EURASIP Journal on Applied Signal Processing 2005:8, 1159–1173

Clustering Time Series Gene Expression Data Based on Sum-of-Exponentials Fitting

Ciprian Doru Giurcăneanu
Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland
Email: ciprian.giurcaneanu@tut.fi

Ioan Tăbuş
Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland
Email: ioan.tabus@tut.fi

Jaakko Astola
Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland
Email: jaakko.astola@tut.fi

Received 8 June 2004; Revised 26 October 2004; Recommended for Publication by Xiaodong Wang

This paper presents a method based on fitting a sum-of-exponentials model to the nonuniformly sampled data, for clustering the time series of gene expression data. The structure of the model is estimated by using the minimum description length (MDL) principle for nonlinear regression, in a new form, incorporating a normalized maximum-likelihood (NML) model for a subset of the parameters. The performance of the structure estimation method is studied using simulated data, and the superiority of the new selection criterion over earlier criteria is demonstrated. The accuracy of the nonlinear estimates of the model parameters is analyzed with respect to the Cramér-Rao lower bounds. Clustering examples of gene expression data sets from a developmental biology application are presented, revealing gene grouping into clusters according to functional classes.

Keywords and phrases: nonuniformly sampled data, sum-of-exponentials model, normalized maximum likelihood, time series clustering, gene expression data, developmental biology.

1. INTRODUCTION

The gene expression time profiles are a rich source of information about the dynamics of the underlying genomic network. The experiments are often taken at nonuniform time points, suggested by the biologist's intuition about the time scale of the important changes in the analyzed biological process, for example, a developmental process or administration of a drug. Clustering the time profiles of the thousands of genes recorded by the microarrays is a very important exploratory problem, for which several methods have been proposed in the past [1, 2, 3].

Most of the existing methods, no matter whatever heuristically motivated, or model-based methods [4] do not make use of the time values at which the measurements have been taken, loosing potentially useful information regarding the analyzed waveforms. Some approaches that take into account the temporal structure in gene expression data are based on hidden Markov model [5], spline approximation [6], or on analysis of temporal variation [7]. In [8], an autoregressive model is used for the gene expression time series, and the clustering is performed with a Bayesian criterion which measures the similarity between two time series. A comprehensive study on various clustering methods applied to gene expression data that are time series can be found in [9].

A general methodology for modelling the time series collected at nonuniform time points has been presented in [10], where the sum-of-exponentials model was used for getting estimates of the gains and time constants, and then a generalized correlation coefficient was introduced based on the cost of describing all relevant parameters of the waveforms interpolated at an equidistant grid. The generalized correlation coefficient was intended for various applications, including gene prediction in genetic networks and disease classification. The sum-of-exponentials model is appealing since it can be interpreted as the transient output of a linear dynamical system evolving from a first stationary regime to another one.

The work in [10] was first extended in a preliminary version of this paper [11], by elaborating on the first critical stage, that of fitting the sum-of-exponentials model. We introduce a new minimum description length (MDL) criterion for selecting the number of exponentials in the model, and provide a statistical analysis of the fitting accuracy. A fitting procedure combining several known methods is also proposed and is found experimentally to have an accuracy close to Cramér-Rao lower bound.

We apply the proposed methods for finding the dynamical parameters in the models of the time series from two experimental data sets containing gene expressions measured during the development of mouse cerebellum and dentate gyrus. The estimated parameters are subsequently used in a clustering algorithm.

The remainder of the paper is organized as follows. In the next section, we outline the algorithm for fitting a sum of exponentials to nonuniformly sampled data. The expression of Cramér-Rao lower bound is obtained, and the new MDL criterion is introduced for estimating the structure parameter. Based on sum-of-exponentials model, a new clustering procedure is proposed in Section 3, then is tested with simulated data. An experiment with data from developmental biology is conducted in Section 4, and the enrichment of functional categories in clusters found by the proposed method is investigated with statistical tests.

2. FITTING A SUM OF EXPONENTIALS TO NONUNIFORMLY SAMPLED DATA

2.1. Motivation

A linear differential equations model for the concentrations of mRNA and proteins was introduced in [12]. In [13], since usually only the concentrations of mRNA are measured, the differential equations model was modified to the form

$$\frac{d}{dt}\mathbf{y}(t) = \mathbf{M}y(t), \tag{1}$$

where $\mathbf{y}(t)$ contains only mRNA concentrations as a function of time, and the constant matrix $\mathbf{M}$ describes interactions between various genes. Biological considerations lead to constraining the eigenvalues and the eigenvectors of matrix $\mathbf{M}$ to be real-valued. Supplementary, the eigenvalues of $\mathbf{M}$ are assumed to be negative to ensure that $\exp(\mathbf{M}t) \to \mathbf{0}$ as $t \to \infty$. It is well known that the solution for (1) is given by $\mathbf{y}(t) = \exp(\mathbf{M}t)\mathbf{y}(0)$, where $\mathbf{y}(0)$ is the vector of measurements at time moment zero. The solution can be expressed as a sum-of-exponential terms multiplied by polynomial functions of t, which is the basic reason for our choosing of the sum of exponentials as a dynamic model for gene expressions. In [12] it was argued that only the case when all these polynomials are constant (have degree zero) has biological relevance. More on linear differential equations models for gene expression data can be found in [14, 15].

2.2. Problem formulation

We consider an estimation procedure for the following nonlinear regression model:

$$y(t|\theta_\gamma) = \sum_{j=1}^{p} \alpha_j \exp(-\beta_j t), \tag{2}$$

where the model parameters are $\theta_\gamma = \{\alpha_j, \beta_j \mid j = 1, 2, \ldots, p\}$, and the structure parameter is $\gamma = 2p$. The α_j's are real-valued and β_j's are taken, without loss of generality, to verify $\beta_1 > \beta_2 > \cdots > \beta_p > 0$. The noisy signal

$$z(t) = y(t|\theta_\gamma) + \varepsilon(t) \tag{3}$$

is observed at the nonnegative time points $t_1 < t_2 < \cdots < t_n$, which are not equally spaced. The vector of measurements is $\mathbf{z} = [z(t_1) \cdots z(t_n)]^\top$. The noise $\varepsilon(t)$ is assumed to be stationary and to have finite variance. For a fixed structure γ, we define the LS estimates of the parameters

$$\hat{\theta}_\gamma = \arg\min_{\theta_\gamma} \sum_{i=1}^{n} [z(t_i) - y(t_i|\theta_\gamma)]^2 \tag{4}$$

and introduce the residual sum of squares as the sum in (4) evaluated at the LS parameters

$$\mathrm{RSS}(\gamma) = \sum_{i=1}^{n} [z(t_i) - y(t_i|\hat{\theta}_\gamma)]^2. \tag{5}$$

If time moments are equally spaced, the estimation problem can be rephrased as a simple linear least-squares problem, but even then important difficulties arise when fitting a sum of exponentials: choosing initial values and ill-conditioning when two or more β_j's are close [16]. Since the problem is complex, many algorithms have been proposed to solve it, beginning with the Prony's method introduced as early as 1795. The method was originally used for fitting an exponential model to uniformly sampled experimental data, and consists in solving a set of linear equations for the recurrence equation that the signals satisfy. It was shown in [17] that Prony's method is close to Pisarenko's method, which was analyzed and further improved in signal processing community. Many modified Prony algorithms have been also proposed, see, for example, [16, 18]. A survey on various algorithms for fitting a sum of exponentials can be found in [19], where a special section is dedicated to minimizing the LS criterion by using standard optimization techniques. One such technique is the Al-Baali-Fletcher algorithm [20], which is a hybrid method in the sense that during the iterations the algorithm switches between GN (Gauss-Newton) and BFGS (Broyden-Fletcher-Goldfarb-Shanno) for the estimation of the Hessian matrix.

When fitting a sum of exponentials by minimizing an LS criterion, a critical part is the choice of initial values for the α_j and β_j parameters. An algorithm for finding initial values in the particular case when all α_j coefficients are strictly positive is given in [21]. In the general case when α_j's are not constrained, the grid search is generally applied [22].

2.3. An estimation procedure for the nonlinear regression model

Fitting an exponential model to gene expressions is hard since the number of available measurements is small and they are nonuniformly sampled. We resort to a grid search for initializing the parameters. For simplicity of notation, we define two vectors of parameters, namely, $\mathbf{a} = [\alpha_1\ \alpha_2 \cdots \alpha_p]^\top$ and $\mathbf{b} = [\beta_1\ \beta_2 \cdots \beta_p]^\top$. At each point $\mathbf{b}$ in the grid, the linear parameters $\hat{\mathbf{a}}(\mathbf{b})$ are fitted as shown in Appendix A by minimizing the sum of residual squares. The starting point is chosen to be the pair $(\hat{\mathbf{a}}(\mathbf{b}), \mathbf{b})$ that minimizes the sum of residual squares over all points of the grid defined in the space of $\mathbf{b}$ parameters. The selected pair $(\hat{\mathbf{a}}(\mathbf{b}), \mathbf{b})$ is used to initialize the Al-Baali-Fletcher optimization algorithm [20]. We employ the Matlab implementation of this algorithm as provided by the Tomlab software, which is publicly available at http://www.mdh.se/ima/personal/khm01/tom/.

