Clustering Multivariate Longitudinal Data: Hidden Markov of Factor Analyzers

Antonello Maruotti and Francesca Martella

Abstract Parsimonious Hidden Markov of Factor Analyzers models are developed by using a modified factor analysis covariance structure. This framework can be seen as a extension of the Parsimonious Gaussian mixture models (PGMMs) accounting for heterogeneity in a longitudinal setting. In particular, a class of 12 models are introduced and the maximum likelihood estimates for the parameters in these models are found using an AECM algorithm. The class of models includes parsimonious models that have not previously been developed. The performance of these models is discussed on a benchmark gene expression data. The results are encouraging and would deserve further discussion.

Key words: Clustering Longitudinal Data, Factor Analyzers, Hidden Markov Models, Dimensionality reduction

1 Introduction

In a longitudinal setting, repeated measurements are collected on the same (independent) units over several periods of time. Standard methods for longitudinal data analysis focus on the dependence of the variables on covariates, serial dependence, and heterogeneity in the individuals/units (see, e.g., [9]). A growing interest has been recently devoted to appropriately account for heterogeneity across the individual sequences (see e.g. [16]). To capture heterogeneity in a longitudinal setting, it is common to assume the existence of a latent process, driving and characterizing

Antonello Maruotti
Dipartimento di Isitituzioni Pubbliche, Economia e Societá, Universitá di Roma Tre, Via Gabriello Chiabrera, 199 - 00145 Roma e-mail: antonello.maruotti@uniroma3.it

Francesca Martella
Dipartimento di Scienze Statistiche, Sapienza Universitá di Roma, P.le Aldo Moro, 5 - 00185 Roma e-mail: francesca.martella@uniroma1.it

different data generation mechanisms ([6, 10, 12] provide interesting reviews on this topic in different contexts).

Recently, [16] introduce a model-based clustering technique for clustering longitudinal in a finite mixture framework. Each longitudinal sequence can be considered as a single *object* or *entity* belonging to one of the mixture components and all the individual sequences within the same component are characterized by the same generating mechanism. Other approaches have been developed by using *hierarchical* models ([4, 5, 11, 1, 13]) at the cost of an increasing computational burden.

In the following we are going to consider a multivariate Gaussian hidden Markov model (HMM; see [23] for a general introduction on HMMs), which can be seen as an extension of the finite mixture model [14] where individuals are allowed to move between the (hidden) components during the period of observation. Starting from the Parsimonious Gaussian mixture models (PGMMs) introduced by [15] and further extended by [17], we introduce a hidden Markov of factor analyzers by specifying a modified factor analysis covariance structure, including the possibility of imposing constraints which leads to a family of 12 models, including parsimonious models.

Parameter estimates can be obtained by an Alternating Expectation Conditional Maximization algorithm (AECM, [18]) in a HMM framework by adapting the well-know forward-backward algorithm ([3, 21]). The hidden Makov framework of factor analyzers is illustrated in the clustering of a representative dataset in the microarray literature: the yeast galactose data of [8]. The paper is organized as follows. Section 2 introduces the model by specifying some preliminaries on HMMs and providing extensions of the basic HMM in a multivariate clustering setting. Computational details are briefly described in Section 3, while Section 4 provides an illustrative example of the proposed models.

2 Model-based clustering of longitudinal data

In this section we firstly introduce the basic notation and the main assumptions on HMMs. Afterwards, we introduce in detail the hidden Markov of factor analyzers, pointing out the considered covariance structures and the computational aspects related to the estimate of model parameters.

2.1 Hidden Markov models

In a basic HMM for longitudinal data, the existence of two processes is assumed: an unobservable finite-state first-order Markov chain, S_{it}, $i = 1,\ldots,n$, $t = 0,\ldots,T$ with state space $\mathscr{S} = \{1,\ldots,m\}$ and an observed process, $\mathbf{Y}_{it} = \{Y_{it1}, Y_{it2}, \ldots, Y_{itJ}\}$, where Y_{itj} denotes the j-th response variable for individual i at time t (similarly for S_{it}).

