback for the mixtures of LMs with unregularized estimation is deliberately chosen, because fitting all models with different number of components and varying number of degrees of freedom for all components is simply computationally infeasible. In addition, the initialization would be still more crucial if the *a priori* symmetry of components is destroyed.

With respect to selecting the number of components, two different estimation strategies are employed. First, the classification into five different groups provided by Spellman *et al.* (1998) is used for the initialization of the (H)EM algorithm. Second, the number of components are varied to take all integer values from 4 to 20 and random initialization with 10 short runs followed by one long run for the best solution of the (H)EM algorithm is used. The best model is selected using the BIC. In addition, the BIC is also used to choose the suitable number of degrees of freedom for the B-splines for the mixtures of LMs and the maximum number of degrees of freedom for the thin plate regression splines for the mixtures of LAMs.

First, the performance is evaluated using the Spellman *et al.* (1998) classification into five groups for initialization of the (H)EM algorithm. In this setting, the number of components is fixed and only the degrees of freedom are selected using the BIC. For the finite mixtures of LMs with unregularized estimation using B-splines, the selected degrees of freedom including the intercept are 18, which is the maximum possible degrees of freedom for 18 time points. For the mixture of LAMs with regularized estimation, the number of degrees of freedom including the intercept for the five components are as follows: 17.5, 17.5, 14.2, 16.3, 17.4. The BIC values are 26 315 for the LMs with unregularized estimation and 26 263 for the LAMs with regularized estimation. This indicates that with respect to the BIC, the finite mixture model of LAMs with regularized estimation is preferred. Given that the fitted mixture of LAMs has about the same amount of flexibility selected for each of the components the similarity in results is not surprising.

Figure 3 compares the three partitions into five groups determined using (i) Spellman *et al.* (1998), (ii) mixtures of LMs with unregularized estimation and (iii) mixtures of LAMs with regularized estimation. The cluster means are inserted for the Spellman *et al.* (1998) partition as well as the estimated smooth curves for the mixture components. The smooth curves are evaluated at (i) the observed time points which are at intervals of 7 min (light gray line) and (ii) at time points at intervals of 1 min (dark gray line). The partitions determined using the finite mixture approach are about the same for LMs with unregularized estimation as well as LAMs with regularized estimation. In fact, only two genes (0.3%) were differently assigned. The estimated smooth curves evaluated only at the observed time points are equally similar. However, evaluation of the estimated curves on a finer grid shows the overfitting of the mixtures of LMs with unregularized estimation.

If the number of components as well as the number of degrees of freedom are selected using the BIC, the best model for the LMs with unregularized estimation has 13 components and for LAMs with regularized estimation 16 components. The selected number of degrees of freedom including the intercept for the LMs with unregularized estimation are 11 and the selected maximum number of degrees of freedom including the intercept for the LAMs with regularized estimation are 18. The effective degrees of freedom for each of the components for the mixtures of LAMs with regularized estimation varies distinctively from 2.0 to 17.6. The BIC criterion is equal to 25 378 for the best model of the mixtures of LMs with unregularized estimation and 25 356 for the mixture of LAMs with regularized estimation indicating that the mixture of LAMs with regularized estimation provides a slightly better model fit. The component sizes range from 0.03 to 0.16 for the best mixture of LMs with unregularized estimation and from 0.01 to 0.13 for the best mixture of LAMs with regularized estimation.

The implied partitions of the two models are shown in Figures 4 and 5. The expression patterns over time clearly differ between

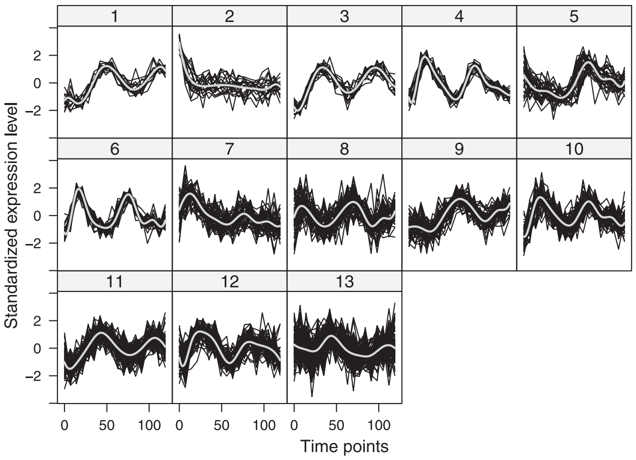

Fig. 4. Partition induced by the best finite mixture of LMs with unregularized estimation selected by the BIC together with the fitted cluster curves.