Cramér-Rao lower bound (CRLB)

For investigating some statistical aspects of the estimation problem, we resort to the computation of the Cramér-Rao lower bound (CRLB). We denote by $\mathbf{F}^{-1}(\theta_\gamma)$ the inverse of the Fisher information matrix for the signal model when the parameters are θ_γ. Let $\hat{\theta}_j$ be an unbiased estimator for the j'th component of θ_γ. A classical result from statistics [23] states that the CRLB for the variance $\mathrm{var}(\hat{\theta}_j)$ is given by $[\mathbf{F}^{-1}(\theta_\gamma)]_{jj}$, where the index jj designates the entry of the matrix for which both the row and the column are equal to j.

The independence assumption is rather strong for gene expression, and further studies will be needed in order to estimate a correlated model for the noise, especially when the number of data points available will increase. Under the hypothesis of white Gaussian noise, we find in Appendix B closed-form expressions for the entries of the Fisher information matrix. First we obtain these expressions for the set of parameters

$$\theta'_\gamma = \{\alpha_1, \alpha_2, \ldots, \alpha_p, \delta_1, \delta_2, \ldots, \delta_p\}, \tag{6}$$

where $\delta_j = \exp(-\beta_j)$, $1 \le j \le p$, denote the decay rates. Since time constants are more important for the interpretation of results obtained with gene expression data, we further develop the calculus for the Fisher information matrix when the set of parameters is given by

$$\theta''_\gamma = \{\alpha_1, \alpha_2, \ldots, \alpha_p, \tau_1, \tau_2, \ldots, \tau_p\}. \tag{7}$$

In the definition above, we use τ_j for the time constants, namely, $\tau_j = T_0/\beta_j$, $1 \le j \le p$, where T_0 is the greatest common divisor of $\{t_2 - t_1, t_3 - t_2, \ldots, t_n - t_{n-1}\}$, and $t_1, t_2, \ldots, t_n$ are assumed to be integer-valued. It is obvious how to obtain the conversion between the time constants and the decay terms: $\tau_j = -T_0/\ln \delta_j$ and $\delta_j = \exp(-T_0/\tau_j)$.

Under the hypothesis of white Gaussian noise with zero mean and variance σ^2, the expression of the log-likelihood function is given by

$$\Lambda(\mathbf{z}|\theta_\gamma) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[z(t_i) - y(t_i|\theta_\gamma)\right]^2. \tag{8}$$

It is immediate to observe that the ML estimator for the nonlinear regression model is the one given by (4). In general, the ML estimator is optimal since asymptotically it is unbiased and achieves the CRLB [23], which is a highly desired property. We are interested to assess the estimation results for finite samples, and especially when the number of measurements is small. Even in these cases, it is customary to compare the variance of estimates with CRLB, but two important facts have to be considered when interpreting the results [24]: (a) there exist biased estimators for which the variance is even smaller than CRLB, as shown in the examples discussed in [25, 26]; (b) the ML estimator achieves *asymptotically* the CRLB, but an estimator which achieves the lower bound may not exist for small samples.

Structure parameter estimation

In the discussion above, we have assumed that the structure parameter $\gamma = 2p$ is known, or equivalently the number of exponential terms in (2) is given. This is not the case in practical applications, thus we need to estimate the value of γ, which amounts to select this value from a finite set of positive even integers. The selection is usually performed based on well-known criteria as MDL or AIC [19]. When applying the form of MDL principle called two-stage description length [27], the structure parameter is given by $\gamma^* = \arg\min_{2\le\gamma\le\gamma_{\max}} \mathrm{MDL}(\gamma)$, where

$$\mathrm{MDL}(\gamma) = \frac{n}{2}\log \mathrm{RSS}(\gamma) + \frac{\gamma}{2}\log n. \tag{9}$$

We use the notation $\log(\cdot)$ to denote the logarithm base two. The MDL criterion represents the ideal code length for transmitting the values of measurements $z(t_1), z(t_2), \ldots, z(t_n)$ from a hypothesized encoder to a decoder. For a fixed structure γ, the parameters θ_γ are estimated as described above, and each parameter is encoded by using $(1/2)\log n$ bits, which leads to a total cost that equals the second term in (9). The first term represents the number of bits necessary for encoding $z(t_1), z(t_2), \ldots, z(t_n)$ given the estimated values $\hat{\theta}_\gamma$; $\mathrm{RSS}(\gamma)$ is calculated as in (5).

We propose to apply a different coding scenario that allows the use of recent advances in universal modeling, namely, the normalized maximum-likelihood (NML) estimator. The key observation is that once the estimated values for β's are known both at the encoder and at the decoder sites, the modeling problem reduces to a linear regression model as shown in Appendix A. Therefore, it is straightforward to use for the ideal code length the NML criterion introduced in [28]. It remains only to find a method for transmitting the estimated values of β_j's from the encoder to the decoder. A natural solution is to encode every β_j parameter by using the asymptotically optimal number of bits, namely, $(1/2)\log n$ bits. We obtain now the $\mathrm{nMDL}(\gamma)$ criterion as a sum of two terms: the first one is given by NML formula from [29], and the second one is $(\gamma/4)\log n$, the cost for transmitting the β_j's. Therefore, we obtain

$$\begin{aligned}\mathrm{nMDL}(\gamma) = {} & \frac{n-\gamma/2}{2}\log\frac{\mathrm{RSS}(\gamma)}{n} + \frac{\gamma}{4}\log\frac{\hat{\mathbf{a}}^\top\hat{\mathbf{B}}^\top\hat{\mathbf{B}}\hat{\mathbf{a}}}{n} \\ & - \log\Gamma\left(\frac{n-\gamma/2}{2}\right) - \log\Gamma\left(\frac{\gamma}{4}\right) + \frac{\gamma}{4}\log n,\end{aligned} \tag{10}$$

where $\hat{\mathbf{a}} = [\hat{\alpha}_1\ \hat{\alpha}_2 \cdots \hat{\alpha}_{\gamma/2}]^\top$, the entries of the matrix $\hat{\mathbf{B}}$ are $b_{ij} = \exp(-t_i\hat{\beta}_j)$, $1 \le i \le n$, $1 \le j \le \gamma/2$, and $\Gamma(\cdot)$ denotes the usual *Gamma* function.

The performances of MDL and nMDL criteria are compared in Section 3.2 for simulated data.

3. GENE CLUSTERING

3.1. New clustering algorithm

Assume that, applying the procedure described above, we have fitted a sum of exponentials to the time profile measured for a certain gene. Finding similarities between the expressions of this gene and another gene expressions reduces to a comparison between the two sets of the estimated parameters. At this point, a large family of comparison criteria can be considered. For example, we can first compare the estimated structure parameters, and if both model orders are the same we can further compare the gains and the time constants, respectively. Since generally a microarray data set contains measurements for thousands of genes, it is prohibitive to consider all possible pairs of genes for finding similarities.

We decide that two different genes share common regulation if the set of time constants is the same for both of them, and we cluster the genes together. Observe that the proposed similarity measure for genes ignores the gains. We do not know the true values of the time constants, and the estimated values are all different with probability one. We model the time constants estimated for all genes from a microarray data set as outcomes of a Gaussian mixture model, and cluster them in N_{TC} clusters with classification-expectation-maximization (CEM) algorithm [30]. The centroids found by CEM are denoted by T_i where index i takes values between 1 and N_{TC}. The centroids are increasingly ordered, namely, $T_i < T_j$ when $1 \le i < j \le N_{\mathrm{TC}}$. For clustering the time constants, we have pooled them together no matter the model order inferred for every gene.

We consider that, for a particular gene, the model order given by an information theoretic criterion like MDL or nMDL is p^*, and the estimated time constants are $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{p^*}$. We associate to this gene the sequence $T_{(1)} < T_{(2)} < \cdots < T_{(\pi)}$ determined by the centroids of the clusters to which the CEM algorithm assigns the time constants $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{p^*}$. If two or more time constants of the considered gene are assigned to the same cluster, then the corresponding centroid occurs only once in the sequence of centroids, and consequently $1 \le \pi \le p^*$. We cluster together two genes if the same sequence of centroids $T_{(1)} < T_{(2)} < \cdots < T_{(\pi)}$ is associated to both genes.

Therefore, we first cluster the time constants, and then we use the result to further cluster the genes. It is interesting to investigate the relationship between N_{TC}, the number of clusters for time constants, and N_{GC}, the number of gene clusters. Under the hypothesis that the information theoretic criterion selects the number of exponentials from the set $\{1, 2, \ldots, p_{\max}\}$, it is easy to prove that $1 \le N_{\mathrm{GC}} \le \sum_{i=1}^{p_{\max}} \binom{N_{\mathrm{TC}}}{i}$, where $\binom{N_{\mathrm{TC}}}{i} = 0$ for $i > N_{\mathrm{TC}}$. For the case $p_{\max} \ge N_{\mathrm{TC}}$, the inequality becomes $1 \le N_{\mathrm{GC}} < 2^{N_{\mathrm{TC}}}$. To illustrate the situation when $p_{\max} < N_{\mathrm{TC}}$, we choose $p_{\max} = 3$ and $N_{\mathrm{TC}} = 5$. For this selection, the number of gene clusters can potentially be as large as 25.