We assume that the distribution of $\mathbf{Y}_{it}$ depends only on S_{it}, specifically the $\mathbf{Y}_{it}$, $t = 1,\ldots,T$, are conditionally independent given the S_{it}:

$$f(\mathbf{Y}_{it} = \mathbf{y}_{it} \mid \mathbf{Y}_{i0} = y_{i0},\ldots,\mathbf{Y}_{it-1} = y_{it-1}, S_{i0} = s_{i0},\ldots,S_{it} = s_{it}) = f(\mathbf{Y}_{it} = \mathbf{y}_{it} \mid S_{it} = s_{it}) \quad (1)$$

Typically it is assumed that the state-dependent distributions, i.e. the distributions of $\mathbf{Y}_{it}$ given S_{it}, come from a parametric family of continuous or discrete distributions. Thus, the unknown parameters in a HMM involve both the parameters of the Markov chain and those of the state-dependent distributions of the random variables $\mathbf{Y}_{it}$. In particular, the parameters of the Markov chain are the elements of the transition probability matrices $\mathbf{Q} = \{q_{itlk}\}$, where $q_{itlk} = \Pr(S_{it} = k \mid S_{it-1} = l), l,k \in \mathscr{S}$ is the probability that individual i visits state k at time t given that at time $t-1$ he/she was in state l, and the initial probabilities $\delta = \{\delta_{il}\}$, where $\delta_{il} = \Pr(S_{i0} = l)$, i.e. the probability of being in state l at time 0. The simplest model in this framework is the homogeneous HMM, which assumes common transition and initial probabilities, i.e. $q_{itlk} = q_{lk}$ and $\delta_{il} = \delta_l$. We will focus on homogeneous HMMs to simplify the discussion, but of course the hidden Markov chain can be assumed to be non-homogeneous: the transition probabilities may be individual and/or time varying and modeled via a logit function of explanatory variables.

2.2 Hidden Markov of Factor Analyzers

Consider an HMM with $\mathbf{Y}_{it}$ being multidimensional with the conditional distribution of $\mathbf{Y}_{it}$ given $S_{it} = s_{it}$ being $N(\mu_{s_{it}}, \Sigma_{s_{it}})$, i.e. multivariate Gaussian with state-dependent mean, $\mu_{s_{it}}$, and covariance matrix $\Sigma_{s_{it}}$. In line with the more general mixture of factor analyzers framework, we assume that conditionally to the s_{it}-th state, the random vector $\mathbf{y}_{it}$ is modelled using a H-dimensional vector of latent factors $\mathbf{w}_{is_{it}}$ (typically $H \ll J$) as $\mathbf{y}_{it} = \mu_{s_{it}} + \Lambda_{s_{it}}\mathbf{w}_{is_{it}} + \mathbf{e}_{it}$, where $\Lambda_{s_{it}}$ is a $J \times H$ matrix of factor weights, the latent variables $\mathbf{w}_{is_{it}} \sim MVN(\mathbf{0}, \mathbf{I}_H)$, and $\mathbf{e}_{it} \sim MVN(\mathbf{0}, \Psi_{s_{it}})$, where $\Psi_{s_{it}}$ is a $J \times J$ diagonal matrix. Thus, conditionally on the s_{it}-th state, the density of $\mathbf{y}_{it}$ is $MVN(\mathbf{0}, \Lambda_{s_{it}}\Lambda'_{s_{it}} + \Psi_{s_{it}})$. Therefore, the marginal density of a hidden Markov of factor analyzers is given by:

$$f(\mathbf{y}_i) = \sum_{\mathscr{S}^T} \delta_{s_{i0}} \prod_{t=1}^{T} q_{s_{it-1}s_{it}} \prod_{t=0}^{T} \frac{\exp\left[-\frac{1}{2}(\mathbf{y}_{it} - \mu_{s_{it}})'(\Lambda_{s_{it}}\Lambda'_{s_{it}} + \Psi_{s_{it}})^{-1}(\mathbf{y}_{it} - \mu_{s_{it}})\right]}{(2\pi)^{J/2} \mid \Lambda_{s_{it}}\Lambda'_{s_{it}} + \Psi_{s_{it}} \mid^{1/2}} \quad (2)$$

where $\sum_{\mathscr{S}^T}$ denotes summation over all realizations $s_i t$, $t = 0,\ldots,T$, for individual i.