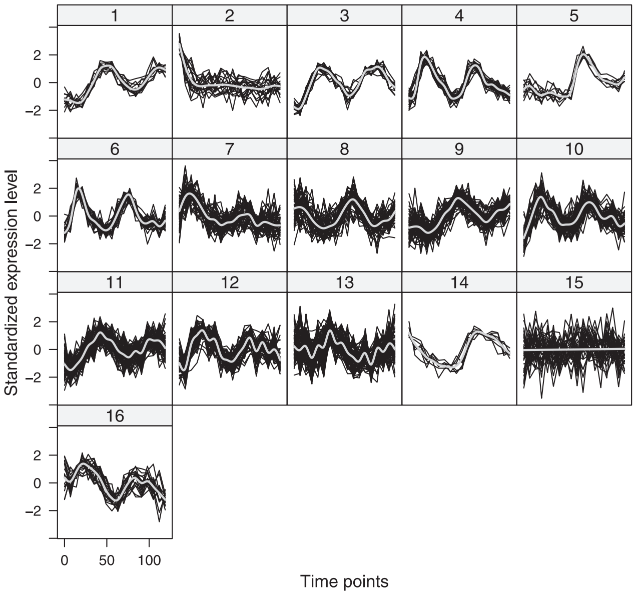

Fig. 5. Partition induced by the best finite mixture of LAMs with regularized estimation selected by the BIC together with the fitted cluster curves.

the components with respect to smoothness and variability. While the mixture of LAMs with regularized estimation allows to have components with different smoothness, the mixture of LMs with unregularized estimation needs to determine a compromise suiting all components. For some components, the selected smoothing seems to be too strong and a more flexible curve would better fit the expression patterns of the genes assigned to these components. The mixture of LAMs with regularized estimation does not need to select a smoothing parameter, which is appropriate for all components simultaneously. If the genes are split into more than five components, this is a strong advantage for this model class compared to mixtures of LMs with unregularized estimation. A further advantage of the best model fitted by mixtures of LAMs with regularized estimation is that a noise component is also formed where only a linear function is fitted effectively (see Component 15) and no change of the gene expression over time is observable. This component indicates that the 800 genes identified by Spellman *et al.* (1998) to be cell cycle regulated still contain genes not showing the expected periodicity in the alpha factor arrest.

5 SUMMARY AND OUTLOOK

Mixtures of LAMs with regularized estimation provide a convenient alternative and extension to mixtures of LMs using B-splines. Estimation within an ML framework is possible using an EM-type algorithm where a suitable smoothing parameter is selected iteratively between the E- and M-step in an additional H-step. The results on artificial data and the yeast cell cycle dataset are promising. Especially for the yeast cell cycle dataset, the advantage of automatically selecting different degrees of freedom for the components leads to superior results and allows to easily fit components with different degrees of smoothness.

The proposed model class provides a computationally more efficient way of determining the flexibility needed for the smoothing splines in each of the components. Already under the assumption that the flexibility needed is the same in all components a considerable number of models needs to be estimated and compared to choose the suitable number of degrees of freedom for finite mixtures of LMs, whereas the maximum number of degrees of freedom allowed when fitting finite mixtures of LAMs using regularized estimation is less crucial. If the smoothness of the fitted curves is allowed to vary over components, the number of different finite mixture models of LMs, which need to be compared, would be prohibitively large.

The proposed model class can also be used if the number of time points is only small. However, the computational advantages are less important, because a complete enumeration of all possible combinations of flexibility allowed in the components is more likely to be computationally feasible. Furthermore, the area of application is not restricted to time-course gene expression data. The proposed method could, for example, also be used to model stock prices over time.

In the future, the extension of regularized estimation of the model of finite mixtures of linear additive models to the regularized estimation of mixtures of linear additive mixed models could be considered. We assume that the performance, in general, as well as in comparison to linear mixed models should essentially be the same. Estimation, however, is more complex and the implementation is complicated by the fact that available standard tools for fitting linear mixed models do not allow for weighted ML estimation. This is necessary because the random effects are on the individual level and therefore they are integrated out before the individual log-likelihoods are weighted with the *a posteriori* probabilities in the M-step.

Funding: Austrian Science Fund (FWF): Hertha-Firnberg grant (T351-N18); Elise-Richter grant (V170-N18).

Conflict of Interest: none declared.

REFERENCES

Androulakis,I. *et al.* (2007) Analysis of time-series gene expression data: Methods, challenges, and opportunities. *Ann. Rev. Biomed. Eng.*, **9**, 205–228.

Biernacki,C. *et al.* (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Stat. Data Anal.*, **41**, 561–575.

Celeux,G. *et al.* (2005) Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Stat. Model.*, **5**, 243–267.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM-algorithm. *J. R. Stat. Soc. B*, **39**, 1–38.

Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Grün,B. and Leisch,F. (2008) FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softwr.*, **28**, 1–35.

Hastie,T. and Tibshirani,R. (1990) *Generalized Additive Models*, vol. 43 of *Monographs on Statistics and Applied Probability*. 1st edn. Chapman and Hall, London .

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Kneib,T. (2006) *Mixed Model Based Inference in Structured Additive Regression*. PhD Thesis, Institut für Statistik, Ludwig-Maximilians-Universität München.

Leisch,F. (2004) FlexMix: a general framework for finite mixture models and latent class regression in R. *J. Stat. Softwr.*, **11**, 1–18.