For completeness, we list in Algorithm 1 the newly introduced gene clustering algorithm.

3.2. Experimental results with simulated data

For validating the proposed method, we test it with carefully crafted data. Note that fitting sum of exponentials to the measured data is the crucial step of the procedure, in the sense that unreliable estimates for time constants can lead to false conclusions on the similarity of the genes. This is the reason for which we generate data according to a prototype that was introduced in [31] and used since then as a benchmark to evaluate the performances of various estimation algorithms. The model used in [31] to generate data is the same as the one in (2) with the following parameters: $p = 3$, $\alpha_1 = 0.6$, $\alpha_2 = 0.3$, $\alpha_3 = 0.1$, and $\beta_1 = 0.1$, $\beta_2 = 0.01$, $\beta_3 = 0.001$. Their proposed task was to estimate the parameters from 20 measurements nonuniformly sampled at time points between 0 and 6000, where no noise was added, but every measurement rounded to four significant digits. It is obvious that their goal is the same as in the estimation problem treated in Section 2, but the solution proposed by [31] applies only when all α_j's are strictly positive.

We extend this example by considering more linear combinations of the same exponential terms. To fix the ideas, in this section, we denote by G_i, $1 \le i \le 10$, a gene prototype, and not simply a gene. In Table 1 ten different gene prototypes are shown by indicating for each of them the values of p and α_j's. Note that for all prototypes the β_j's are the same as in the example from [31]. Beginning from a particular prototype G_i, we generate measurements for a gene by adding i.i.d. noise to the waveform given by G_i at the time points $0, 1, 2, 3, 4, 5, 10, 30, 60, 150, 300, 400$. We employ Gaussian noise with zero mean and variance σ^2. For G_1, we consider only the first 9 time points from the set of 12 time points listed above. The reason is that G_1 takes values very close to zero when time $t \ge 150$. Therefore, for genes generated according to prototype G_1, only 9 nonequidistant measurements are used in estimation, and for prototypes $G_2, \ldots, G_{10}$ the estimation of parameters is based on 12 nonequidistant measurements.

To test the capabilities of the method discussed in Section 2, we estimate the parameters of the model (3) from gene measurements simulated according to the previous scenario. For every prototype in Table 1, we generate measurements for 50 different genes, or equivalently, we consider 50 different noise realizations. The "true" time constants span a very large domain from 1 to 1000. This makes difficult the task of defining the domain in the grid-search algorithm. The reported estimation results are obtained when, for every time constant, the search domain is limited to $[1, 1200]$, and the search step is 20. The computational burden is decreased by assuming a priori that the set of estimated time constants for every gene is ordered. For improving the numerical conditioning of the algorithm, we force the difference between every two time constants to be larger than 20. It is clear that

```
Input: a data set containing gene expressions that are time series. It is not necessary
that the time sampling points are the same for all genes, and the number of samples
can vary from one gene to another.
1. For every gene,
        the available measurements are denoted by $y(t_1), \dots, y(t_n)$, where the time points
        $t_1, \dots t_n$ are generally nonequidistant.
        For $p = 1 : p_{\max}$,
                fit a sum of $p$ exponentials to the data $y(t_1), \dots, y(t_n)$ (Section 2.3);
                evaluate the nMDL criterion (10).
        End
        Choose $p^*$ to be the value of $p$ that minimizes the nMDL criterion. The estimated
        parameters are the gains $\hat{\alpha}_1, \dots, \hat{\alpha}_{p^*}$, and the time constants are $\hat{\tau}_1, \dots, \hat{\tau}_{p^*}$.
        If the goodness-of-fit criterion (11) is satisfied,
                include the time constants $\hat{\tau}_1, \dots, \hat{\tau}_{p^*}$ in the set $\mathcal{T}$.
        Else
                Label the current gene with zero.
        End
    End
2. After eliminating the outliers, group the objects from $\mathcal{T}$ into $N_{\mathrm{TC}}$ clusters by applying
the CEM algorithm (Section 3.1).
3. For each gene that is not labeled with zero, replace every time constant $\hat{\tau}_i$, $1 \le i \le p^*$,
with the centroid of the cluster to which $\hat{\tau}_i$ was assigned.
4. Group together into the same cluster all genes whose time constants are assigned to
the same set of centroids.
Output: each gene is labeled according to which cluster it belongs. For the genes
with label zero, the sum-of-exponentials model does not fit well, therefore they
are not included in any cluster.
```

Algorithm 1: Gene clustering algorithm.

Table 1: The parameters for the ten gene prototypes used in experiments with simulated data. For all gene prototypes, the model is given in (2), and $\beta_1 = 0.1$, $\beta_2 = 0.01$, $\beta_3 = 0.001$.

G	p	α_1	α_2	α_3
G_1	1	1.0	0.0	0.0
G_2	1	0.0	1.0	0.0
G_3	1	0.0	0.0	1.0
G_4	2	−0.5	1.5	0.0
G_5	2	0.0	0.6	0.4
G_6	2	0.6	0.0	0.4
G_7	2	−0.5	0.0	1.5
G_8	2	0.6	0.4	0.0
G_9	3	0.6	0.3	0.1
G_{10}	3	0.8	−0.6	0.8

the smaller the search step, the better the accuracy of initialization points for the optimization algorithm, but a small value for the step search means also a significant computational burden. In the analyzed case, a value of 20 is a good tradeoff between accuracy and complexity.

We run the grid-search algorithm and then the optimization algorithm, and report in Table 2 the results obtained when considering 50 trials for every gene prototype. The noise standard deviation is 10^{-3}. To have a better image on signal-to-noise ratio, remark that the values of each gene prototype varies between one for time moment zero, and asymptotic value zero. For the results in Table 2, the bias is small and the variance is close to CRLB. These estimations for coefficients α_j and time constants $\tau_j = 1/\beta_j$ are obtained by assuming that the true value of p is known. Using the same simulated data sets, we compare in Table 3 the estimations of the structural parameter obtained by MDL and nMDL criteria. Observe for $\sigma = 10^{-3}$ that nMDL estimates are 100% correct for seven out of ten prototypes, and the proportion of correct estimates is never smaller than 82%. At this level of noise, nMDL does not perform worse than MDL criterion in any of the cases. When the level of noise is increasing, the proportion of correct estimations declines for both MDL and nMDL, but overall we can conclude that nMDL is superior.

To complete the experiments, we have to cluster the estimated time constants, and for this task we use the Matlab programs which are available at http://www.cs.ucl.ac.uk/staff/D.Corney/ClusteringMatlab.html and http://www.ncrg.aston.ac.uk/netlab/. We try to mimic the real situations when more copies are available for the same microarray. We assume that the microarray contains measurements for 20 genes and 25 copies of it are available. More precisely, we randomly distribute the genes from the data sets already employed in the previous experiments such that to have 25 different copies of the same microarray, and every copy to contain exactly two genes from each prototype.

In the experiments, we apply two different procedures: (a) cluster the time constants estimated for the genes that belong to a microarray copy, and ignore the estimations obtained for the other microarray copies ("one clustering for

Table 2: Parameters, Cramér-Rao lower bounds, and estimation results for 50 trials when ten different gene prototypes are considered. Noise standard deviation is $\sigma = 10^{-3}$.

G	Parameter	α_1	α_2	α_3	τ_1	τ_2	τ_3
G_1	Actual	1.0	—	—	10	—	—
	Average	1.01	—	—	9.81	—	—
	Std. dev.	0.0389	—	—	0.76	—	—
	$\sqrt{\text{CRLB}}$	0.0007	—	—	0.02	—	—
G_2	Actual	—	1.0	—	—	100	—
	Average	—	1.00	—	—	100.52	—
	Std. dev.	—	0.0010	—	—	0.55	—
	$\sqrt{\text{CRLB}}$	—	0.0004	—	—	0.20	—
G_3	Actual	—	—	1.0	—	—	1000
	Average	—	—	1.00	—	—	1000.18
	Std. dev.	—	—	0.0003	—	—	2.93
	$\sqrt{\text{CRLB}}$	—	—	0.0003	—	—	2.99
G_4	Actual	−0.5	1.5	—	10	100	—
	Average	−0.50	1.50	—	10.01	100.00	—
	Std. dev.	0.0147	0.0160	—	0.45	0.91	—
	$\sqrt{\text{CRLB}}$	0.0026	0.0027	—	0.08	0.25	—
G_5	Actual	—	0.6	0.4	—	100	1000
	Average	—	0.60	0.40	—	100.64	1025.11
	Std. dev.	—	0.0119	0.0120	—	1.86	71.71
	$\sqrt{\text{CRLB}}$	—	0.0122	0.0124	—	1.89	72.85
G_6	Actual	0.6	—	0.4	10	—	1000
	Average	0.60	—	0.40	10.00	—	1001.49
	Std. dev.	0.0011	—	0.0010	0.05	—	11.42
	$\sqrt{\text{CRLB}}$	0.0011	—	0.0010	0.05	—	11.04
G_7	Actual	−0.5	—	1.5	10	—	1000
	Average	−0.50	—	1.50	9.99	—	1000.50
	Std. dev.	0.0011	—	0.0010	0.06	—	3.24
	$\sqrt{\text{CRLB}}$	0.0011	—	0.0010	0.06	—	2.95
G_8	Actual	0.6	0.4	—	10	100	—
	Average	0.60	0.40	—	9.92	99.56	—
	Std. dev.	0.0122	0.0169	—	0.59	3.66	—
	$\sqrt{\text{CRLB}}$	0.0026	0.0027	—	0.07	0.95	—
G_9	Actual	0.6	0.3	0.1	10	100	1000
	Average	0.60	0.30	0.11	9.97	97.78	960.23
	Std. dev.	0.0051	0.0106	0.0143	0.12	6.36	220.76
	$\sqrt{\text{CRLB}}$	0.0052	0.0174	0.0215	0.10	8.76	476.81
G_{10}	Actual	0.8	−0.6	0.8	10	100	1000
	Average	0.80	−0.60	0.80	10.01	100.01	1003.13
	Std. dev.	0.0052	0.0172	0.0218	0.07	4.54	58.79
	$\sqrt{\text{CRLB}}$	0.0052	0.0174	0.0215	0.07	4.38	59.60