Note that the proposed model can be seen as an extension of the mixture of factor analyzers model by allowing time dependence and, following the idea in [17], constraints across groups on the $\Lambda_{s_{it}}$ and $\Psi_{s_{it}}$ matrices and on whether or not $\Psi_{s_{it}} = \psi_{s_{it}}\Xi_{s_{it}}$, where $\psi_{s_{it}} \in \mathbb{R}^+$ and $\Xi_{s_{it}} = diag\{\xi_1,\ldots,\xi_J\}$ such that $|\Xi_{s_{it}}| = 1$. The

full range of possible constraints provides a class of 12 different Hidden Markov of Factor Analyzers models, which are given in Table 1. Note that CCCC and CCCU assume the equal isotropic noise whereas UCCC and CUUU assume the unequal isotropic noise. The other eight covariance structures incorporating constraints on the loading matrices dramatically reduces the number of covariance parameters and lead to parsimonious models.

Table 1 Covariance structure in a hidden Markov of factor analyzers framework

Model ID	$\Lambda_{s_{it}} = \Lambda$	$\Xi_{s_{it}} = \Xi$	$\psi_{s_{it}} = \psi$	$\Xi_{s_{it}} = \mathbf{I}_J$	
CCCC	Constrained	Constrained	Constrained	Constrained	$\Sigma = \Lambda\Lambda' + \psi\mathbf{I}_J$
CCCU	Constrained	Constrained	Constrained	Unconstrained	$\Sigma = \Lambda\Lambda' + \psi\Xi$
CCUC	Constrained	Constrained	Unconstrained	Constrained	$\Sigma_{s_{it}} = \Lambda\Lambda' + \psi_{s_{it}}\mathbf{I}_J$
CUUU	Constrained	Unconstrained	Unconstrained	Unconstrained	$\Sigma_{s_{it}} = \Lambda\Lambda' + \psi_{s_{it}}\Xi_{s_{it}}$
UCCC	Unconstrained	Constrained	Constrained	Constrained	$\Sigma_{s_{it}} = \Lambda_{s_{it}}\Lambda'_{s_{it}} + \psi\mathbf{I}_J$
UCCU	Unconstrained	Constrained	Constrained	Unconstrained	$\Sigma_{s_{it}} = \Lambda_{s_{it}}\Lambda'_{s_{it}} + \psi\Xi$
UCUC	Unconstrained	Constrained	Unconstrained	Constrained	$\Sigma_{s_{it}} = \Lambda_{s_{it}}\Lambda'_{s_{it}} + \psi_{s_{it}}\mathbf{I}_J$
UUUU	Unconstrained	Unconstrained	Unconstrained	Unconstrained	$\Sigma_{s_{it}} = \Lambda_{s_{it}}\Lambda'_{s_{it}} + \psi_{s_{it}}\Xi_{s_{it}}$
CCUU	Constrained	Constrained	Unconstrained	Unconstrained	$\Sigma_{s_{it}} = \Lambda\Lambda' + \psi_{s_{it}}\Xi$
UCUU	Unconstrained	Constrained	Unconstrained	Unconstrained	$\Sigma_{s_{it}} = \Lambda_{s_{it}}\Lambda'_{s_{it}} + \psi_{s_{it}}\Xi$
CUCU	Constrained	Unconstrained	Constrained	Unconstrained	$\Sigma_{s_{it}} = \Lambda\Lambda' + \psi\Xi_{s_{it}}$
UUCU	Unconstrained	Unconstrained	Constrained	Unconstrained	$\Sigma_{s_{it}} = \Lambda_{s_{it}}\Lambda'_{s_{it}} + \psi\Xi_{s_{it}}$

3 Computational details

Even if this form of the likelihood has several appealing properties, as it stands expression (2) is of little or no computational use, because it involves a sum over m^T terms for each unit i and cannot be directly evaluated. It quickly becomes infeasible to compute even for small values of m as T grows to moderate size. Clearly, a more efficient procedure is needed to perform the calculation of the likelihood function. This issue may be addressed via the so-called forward variables ([3, 21]). To estimate model parameters, an Alternating Expectation Conditional Maximization (AECM) algorithm introduced by [18] is used. This algorithm is an extension of the EM algorithm using different definitions of missing data at different stages. The AECM algorithm tends to be preferred to its alternatives due to its robustness and ease of application in various scenarios, especially when the model parameters are constrained. For homogeneous HMMs, the AECM reduces to an iterative procedure with simple, closed form expressions for parameter estimates at each iteration. It is based on complete-data log-likelihood, i.e., the log-likelihood of the observations (the incomplete data) plus the states (the missing data). Before deriving the complete data log-likelihood, we define $u_{itl} = I(S_{it} = l)$ as an indicator variable equal to