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

Maugis,C. *et al.* (2009) Variable selection for clustering with Gaussian mixture models. *Biometrics*, **65**, 701–709.

Ng,S.K. *et al.* (2006) A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, **22**, 1745–1752.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scharl,T. and Leisch,F. (2009) gcExplorer: interactive exploration of gene clusters. *Bioinformatics*, **25**, 1089–1090.

Scharl,T. *et al.* (2010) Mixtures of regression models for time-course gene expression data: Evaluation of initialization and random effects. *Bioinformatics*, **26**, 370–377.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Wood,S.N. (2006) *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Press, Boca Raton.

Wood,S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B*, **73**, 3–36.

APPENDIX A

A1. COMPUTATIONAL DETAILS

All computations are performed in the statistical computing environment R version 2.13.1 (R Development Core Team, 2011) with the packages **flexmix** 2.3-6, **multcomp** 1.2-7, **survival** 2.36-9, **mvtnorm** 0.9-9991, **modeltools** 0.2-18, **latice** 0.19-33, **grid** 2.13.1, **tools** 2.13.1. The EM algorithm for ML estimation of finite mixture models is implemented in the R package **flexmix** (Grün and Leisch, 2008; Leisch, 2004). For mixtures of linear additive models with regularized estimation, `FLXMRmgv()` is used as model driver for the H- and M-step. `FLXMRmgv()` uses functionality for regularized fitting of generalized additive models from the R package **mgcv** (Wood, 2011). The datasets for the simulation study using artificial data were conveniently generated using functions `gcSim()` and `gcData()` from the R package **gcExplorer** (Scharl and Leisch, 2009). The yeast cell cycle dataset is available in the Bioconductor package **yeastCC**.

B1. LIKELIHOOD MAXIMIZED IN THE H-STEP

In the H-step, the random effects are integrated out. This results in the following likelihood where only λ_k is assumed to be a parameter, β_k and σ_k^2 depend on λ_k and Y, X, Z and $\hat{c} = (\hat{c}_k)_{k=1,\dots,K}$ with $\hat{c}_k = (\hat{c}_{ki})_{j \in J_i, i=1,\dots,n}$ are assumed given.

$$\Delta_h(\lambda_k | Y, X, Z, \hat{c}) =$$

$$\int \prod_{i=1}^{n} \left[\prod_{j=1}^{J_i} f(y_{ij} | x_{ij}\beta_k + z_{ij} b_k, \sigma_k^2) \right]^{\hat{c}_{ik}} f(b_k | 0, \lambda_k^{-1} I) db_k$$

Using the following identity

$$\left[f(y_{ij} | x_{ij}\beta_k + z_{ij} b_k, \sigma_k^2) \right]^{\hat{c}_{ik}} = (2\pi\sigma_k^2)^{\frac{(1-\hat{c}_{ik})}{2}} \cdot$$

$$\left[f(\sqrt{\hat{c}_{ik}} y_{ij} | \sqrt{\hat{c}_{ik}} x_{ij} \beta_k + \sqrt{\hat{c}_{ik}} z_{ij} b_k, \sigma_k^2) \right]$$

as well as the fact that the marginal distribution for a linear random effects model is a multivariate normal distribution we have

$$\Delta_h(\lambda_k | Y, X, Z, \hat{c}) = \left(2\pi\sigma_k^2 \right)^{\alpha} f(\hat{Y}_k | \hat{X}_k \beta_k, \hat{Z}_k \hat{Z}_k^\top \lambda_k^{-1} + \sigma_k^2 I),$$

where

$$\alpha = \frac{1}{2} \sum_{i=1}^{n} J_i (1 - \hat{c}_{ik}).$$

I denotes the identity matrix of suitable dimension, $\hat{Y}_k = (\sqrt{\hat{c}_{ik}} y_{ij})_{ij}$ and $\hat{X}_k$ and $\hat{Z}_k$ are analogously defined. The diag$(\hat{c}_k^{-1})$ denotes diagonal matrix with $\hat{c}_k^{-1}$ in the diagonal. Transforming this back to the original variables gives

$$\Delta_h(\lambda_k | Y, X, Z, \hat{c}) = \left(2\pi\sigma_k^2 \right)^{\alpha} \left(\prod_{i=1}^{n} \sqrt{\hat{c}_{ik}^{-1}}^{J_i} \right) \cdot$$

$$f(Y | X\beta_k, ZZ^\top \lambda_k^{-1} + \sigma_k^2 \mathrm{diag}(\hat{c}_k^{-1})).$$

C1. ALGORITHMIC IMPLEMENTATION

In the following, the building blocks of the implementation are outlined which make use of functionality from packages **flexmix** and **mgcv**.

(1) Data pre-processing: the vector of responses, the model matrix for the fixed effects as well as the smoothers are determined using functionality from package **mgcv**.

(2) E-Step: given the current parameter estimates, the *a posteriori* probabilities are determined by predicting the mean values and evaluating the likelihood.

(3) H- and M-Step: for each component separately λ and the corresponding parameters are jointly determined with the fit function from package **mgcv** for the weighted data.