each copy”); (b) cluster all time constants, estimated for all microarray copies (“one clustering for all copies”). For both procedures, we assume that the number of clusters is 3, and report the results in Tables 4, 5, 6, and 7. We mention that in our implementation, all estimated time constants that have values smaller than 1, or larger than 1200 are considered outliers. For clustering a set of time constants, we first eliminate the outliers, and then run the CEM algorithm starting from

Table 3: The percentage of estimations for model order when applying MDL and nMDL criteria. The value σ of noise standard deviation is given for every experiment. The symbol $*$ is used to indicate the "true" model order.

G		$\sigma = 10^{-3}$		$\sigma = 10^{-2}$		$\sigma = 10^{-1}$	
		MDL	nMDL	MDL	nMDL	MDL	nMDL
G_1	1*	**50**	**82**	20	**58**	**54**	**84**
	2	12	10	**46**	34	42	16
	3	38	8	34	8	4	0
G_2	1*	**84**	**100**	**90**	**100**	**82**	**100**
	2	10	0	8	0	10	0
	3	6	0	2	0	8	0
G_3	1*	**94**	**100**	**76**	**100**	**88**	**100**
	2	2	0	18	0	10	0
	3	4	0	6	0	2	0
G_4	1	0	0	0	0	32	**68**
	2*	**86**	**100**	**56**	**90**	**42**	30
	3	14	0	44	10	26	2
G_5	1	0	0	0	0	**58**	**94**
	2*	**94**	**100**	**84**	**100**	28	4
	3	6	0	16	0	14	2
G_6	1	0	0	0	0	0	2
	2*	**92**	**100**	**96**	**100**	**82**	**94**
	3	8	0	4	0	18	4
G_7	1	0	0	0	0	2	6
	2*	**90**	**98**	**86**	**98**	**86**	**92**
	3	10	2	14	2	12	2
G_8	1	0	0	0	0	38	**84**
	2*	**64**	**92**	**54**	**68**	**52**	16
	3	36	8	46	32	10	0
G_9	1	0	0	0	0	26	**74**
	2	0	0	12	36	**54**	24
	3*	**100**	**100**	**88**	**64**	20	2
G_{10}	1	0	0	0	0	0	12
	2	0	0	20	20	28	**46**
	3*	**100**	**100**	**80**	**80**	**72**	42

40 randomly chosen initialization points. Among the 40 resulting solutions, we select the partition that minimizes the sum of squared errors.

The results in Tables 4 and 5 show how well the estimated time constants have been allocated to clusters. Convention is that numbers represented with shades are counts of how many times the time constants are properly assigned to clusters. In our settings, for an ideal method, all counts represented with shades in Tables 4 and 5 are equal to 50, and all other counts in these tables take value zero. Now we can easily observe in Table 4 that the results yielded by the proposed method when clustering separately every copy, for $\sigma = 10^{-3}$, are very close to the best possible results. One single misclassification occurs for a gene generated according to prototype G_8. We investigate closely the estimations for G_8 when $\sigma = 10^{-3}$: percentages in Table 3 indicate that nMDL estimates correctly the number of clusters for 92% of genes, and overestimates the order for the rest of 8% genes. As we have generated 50 different realizations for G_8, it means that the model order was estimated to be three for four genes. Based on this observation, we could expect that the row corresponding to G_8 in Table 4 contains value 50 (represented with shades) for the first two counters, and value 4 for the third counter. The value of the last counter is 1, which can be explained as follows: in the case of three G_8 genes for which the order was overestimated, the largest time constant was grouped by CEM together with the second largest time constant. The discussion on this particular example gives hints for the interpretation of the data in Tables 4 and 5, and emphasizes the importance of using an accurate estimator for the structure parameter. When comparing the content of the two tables, we note again the superiority of the nMDL criterion. We observe also that performing "one clustering for all copies" does not improve the grouping of genes.

Table 4: Results obtained when clustering the time constants estimated for 25 microarray copies when each copy contains the measurements of exactly 2 genes from every prototype $G_1 - G_{10}$. The values represented with shades are counts for the time constants that are properly assigned to clusters, and the numbers represented without shades count for misclassified time constants. The structure parameter is estimated by applying the nMDL criterion.

$\sigma = 10^{-3}$

G	One clustering for each copy			One clustering for all copies		
	$\#T_1$	$\#T_2$	$\#T_3$	$\#T_1$	$\#T_2$	$\#T_3$
G_1	50	0	0	49	7	0
G_2	0	50	0	0	50	0
G_3	0	0	50	0	0	50
G_4	50	50	0	50	50	0
G_5	0	50	50	0	50	50
G_6	50	0	50	50	0	50
G_7	50	0	50	50	1	50
G_8	50	50	1	50	50	1
G_9	50	50	50	50	50	50
G_{10}	50	50	50	50	50	50

$\sigma = 10^{-2}$

G	One clustering for each copy			One clustering for all copies		
	$\#T_1$	$\#T_2$	$\#T_3$	$\#T_1$	$\#T_2$	$\#T_3$
G_1	50	8	7	50	11	8
G_2	2	48	0	0	50	0
G_3	0	0	50	0	0	50
G_4	42	48	2	38	50	2
G_5	2	50	45	0	50	50
G_6	50	0	50	50	0	50
G_7	50	0	50	50	1	50
G_8	46	48	4	44	50	5
G_9	47	49	18	49	33	50
G_{10}	47	43	48	50	40	50

$\sigma = 10^{-1}$

G	One clustering for each copy			One clustering for all copies		
	$\#T_1$	$\#T_2$	$\#T_3$	$\#T_1$	$\#T_2$	$\#T_3$
G_1	50	2	0	50	1	0
G_2	8	42	0	0	50	0
G_3	0	10	40	0	33	17
G_4	18	42	0	16	49	0
G_5	4	46	3	3	48	2
G_6	49	14	36	49	24	26
G_7	47	10	40	45	31	19
G_8	45	10	2	48	6	1
G_9	42	12	8	47	9	7
G_{10}	43	15	42	44	23	35

Table 5: Counting the well-classified and misclassified time constants when the data sets are the same as for the results reported in Table 4, and the structure parameter is estimated with MDL criterion. The convention for using shades is the same as in Table 4.

$\sigma = 10^{-3}$

G	One clustering for each copy			One clustering for all copies		
	$\#T_1$	$\#T_2$	$\#T_3$	$\#T_1$	$\#T_2$	$\#T_3$
G_1	50	3	0	40	25	0
G_2	6	50	0	0	50	0
G_3	2	2	50	0	3	50
G_4	50	50	0	48	50	0
G_5	2	50	50	0	50	50
G_6	50	1	50	50	4	50
G_7	50	2	50	50	5	50
G_8	50	50	2	47	50	2
G_9	50	50	50	50	50	50
G_{10}	50	50	50	50	50	50

$\sigma = 10^{-2}$

G	One clustering for each copy			One clustering for all copies		
	$\#T_1$	$\#T_2$	$\#T_3$	$\#T_1$	$\#T_2$	$\#T_3$
G_1	49	14	12	49	28	13
G_2	4	48	2	0	50	2
G_3	8	4	50	4	10	50
G_4	43	48	5	36	50	5
G_5	7	50	45	0	50	50
G_6	50	2	50	50	2	50
G_7	50	2	50	48	7	50
G_8	47	47	6	41	50	7
G_9	49	49	24	50	45	49
G_{10}	48	40	50	42	48	50

$\sigma = 10^{-1}$

G	One clustering for each copy			One clustering for all copies		
	$\#T_1$	$\#T_2$	$\#T_3$	$\#T_1$	$\#T_2$	$\#T_3$
G_1	49	6	5	50	5	5
G_2	17	41	0	8	50	0
G_3	6	17	33	6	33	17
G_4	33	39	4	34	47	4
G_5	17	45	9	17	48	5
G_6	50	21	29	50	26	24
G_7	49	18	34	47	32	19
G_8	49	18	10	49	19	9
G_9	49	16	16	49	19	16
G_{10}	49	25	38	50	31	33

Table 6: The centroids for clusters of estimated time constants. For the scenario "one clustering for each copy," the mean and standard deviation are reported for every centroid. The structure parameter is estimated with the nMDL criterion.