1 if unit i is in state l at time t and 0 otherwise, and $v_{itlk} = I(S_{it} = k, S_{it-1} = l)$ as an indicator variable equal to 1 if unit i is in to state l at time $t-1$ and in state k at time t, 0 otherwise. Moreover, we partition the vector of unknown parameters Φ in (Φ_1 , Φ_2); Φ_1 contains the transition probabilities q_{lk} and δ_l and $\mu_{s_{it}}$. The Φ_2 contains the elements of $\Lambda_{s_{it}}$, $\Psi_{s_{it}}$ and $\mathbf{w}_{is_{it}}$. At the first stage of the algorithm, we define the state labels as missing data, and the complete data log-likelihood function has the following form:

$$\ell_{c_1}(\theta) = \sum_{i=1}^{n} \left\{ \sum_{l=1}^{m} u_{i0l} \log \delta_l + \sum_{t=1}^{T} \sum_{l=1}^{m} \sum_{k=1}^{m} v_{itlk} \log q_{lk} + \sum_{t=0}^{T} \sum_{l=1}^{m} u_{itl} \log f(\mathbf{y}_{it} \mid S_{it} = l) \right\}, \tag{3}$$

where

$$f(\mathbf{y}_{it} \mid S_{it} = l) = \frac{\exp\left[-\frac{1}{2}(\mathbf{y}_{it} - \mu_l)'(\Lambda_l \Lambda_l' + \Psi_l)^{-1}(\mathbf{y}_{it} - \mu_l)\right]}{(2\pi)^{J/2} \mid \Lambda_l \Lambda_l' + \Psi_l \mid^{1/2}}$$

Thus, the first E-step consists of calculating the conditional expectation of expression (3) by replacing all the quantities u_{itl} and v_{itlk} with their conditional expectations $\hat{u}_{itl}$ and $\hat{v}_{itlk}$, given the current values of the parameters and the observed data. On the other hand, at the first CM-step, the expected complete-data log-likelihood is maximized with respect to μ_l, δ_l and q_{lk} obtaining:

$$\hat{\mu}_l = \frac{\sum_{i=1}^{n} \sum_{t=0}^{T} \hat{u}_{itl} \mathbf{y}_{it}}{\sum_{i=1}^{n} \sum_{t=0}^{T} \hat{u}_{itl}}, \quad \hat{\delta}_l = \frac{\sum_{i=1}^{n} \hat{u}_{i0l}}{n}; \quad \hat{q}_{lk} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \hat{v}_{itlk}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{k=1}^{m} \hat{v}_{itlk}}.$$

At the second stage of the AECM algorithm, we use $\hat{\mu}_l$, $\hat{\delta}_l$ and $\hat{q}_{lk}$ obtained above, when estimating Λ_l and Ψ_l and consider the state labels and the factors to be the missing data. Therefore, the complete data log-likelihood is

$$\ell_{c_2}(\theta) = \sum_{i=1}^{n} \left\{ \sum_{l=1}^{m} u_{i0l} \log \delta_l + \sum_{t=1}^{T} \sum_{l=1}^{m} \sum_{k=1}^{m} v_{itlk} \log q_{lk} + \sum_{t=0}^{T} \sum_{l=1}^{m} u_{itl} \log f(\mathbf{y}_{it} \mid S_{it} = l, \mathbf{w}_{il}) + \sum_{t=0}^{T} \sum_{l=1}^{m} \log u_{itl} f(\mathbf{w}_{il}) \right\}. \tag{4}$$

In a similar manner as before, the estimates of Λ_l and Ψ_l can be easily derived under the different imposed constraints (not shown her for sake of brevity). The AECM algorithm iteratively updates the parameters until convergence to maximum likelihood estimates of the parameters. As a by-product of the estimation procedure we have the possibility of classifying genes on the basis of their posterior probability estimates $\hat{u}_{itl}$. In fact, the i-th gene can be classified to the l-th group (component of

the estimated mixture) if $\hat{u}_{itl} = \max(\hat{u}_{it1}, \hat{u}_{it2}, \ldots, \hat{u}_{itm})$. It is worth noticing that each group is characterized by homogeneous values of the estimated parameters.