$\sigma = 10^{-3}$			
T	One clustering for each copy		One clustering for all copies
	Average	Std. dev.	
$T1$	10.20	0.45	10.00
$T2$	99.33	1.47	96.14
$T3$	999.09	27.99	999.13
$\sigma = 10^{-2}$			
T	One clustering for each copy		One clustering for all copies
	Average	Std. dev.	
$T1$	11.49	6.76	9.78
$T2$	122.95	94.04	81.01
$T3$	1001.73	72.74	912.52
$\sigma = 10^{-1}$			
T	One clustering for each copy		One clustering for all copies
	Average	Std. dev.	
$T1$	21.07	12.37	18.03
$T2$	259.17	175.06	389.54
$T3$	1094.89	94.35	1192.57

Table 7: The centroids for clusters of estimated time constants. For the scenario "one clustering for each copy", the mean and standard deviation are reported for every centroid. The structure parameter is estimated with the MDL criterion.

$\sigma = 10^{-3}$			
T	One clustering for each copy		One clustering for all copies
	Average	Std. dev.	
$T1$	11.71	1.42	10.01
$T2$	96.77	4.72	80.40
$T3$	998.62	30.02	998.78
$\sigma = 10^{-2}$			
T	One clustering for each copy		One clustering for all copies
	Average	Std. dev.	
$T1$	11.93	6.79	9.74
$T2$	109.82	62.52	69.30
$T3$	998.60	93.71	929.22
$\sigma = 10^{-1}$			
T	One clustering for each copy		One clustering for all copies
	Average	Std. dev.	
$T1$	20.17	8.78	17.63
$T2$	302.08	140.30	374.81
$T3$	1140.30	75.26	1191.00

The data have been generated such that the "true" time constants for every gene, in all microarray copies, belong to the set $\{10, 100, 1000\}$. We compare next the values $10, 100, 1000$, with the centroids found by clustering the estimated time constants. In Tables 6 and 7 are shown the centroids obtained when applying the scenario "one clustering for all copies". Since "one clustering for each copy" leads to 25 different estimations for every centroid, we report in Tables 6 and 7 the computed mean and standard deviation. Remark in the case when the structure parameter is estimated with nMDL, and noise standard deviation is $\sigma = 10^{-3}$, that the centroids are close to the "true" values. When $\sigma = 10^{-1}$, the centroids corresponding to 10 and 100 take values larger than expected.

The clustering results in Tables 4, 5, 6, and 7 are a good measure of the accuracy for the proposed method. Encouraged by these results, we apply next the clustering algorithm for data sets from developmental biology.

4. CLUSTERING THE GENE EXPRESSION DATA SAMPLED DURING POSTNATAL DEVELOPMENT OF MOUSE DENTATE GYRUS AND CEREBELLUM

4.1. Data sets from developmental biology

We apply the newly introduced clustering method for measurements obtained in experimental studies from developmental biology. The data are available at http://physiolgenomics.physiology.org/cgi/content/full/8/2/131/DC1/2, and represent expressions of 1412 genes measured at the same time points in two different experiments. The first experiment [1] is focused on studying the postnatal development of mouse cerebellar cortex, and the second one [3] analyzes the postnatal development of the dentate gyrus of mouse hippocampus. Since the cerebellar cortex and the dentate gyrus have common features, in [1], comparisons are performed between the time profiles obtained in both experiments.

The measurements are sampled at six time points, namely, 2, 4, 8, 12, 21, and 42 days after birth. In [3], the comparison of gene expressions relies in Euclidean distance. A statistical analysis is also conducted: first the genes are grouped by using Ward's hierarchical clustering method [32], and then the enrichment of functional categories in each cluster is investigated. As functional class labels have been associated with most of the genes, there is a significant interest on finding a correspondence between time profile of gene expressions, and the functional role played by each gene. Once the clustering is performed, it remains to decide if a particular functional category $\mathcal{F}$ appears unusually often within a particular cluster $\mathcal{C}$.

4.2. Fitting the sum-of-exponentials model

First we apply the algorithm described in Section 2 for fitting sum of exponentials to the 1412 gene profiles measured during postnatal development of mouse cerebellum [1, 3]. The optimal number of exponentials in each sum is selected from the set $p \in \{1, 2, 3\}$ by applying the nMDL criterion. For every time constant, the grid-search domain is taken as $[1, 35]$, and the search step is 0.3.

4.2.1. Goodness-of-fit testing

Our first aim is to test whether the developmental biology data fit well the sum-of-exponentials model. Among the rich family of testing methods, we choose a criterion that is intuitive and very simple to implement. For each gene, we decide that the estimation is reliable only when the "signal-to-noise ratio" is high enough, or equivalently, when

$$\frac{\mathrm{RSS}\,(\gamma^*)}{\sum_{i=1}^{n} z(t_i)^2} < \mathrm{Th}_{\mathrm{RSS}}, \tag{11}$$

where the notations are like in (5). The threshold $\mathrm{Th}_{\mathrm{RSS}}$ is chosen in the interval $(0, 0.5)$ based on a procedure described in the sequel. To investigate the robustness of this criterion, we compare it with another one that exploits in a different way the information in the errors observed when fitting the sum-of-exponentials model to gene expression data. To illustrate this second criterion we use the data set measured during postnatal development of mouse cerebellum. For an arbitrary gene in the analyzed data set, we define for the jth sample the relative error $\varepsilon_j^r = |z(t_j) - y(t_j|\hat{\theta}_\gamma)|/[\sum_{i=1}^{n}(z(t_i) - y(t_i|\hat{\theta}_\gamma))^2]^{1/2}$, where we use the same notations as in equation (5). Remark from the definition that positive and negative errors are mapped together. We collect all errors computed with the expression above for this data set, and further group them based on their magnitudes. The errors are assumed to be outcomes from a Gaussian mixture with two components: the first group is denoted $\mathcal{R}_g$ and contains the residuals with small magnitudes that correspond to the case of good fit between the measured data and the sum-of-exponentials model. The rest of the residuals are assigned to the group conventionally denoted by $\mathcal{R}_b$. Since we label each error with $\mathcal{R}_g$ or $\mathcal{R}_b$, it follows that the larger the number of $\mathcal{R}_g$ labels for a gene record, the higher the quality of fit for the gene. To fix the ideas we note that the data set we study contains records for 1412 genes and six measurements are available for each gene. For simplicity we ignore one gene for which all six measurements take the same value. Therefore the total number of errors is 8466. After applying the CEM algorithm initialized from 20 different points and selecting the best solution, the errors are split in the groups $\mathcal{R}_g$ and $\mathcal{R}_b$. We note that $\mathcal{R}_g$ contains 6164 errors for which the mean is 0.054 and the variance is 0.003. The rest of 2302 errors are grouped in $\mathcal{R}_b$, their mean is 0.390 and the variance is 0.045.

For $m \in \{1, \ldots, 6\}$, we assign a gene to the set denoted $\mathcal{R}_g^m$ if at least m of its error labels are $\mathcal{R}_g$. For each m, the genes not assigned to $\mathcal{R}_g^m$ are included in $\mathcal{R}_b^m$. This leads to a goodness-of-fit criterion: for a given value of m, accept that the sum-of-exponentials model fits well a gene if the gene belongs to $\mathcal{R}_g^m$. For example, when choosing $m = 6$, the criterion is very selective in the sense that requires all errors corresponding to the analyzed gene to be small in magnitude. Since we are interested only in those genes for which the number of small magnitude errors exceeds the number of large magnitude errors, we restrict the analysis to $m \in \{4, 5, 6\}$, and note that the cardinalities of the selected sets are $|\mathcal{R}_g^6| = 378$, $|\mathcal{R}_g^5| = 654$, $|\mathcal{R}_g^4| = 1036$.

Table 8: The goodness-of-fit of sum-of-exponentials model is evaluated with two different criteria for postnatal development of mouse cerebellum gene expression data. The first criterion compares with a threshold $\mathrm{Th}_{\mathrm{RSS}}$ the ratio between the sum of squared errors and the sum of squared measurements, for each gene. For the second criterion, the small magnitude errors are collected in $\mathcal{R}_g$, and the rest of errors are included in $\mathcal{R}_b$. Then $\mathcal{R}_g^m$, $1 \le m \le 6$, is the set of all genes for which at least m errors belong to $\mathcal{R}_g$. For each m, the complementary set of $\mathcal{R}_g^m$ is $\mathcal{R}_b^m$. The contingency tables of the gene partitions determined by the two criteria are shown for different values of $\mathrm{Th}_{\mathrm{RSS}}$ and for $m \in \{4, 5, 6\}$.