4 The yeast galactose data

To discuss the empirical performance of the proposed model, we use a typical gene expression dataset where the expression levels are measured at many time points or under different conditions to elucidate genetic networks or some important biological process. Specifically, this dataset has been used to study integrated genomic and proteomic analyses of a systemically perturbed metabolic network ([8]). The experiments included single gene deletion involving nine of the key genes (GAL1, GAL2 GAL3 GAL4, GAL5(PGM2), GAL6(LAP3), GAL7 , GAL10, GAL80) that participate in yeast galactose metabolism. For each experiment, one of the nine genes was deleted, or alternatively, the experiment used a wild-type cell wherein no genes were deleted. For each of those 10 experimental conditions, galactose was available extracellularly in one set of experiments and absent in another set. Thus, there were a total of $T = 20$ different experimental conditions. Since each of those 20 experiments refers $J = 4$ experimental conditions, the overall dataset contains 80 experiments. As in ([22]) and ([19]), we imputed all the missing values using a k-nearest neighbor method. The resulting $n = 205$ gene expression levels reflect four functional categories in the Gene Ontology (GO) listings ([2]). Thus, we applied a hidden Markov of factor analyzers to group genes into m=4 states; we do not discuss fitting for varying numbers of states m, since we would analyze the performance of our proposal in reproducing the known functional categories. Genes are allowed to move among the states during the period of observation. In fact, a gene can be associated with multiple biological functions, due to the fact that genes often have several distinct roles in regulation processes. Therefore the assumption of assigning a gene only to one state (or cluster) is an oversimplification for a biological system. In the following we summarize the potential of the proposed approach. We look at three over twelve factorial parameterizations as illustrative examples. The evolution over time is presented in Figure 1, while a comparison in terms of $BIC = 2 \times \ell + \#parameters \times \log(n)$ and goodness-of-classification with PGMMs is provided in Table 2. We classify each gene in the state maximizing its posterior membership probability deriving the unobserved sequence of states. Figure 1 shows the hidden sequences of hidden states; it is clear that time dependence and heterogeneity play an important role in the classification, since genes seem to change their *behavior*, moving across states over time.

Furthermore, we provide a measure of the quality of the classification by the index

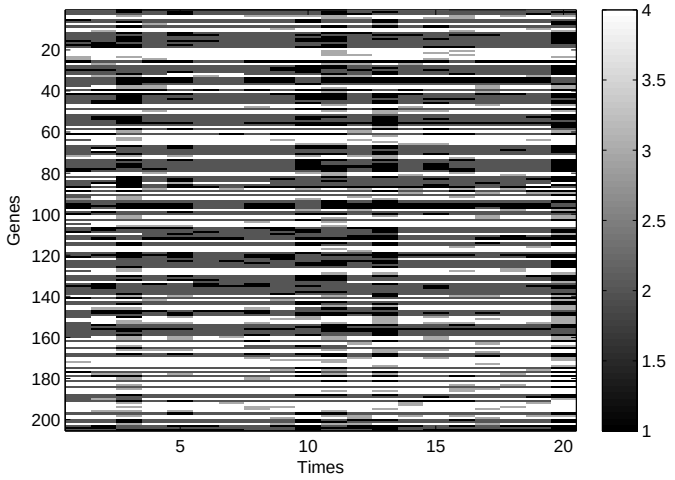
$$S = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T_i} \left(\max(\hat{u}_{it1}, \hat{u}_{it2}, \ldots, \hat{u}_{itm}) - \frac{1}{m} \right)}{(1 - \frac{1}{m}) \sum_{i=1}^{n} T_i}$$

Index S is always between 0 and 1, with 1 corresponding to the situation of absence of uncertainty in the classification, since one of such posterior probabilities is equal to 1 for every individual at every time, with all the other probabilities equal to 0. It helps in identifying if the population clusters are sufficiently well separated. It is worth noting that each state is characterized by homogeneous values of estimated random effects; thus, conditionally on observed covariates values, subjects from that state have a similar propensity to the event of interest. The UCCU HMM is the preferred model, providing the best goodness-of- classification and the best BIC. This confirms the importance of appropriately account for all longitudinal data characteristics.