$\mathrm{Th}_{\mathrm{RSS}}$	Criterion	$\mathcal{R}_g^6$	$\mathcal{R}_b^6$	$\mathcal{R}_g^5$	$\mathcal{R}_b^5$	$\mathcal{R}_g^4$	$\mathcal{R}_b^4$
0.01	$\le \mathrm{Th}_{\mathrm{RSS}}$	**132**	0	132	0	132	**0**
	$> \mathrm{Th}_{\mathrm{RSS}}$	246	1033	522	757	904	375
0.05	$\le \mathrm{Th}_{\mathrm{RSS}}$	**320**	17	337	0	337	**0**
	$> \mathrm{Th}_{\mathrm{RSS}}$	58	1016	317	757	699	375
0.10	$\le \mathrm{Th}_{\mathrm{RSS}}$	**378**	111	478	11	489	**0**
	$> \mathrm{Th}_{\mathrm{RSS}}$	0	922	176	746	547	375
0.15	$\le \mathrm{Th}_{\mathrm{RSS}}$	**378**	246	543	81	622	**2**
	$> \mathrm{Th}_{\mathrm{RSS}}$	0	787	111	676	414	373
0.20	$\le \mathrm{Th}_{\mathrm{RSS}}$	**378**	318	557	139	685	**11**
	$> \mathrm{Th}_{\mathrm{RSS}}$	0	715	97	618	351	364

We can admit with high degree of confidence that a particular gene is properly selected with the criterion (11) if the gene is also included in the set $\mathcal{R}_g^6$. In general, choosing a particular value for $\mathrm{Th}_{\mathrm{RSS}}$ leads to a partition of the genes into two different groups. Similarly, selecting the value of m leads to another two-groups partition of the gene set. The contingency tables, shown in Table 8, are very convenient for comparing the partitions obtained with various values of $\mathrm{Th}_{\mathrm{RSS}}$ and m. Recall that $\mathcal{R}_g^6$ set contains all genes for which the magnitudes of all errors are small. According to the results in Table 8, choosing $\mathrm{Th}_{\mathrm{RSS}}$ to be 0.01 or 0.05 implies that some genes from $\mathcal{R}_g^6$ are considered to have poor fit with sum-of-exponentials model. For $\mathrm{Th}_{\mathrm{RSS}} = 0.05$, 58 genes from $\mathcal{R}_g^6$ are deemed as poor fit for sum-of-exponentials model, while 17 genes from $\mathcal{R}_b^6$ are assumed to fit well the model. When the decision is based on a threshold value $\mathrm{Th}_{\mathrm{RSS}} \ge 0.1$, all 378 genes in the set $\mathcal{R}_g^6$ are qualified as well fitted by the model. Remark from the last column in Table 8 that for $\mathrm{Th}_{\mathrm{RSS}} \le 0.1$ none of the genes with less than four errors in $\mathcal{R}_g$ are selected by the criterion (11). These properties recommend to choose $\mathrm{Th}_{\mathrm{RSS}} = 0.1$. To illustrate the accuracy of modelling gene expressions with sum of exponentials, we plot in Figure 1 the original measurements and the optimal model for two genes.

Based on criterion (11) with $\mathrm{Th}_{\mathrm{RSS}} = 0.1$, we select 489 out of 1412 genes of cerebellum data set, and further cluster

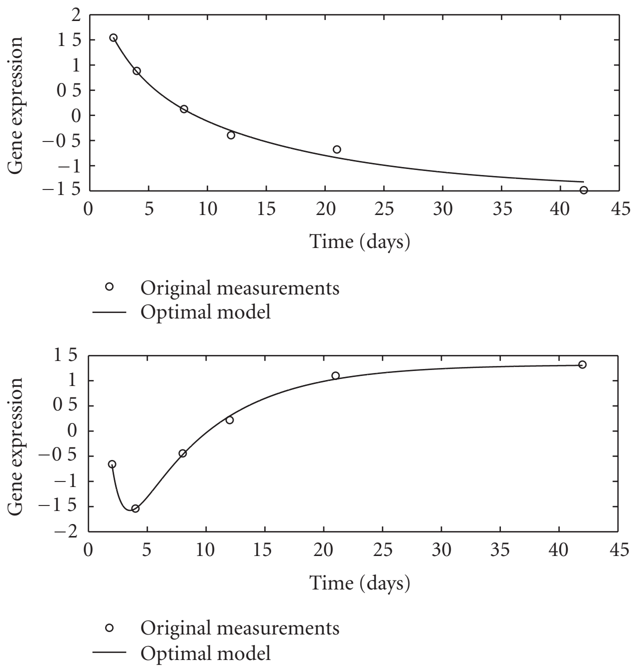

Figure 1: Two genes and their associated models. (a) Time series for a gene measured during postnatal development of mouse cerebellar cortex. The gene has index 297 in http://physiolgenomics.physiology.org/cgi/content/full/8/2/131/DC1/2. The optimal number of exponentials is $p^* = 2$, the gains are $[\hat{\alpha}_1\ \hat{\alpha}_2] = [2.55\ 1.63]$, the time constants are $[\hat{\tau}_1\ \hat{\tau}_2] = [15.20\ 2.80]$, and $\text{RSS}(p^*)/\sum_{i=1}^{6} z(t_i)^2 = 0.010$. (b) Time series of the gene having index 1303 in the URL stated above, measured during the postnatal development of dentate gyrus of mouse hypocampus: the estimated parameters are $p^* = 2$, $[\hat{\alpha}_1\ \hat{\alpha}_2] = [-5.55\ 16.30]$, $[\hat{\tau}_1\ \hat{\tau}_2] = [7.10\ 1.00]$, and $\text{RSS}(p^*)/\sum_{i=1}^{6} z(t_i)^2 = 0.002$.

them. We apply the same procedure for the gene expressions measured during postnatal development of mouse dentate gyrus, and the number of genes for which the model with sum of exponentials fits well is 561 out of 1412.

4.3. Clustering the selected genes

We use CEM to group into $N_{\text{TC}} = 3$ clusters the time constants associated with the 489 genes selected from cerebellum data set. After running the algorithm from 40 different initialization points, the resulting centroids are $T_1 = 1.88$, $T_2 = 6.16$, $T_3 = 15.41$. We note that the time constants are clustered after eliminating the outliers which are the values smaller than 1, or larger than 35. Based on the method described in Section 3, we use the clusters already found for time constants to group the genes, and the resulting number of gene clusters is $N_{\text{GC}} = 6$. Recall that two genes are pooled together in the same cluster when both sets of time constants are associated to the same sequence of centroids. We assign to the gene clusters labels from C_1 to C_6, and list for every cluster the corresponding sequence of centroids: $C_1 : \{T_1\}$, $C_2 : \{T_2\}$, $C_3 : \{T_3\}$, $C_4 : \{T_1, T_2\}$, $C_5 : \{T_1, T_3\}$, $C_6 : \{T_2, T_3\}$. Observe that there is no gene for which the set of time constants contains representatives from all clusters whose centroids are T_1, T_2, T_3.

Table 9: Enrichment of functions in clusters of gene expressions measured during mouse cerebellar development. For each cluster $\mathcal{C}$ and for each functional category $\mathcal{F}$, the number of genes $\nu(\mathcal{C}, \mathcal{F})$ that belong both to $\mathcal{C}$ and $\mathcal{F}$ is given. The enrichment is marked with shades. Statistical tests are based on hypergeometric distribution. The following acronyms are used for the functional categories: GF: growth factors and their receptors, IST: intracellular signal transduction (except kinases), DEV: development, PSOM: proteosome, T: transcription factors, C: carbohydrate metabolism, CY: cytoskeleton, STK: serine/threonine kinase, SY: synaptic component, GR: cell growth-related, B: brain and neuron, PS: ribosomal proteins, GANN: oncogenes and their relates.

Func. categ.	C_1	C_2	C_3	C_4	C_5	C_6	# genes
GF	4	0	1	0	0	0	11
IST	8	1	2	2	2	1	47
DEV	1	5	2	1	0	0	22
PSOM	1	4	0	1	0	0	17
T	3	7	3	0	0	0	37
C	0	3	0	2	0	0	14
CY	7	8	0	0	0	0	48
STK	0	1	3	2	0	0	19
SY	1	0	2	0	1	1	15
GR	2	2	2	0	0	0	15
B	5	15	9	8	3	1	116
PS	8	2	2	16	1	0	50
GANN	2	0	1	2	0	0	11
Total	141	143	67	105	28	5	1412

Then we focus on the time constants corresponding to the 561 genes selected from mouse dentate gyrus data set. After dropping the outliers, the time constants are grouped by CEM in three clusters whose centroids are $T_1 = 1.95$, $T_2 = 5.58$, and $T_3 = 19.38$. Remark that the centroids are close to those determined for the cerebellum data. Moreover, the number of gene clusters is also six, and the sequence of centroids corresponding to every gene cluster is the same as for the cerebellum data.

4.4. Enrichment of functional categories

We briefly revisit some statistical aspects regarding the enrichment of functional categories in experiments with microarray data [33]. To fix the ideas, we assume that the total number of genes on the microarray is M, and only a proportion q of them belongs to functional category $\mathcal{F}$. Therefore, qM genes are in category $\mathcal{F}$, and $(1-q)M$ genes are not in this category, where $0 \le q \le 1$. Suppose that the number of genes in cluster $\mathcal{C}$ is m. We randomly choose m out of M genes, and denote by V the random variable which counts how many genes from $\mathcal{F}$ are among the m selected genes. The probability function of V is modelled by the hypergeometric distribution $H(M, m; q)$ [34]. If $\nu(\mathcal{C}, \mathcal{F})$ denotes the number of genes that belong both to cluster $\mathcal{C}$ and to category $\mathcal{F}$, then $\mathcal{F}$ is enriched in $\mathcal{C}$ when $\nu(\mathcal{C}, \mathcal{F})$ is such that

Table 10: Enrichment of functions in clusters of gene expressions measured during mouse cerebellar development. Statistical tests for enrichment are based on binomial distribution. The acronyms for the functional categories are the same as in Table 9.

Func. categ.	C_1	C_2	C_3	C_4	C_5	C_6	# genes
GF	4	0	1	0	0	0	11
IST	8	1	2	2	2	1	47
DEV	1	5	2	1	0	0	22
PSOM	1	4	0	1	0	0	17
T	3	7	3	0	0	0	37
STK	0	1	3	2	0	0	19
SY	1	0	2	0	1	1	15
GR	2	2	2	0	0	0	15
B	5	15	9	8	3	1	116
PS	8	2	2	16	1	0	50
GANN	2	0	1	2	0	0	11
Total	141	143	67	105	28	5	1412

Table 11: Enrichment of functions in clusters of gene expressions measured during mouse dentate gyrus development. For each cluster $\mathcal{C}$ and for each functional category $\mathcal{F}$, the number of genes $\nu(\mathcal{C},\mathcal{F})$ that belong both to $\mathcal{C}$ and $\mathcal{F}$ is given. The enrichment is marked with shades. Statistical tests are based on hypergeometric distribution. The following acronyms are used for the functional categories: STK: serine/threonine kinase, TES: testis, SY: synaptic component, CR: chromosome component, E: electron transfer, SEC: secretory pathway, B: brain and neuron, L: lipid metabolism, GANN: oncogenes and their relates, PS: ribosomal proteins, A: amino acid metabolism, CHAP: chaperonines.

Func. categ.	C_1	C_2	C_3	C_4	C_5	C_6	# genes
STK	7	1	2	1	2	0	19
TES	5	1	0	4	0	0	16
SY	4	1	1	2	1	1	15
CR	4	1	0	1	0	0	16
E	1	6	1	3	0	0	26
SEC	2	1	2	0	0	0	10
B	16	15	13	9	2	1	116
L	1	3	3	1	0	0	19
GANN	0	2	0	3	0	0	11
PS	7	5	2	7	0	0	50
A	2	0	1	1	3	0	17
CHAP	1	1	2	0	2	0	17
Total	165	152	100	100	36	8	1412

Table 12: Enrichment of functions in clusters of gene expressions measured during mouse dentate gyrus development. Statistical tests for enrichment are based on binomial distribution. The acronyms for the functional categories are the same as in Table 11.

Func. categ.	C_1	C_2	C_3	C_4	C_5	C_6	# genes
STK	7	1	2	1	2	0	19
TES	5	1	0	4	0	0	16
SY	4	1	1	2	1	1	15
CR	4	1	0	1	0	0	16
E	1	6	1	3	0	0	26
B	16	15	13	9	2	1	116
SEC	2	1	2	0	0	0	10
L	1	3	3	1	0	0	19
GANN	0	2	0	3	0	0	11
PS	7	5	2	7	0	0	50
A	2	0	1	1	3	0	17
CHAP	1	1	2	0	2	0	17
Total	165	152	100	100	36	8	1412

$\text{Prob}\{V > \nu(\mathcal{C},\mathcal{F})\} < 0.05$, where

$$\text{Prob}\{V > \nu(\mathcal{C},\mathcal{F})\} = 1 - \sum_{i=0}^{\nu(\mathcal{C},\mathcal{F})} \frac{\binom{qM}{i}\binom{(1-q)M}{m-i}}{\binom{M}{m}}. \quad (12)$$

Since M takes large values for microarray data, numerical difficulties occur when computing the expression above. Relying on the theoretical result [34], which claims that the limit of the hypergeometric distribution $H(M,m;q)$ as $M \to \infty$ is the binomial distribution $\text{Bi}(m,q)$, the probability function of V is modelled by $\text{Bi}(m,q)$, and (12) will be replaced by

$$\text{Prob}\{V > \nu(\mathcal{C},\mathcal{F})\} = 1 - \sum_{i=0}^{\nu(\mathcal{C},\mathcal{F})} \binom{m}{i} q^i (1-q)^{m-i}. \quad (13)$$

We use next both (12) and (13) to investigate the enrichment of functional categories in the examples from the developmental biology. We report in Tables 9, 10, 11, and 12 the results obtained when testing the enrichment of functional categories in clusters found as described above. The following supplementary conditions are added to the tests in (12) and (13): to decide that the functional category $\mathcal{F}$ is enriched in cluster $\mathcal{C}$, it is necessary that at least 10 genes on the microarray belong to $\mathcal{F}$, and at least 2 genes from $\mathcal{C}$ belong to $\mathcal{F}$. To refer to various functions, we use in Tables 9, 10, 11, and 12 the acronyms defined at http://physiolgenomics.physiology.org/cgi/content/full/4/2/155/DC1/2.

Comparing the results in Tables 9 and 10, we remark that almost the same functions are found to be enriched by statistical tests (12) and (13). Applying the binomial distribution model seems to be more restrictive in the sense that the functions C and CY are reported as enriched in Table 9, but not in Table 10.

According to the results shown in Tables 11 and 12, some functions are enriched in the clusters of gene expressions for dentate gyrus data. Observe for this data set that exactly the same functions are found to be enriched when the test is performed with (12), and with (13). Moreover, there exist some functional categories enriched both in clusters of cerebellum data and in clusters of dentate gyrus data: STK, SY, B, PS, and GANN.

The very last observation has a special importance in comparison with the results reported in [1, 3] where enriched functions have been found only for the clusters in cerebellum data. Our clustering procedure allows to find enriched functions for both cerebellum data and dentate gyrus data, and some of these functions are the same for the two data sets. It is widely accepted in biology that cerebellar cortex and the dentate gyrus have common features [3], therefore the findings of the newly introduced algorithm can be explained based on biological knowledge. The biological significance of our results still remains to be further investigated in the future.

5. CONCLUSION

In this paper, we propose a new approach for clustering the gene expression data that are time series. The key step of the algorithm consists in fitting a sum of exponentials to the nonuniformly sampled points of every time series. The optimal number of exponentials is inferred relying on a new information theoretic criterion which is defined based on NML estimator [28, 29]. The estimation method is tested with carefully crafted data, and the results are compared with Cramér-Rao lower bound. The conclusions drawn for simulated data allow to define a criterion that determines when the sum of exponentials model fits well the measured data. The clustering procedure is applied for data sets from developmental biology [1, 3], and the enrichment of functional categories is investigated.

APPENDICES

A. THE SEPARABILITY OF PARAMETERS FOR THE LS PROBLEM

Nonlinear LS problems are in general difficult to solve, and it is of interest to reduce the problem to one of a smaller dimensionality if the optimization can be analytically performed over a subset of variables, for fixed values of the remaining variables. This can be readily achieved for model (2) by restating the LS problem as follows.

Minimize $\|\boldsymbol{\varepsilon}\|^2$, where $\boldsymbol{\varepsilon}$ is the error vector $\boldsymbol{\varepsilon} = \mathbf{z} - \mathbf{B}(\mathbf{b})\mathbf{a}$ and where $\mathbf{b}$ is the vector $[\beta_1, \dots, \beta_p]^\top$, $\mathbf{a}$ is the vector $[\alpha_1, \dots, \alpha_p]^\top$, and $\mathbf{B}$ is the matrix with entries $b_{ij} = \exp(-t_i\beta_j)$, for $1 \le i \le n$ and $1 \le j \le p$. Recall that the vector of measurements is $\mathbf{z} = [z(t_1) \cdots z(t_n)]^\top$. Assume that $n \ge p$, which is the case of interest where the number of exponentials does not exceed the number of measurements.

For a given $\mathbf{b}$, we denote by $\hat{\mathbf{a}}(\mathbf{b})$ the vector that minimizes $\|\boldsymbol{\varepsilon}\|^2$, which results to be the linear LS solution $\hat{\mathbf{a}}(\mathbf{b}) = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{z}$, if the matrix $\mathbf{B}^\top\mathbf{B}$ is nonsingular. The matrix $\mathbf{B}$ has all columns independent, since every $p \times p$ minor is a generalized Vandermonde matrix with nonzero determinant. Thus, the matrix $\mathbf{B}^\top\mathbf{B}$ is positive definite and nonsingular, and $\hat{\mathbf{a}}(\mathbf{b}) = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{z}$ is the unique minimizer of $\|\boldsymbol{\varepsilon}\|^2$ for a given $\mathbf{b}$ vector. Evaluating $\|\boldsymbol{\varepsilon}\|^2$ at the value $\hat{\mathbf{a}}(\mathbf{b})$, we find $\|\boldsymbol{\varepsilon}\|^2 = \mathbf{z}^\top(\mathbf{I} - \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top)\mathbf{z}$, which is now a nonlinear criterion in the reduced set of parameters $\mathbf{b}$, much easier to solve when compared to the original problem having as parameters the entries of the vectors $\mathbf{a}$ and $\mathbf{b}$.