Table 2 Summary results

Model	H	PGMM		HMM	
		BIC	S	BIC	S
UUUU					
	2	14821.53	0.758	16867.12	0.931
	3	14759.23	0.756	16840.02	0.932
UCCU					
	2	14659.44	0.795	16820.41	0.934
	3	14859.23	0.758	16903.45	0.932
UCUC					
	2	7611.761	0.735	10413.41	0.970
	3	12162.73	0.864	14131.45	0.919

Fig. 1 Hidden states sequences for the 205 genes over 20 times



References

1. Alfó M. and Maruotti, A.: A hierarchical model for time dependent multivariate longitudinal data. In: Data Analysis and Classification. Springer Series on Studies in Classification, Data Analysis and Knowledge Organization. C. Lauro, F. Palumbo (eds), Springer-Verlag, 271-279 (2010).
2. Ashburner, M. Ball, C.A. Blake, J.A. Botstein, D. Butler, H. Cherry, J.M. et al. Gene Ontology: tool for the unification of biology. Nat Genet **25**, 25-29.
3. Baum, L.E., Petrie, T., Soules, G. and Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. **41**, 164–171 (1970).
4. Celeux, G., Martin, O. and Lavergne, C.: Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. Stat. Model. **5**, 243–267.
5. De la Cruz-Meía, R., Quintana, F. A. and Marshall, G.: Model-based clustering for longitudinal data. Comp. Stat. and Data Anal. **52**, 1441–1457 (2008).
6. Frühwirth-Schanatter, S.: Panel data analysis: a survey on model-based clustering of time series. Adv. Data Anal. Classif. **5**, 251–280 (2011).
7. Ghahramani, Z. and Hinton, G. E.: The EM algorithm for factor analyzers. Technical Report CRG-TR- 96-1, University of Toronto (1997).
8. Ideker, T. Thorsson, V. Ranish, J.A. Christmas, R. Buhler, J. Eng, J.K. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science **92**, 929-934 (2001).
9. Laird, N.M. and Ware, J.H.: Random effects models for longitudinal data. Biometrics **38**, 963–974 (1982).
10. Martella, F. and Vermunt, J.K.: Model-based approaches to synthesize microarray data: a unifying review using mixture of SEMs. Stat. Methods Med. Res. (2012), doi: 10.1177/0962280211419482
11. Maruotti, A. and Rydén, T.: A semiparametric approach to hidden Markov models under longitudinal observations. Statist. Comput. **19**, 381–393 (2009).
12. Maruotti, A.: Mixed hidden Markov models for longitudinal data: an overview. Int. Stat. Rev. **79**,427–454 (2011).
13. Maruotti, A. and Rocci, R.: A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. Stat. Med. (2012) doi: 10.1002/sim.4478.
14. McLachlan, G.J. and Peel, D.: Finite mixture models. Wiley series in probability and statistics. Wiley, New York (2000).
15. McNicholas P.D. and Murphy T.B.: Parsimonious Gaussian mixture models. Stat. Comput. **18**, 285–296 (2008).
16. McNicholas P.D. and Murphy T.B.: Model-based clustering of longitudinal data. The Canadian Journal of Statistics **38**, no. 1, 153–168 (2010).
17. McNicholas P.D. and Murphy T.B.: Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics **26**, no. 21, 2705–2712 (2010).
18. Meng, X.L. and Van Dyk, D.A.: The EM algorithm -an old folk song sung to a fast new tune. Journal R. Statist. Soc., B **59** 511–567 (1997).
19. Ng, S.K. McLachlan, G.J. Bean, R.W. and Ng, S.W. Clustering replicated microarray data via mixtures of random effects models for various covariance structures. In: Boden Mand Bailey TL (eds.) Conferences in research and practice in information technology. The Australian Computer Society, Sydney, **73** 29-33 (2006).
20. Tipping, T. E. and Bishop, C. M.: Mixtures of probabilistic principal component analysers. Neural Comp. **11**, 443–482 (1999).
21. Welch, L.R.: Hidden Markov models and the Baum-Welch algorithm. IEEE Inf. Theory Soc. Newsl. **53**, 1–13 (2003).
22. Yeung, K.Y. Medvedovic, M. Bumgarner, R.E. Clustering gene-expression data with repeated measurements. Genome Biol, 4(R34) (2003).
23. Zucchini, W. and MacDonald, I.L.: Hidden Markov Models for Time Series: An Introduction Using R. Boca Raton, FL: Chapman & Hall (2009).