B. COMPUTATION OF CRAMÉR-RAO LOWER BOUND (CRLB)

Since the samples are statistically independent, we obtain the following expression of the log-likelihood function, under the hypothesis of white Gaussian noise with zero mean and variance σ^2:

$$\Lambda(\mathbf{z}|\theta'_\gamma) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[z(t_i) - y(t_i|\theta'_\gamma)\right]^2, \quad \text{(B.1)}$$

where $\mathbf{z}$ is the vector of measurements, and the set of parameters θ'_γ is given in (6). In this section, we drop the index γ, and write θ' instead of θ'_γ. The expression of the Fisher information matrix is

$$\begin{bmatrix} \mathbf{J} & \mathbf{0} \\ \mathbf{0} & \dfrac{n}{2\sigma^4} \end{bmatrix}. \quad \text{(B.2)}$$

We focus on the entries of the block J:

$$J_{\ell m} = -E\left[\frac{\partial^2\Lambda(\mathbf{z}|\theta')}{\partial\theta'_\ell\partial\theta'_m}\right] = E\left[\frac{\partial\Lambda(\mathbf{z}|\theta')}{\partial\theta'_\ell}\frac{\partial\Lambda(\mathbf{z}|\theta')}{\partial\theta'_m}\right], \quad \text{(B.3)}$$

where the derivatives are evaluated at the true value of θ', and the expectation is taken with respect to the probability density function of $\mathbf{z}$ conditional to the model parameters [23]. θ'_ℓ and θ'_m are the ℓ'th and the m'th entries of θ', where $1 \le \ell, m \le 2p$. A well-known result on CRLB for general Gaussian case [23] implies

$$J_{\ell m} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{\partial y(t_i|\theta')}{\partial\theta'_\ell}\frac{\partial y(t_i|\theta')}{\partial\theta'_m}. \quad \text{(B.4)}$$

From $y(t_i|\theta') = \sum_{j=1}^{p}\alpha_j\delta_j^{t_i}$, or equivalently, $y(t_i|\theta') = \sum_{j=1}^{p}\theta'_j(\theta'_{j+p})^{t_i}$, we readily obtain

$$\frac{\partial y(t_i|\theta')}{\partial\theta'_\ell} = \begin{cases} (\theta'_{\ell+p})^{t_i}, & 1 \le \ell \le p, \\ \theta'_{\ell-p}t_i(\theta'_\ell)^{t_i-1}, & p+1 \le \ell \le 2p, \end{cases} \quad \text{(B.5)}$$

which leads to

$$J_{\ell m} = \begin{cases} \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(\theta'_{\ell+p})^{t_i}(\theta'_{m+p})^{t_i}, \\ \qquad\qquad 1 \le \ell, m \le p, \\ \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}t_i^2\theta'_{\ell-p}(\theta'_\ell)^{t_i-1}\theta'_{m-p}(\theta'_m)^{t_i-1}, \\ \qquad\qquad p+1 \le \ell, m \le 2p, \\ \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(\theta'_{\ell+p})^{t_i}\theta'_{m-p}t_i(\theta'_m)^{t_i-1}, \\ \qquad\qquad 1 \le \ell \le p, p+1 \le m \le 2p, \\ \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}\theta'_{\ell-p}t_i(\theta'_\ell)^{t_i-1}(\theta'_{m+p})^{t_i}, \\ \qquad\qquad p+1 \le \ell \le 2p, 1 \le m \le p. \end{cases} \quad \text{(B.6)}$$

When the set of parameters for the signal model is θ''_γ (7), let **G** be the block of the Fisher information matrix similar to **J** . For simplicity, we use the notation θ'' instead of θ''_γ. Based on the result from [23] on vector parameter transformations, the relationship between the entries of the matrices **G** and **J** is given by

$$G_{\ell m} = J_{\ell m}\left(\frac{\partial\theta'_\ell}{\partial\theta''_\ell}\right)\left(\frac{\partial\theta'_m}{\partial\theta''_m}\right), \qquad \text{(B.7)}$$

where $1 \leq \ell, m \leq 2p$. For $1 \leq \ell \leq p$, we have $\partial\theta'_\ell/\partial\theta''_\ell = 1$, while for $p+1 \leq \ell \leq 2p$,

$$\frac{\partial\theta'_\ell}{\partial\theta''_\ell} = \frac{\partial\delta_\ell}{\partial\tau_\ell} = \exp\left(-\frac{T_0}{\tau_\ell}\right)\frac{T_0}{\tau_\ell^2} = \delta_\ell\frac{T_0}{\tau_\ell^2} = T_0\frac{\theta'_\ell}{\left(\theta''_\ell\right)^2}. \qquad \text{(B.8)}$$

Based on the equations (B.6) and (B.8), we can easily calculate the entries of the matrices **J** and **G**, and further compute the CRLB.

ACKNOWLEDGMENTS

This work has been supported by Academy of Finland project no. 44876. The material in this manuscript was presented in part at GENSIPS 2004, Workshop on Genomic Signal Processing and Statistics, Baltimore, Maryland, USA, May 26–27, 2004.

REFERENCES

[1] R. Matoba, S. Saito, N. Ueno, C. Maruyama, K. Matsubara, and K. Kato, "Gene expression profiling of mouse postnatal cerebellar development," *Physiol. Genomics*, vol. 4, pp. 155–164, 2000.

[2] J. Ollila and M. Vihinen, "Microarray analysis of B cell stimulation," *Vitam. Horm.*, vol. 64, pp. 77–99, 2002.

[3] S. Saito, R. Matoba, N. Ueno, K. Matsubara, and K. Kato, "Comparison of gene expression profiling during postnatal development of mouse dentate gyrus and cerebellum," *Physiol. Genomics*, vol. 8, pp. 131–137, 2002.

[4] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.

[5] A. Schliep, A. Schonhuth, and C. Steinhoff, "Using hidden Markov models to analyze gene expression time course data," *Bioinformatics*, vol. 19, no. 1, pp. i255–i263, 2003.

[6] M. J. L. de Hoon, S. Imoto, and S. Miyano, "Statistical analysis of a small set of time-ordered gene expression data using linear splines," *Bioinformatics*, vol. 18, no. 11, pp. 1477–1485, 2002.

[7] C. S. Moller-Levet, K.-H. Cho, and O. Wolkenhauer, "Microarray data clustering based on temporal variation: FCV with TSD preclustering," *Applied Bioinformatics*, vol. 2, no. 1, pp. 35–45, 2003.

[8] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proceedings of the National Academy of Sciences of the USA*, vol. 99, no. 14, pp. 9121–9126, 2002.

[9] C. S. Moller-Levet, K.-H. Cho, H. Yin, and O. Wolkenhauer, "Clustering of gene expression time series data," Tech. Rep., Systems Biology and Bioinformatics Group, University of Rostock, Germany, November 2003, http://www.sbi.uni-rostock.de/publications.htm.

[10] I. Tăbuş and J. Astola, "Clustering the non-uniformly sampled time series of gene expression data," in *Proc. Seventh EURASIP-IEEE International Symposium on Signal Processing and Its Applications (ISSPA '03)*, vol. 2, pp. 61–64, Paris, France, July 2003.

[11] C. D. Giurcăneanu, I. Tăbuş, and J. Astola, "Clustering time series gene expression data based on sum-of-exponentials fitting," in *Workshop on Genomic Signal Processing and Statistics (GENSIPS '04)*, Baltimore, Md, USA, May 2004, CD-ROM (4 pages).

[12] T. Chen, H. L. He, and G. M. Church, "Modeling gene expression with differential equations," in *Biocomputing 1999: Pacific Symposium on Biocomputing*, R. Altman, A. Dunker, L. Hunter, T. Klein, and K. Lauderdale, Eds., vol. 4, pp. 29–40, World Scientific Publishing, Mauna Lani, Hawaii, USA, January 1999.

[13] M. de Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano, "Inferring gene regulatory networks from time-ordered gene expression data of Bacillus Subtilis using differential equations," in *Biocomputing 2003: Proc. Pacific Symposium*, R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, Eds., vol. 8, pp. 17–28, World Scientific Publishing, Kauai, Hawaii, USA, January 2003.

[14] C. D. Giurcăneanu, I. Tăbuş, and J. Astola, "Linear algebra results applied to differential equations model for gene expression data," in *The 2nd TICSP Workshop on Computational Systems Biology (WCSB '04)*, pp. 31–32, Helsinki-St.Petersburg, June 2004.

[15] I. Tăbuş, C. D. Giurcăneanu, and J. Astola, "Genetic networks inferred from time series of gene expression data," in *Proc. First International Symposium on Control, Communications and Signal Processing (ISCCSP '04)*, pp. 755–758, Hammamet, Tunisia, March 2004.

[16] M. R. Osborne and K. Smyth, "A modified Prony algorithm for fitting functions defined by difference equations," *SIAM Journal of Scientific and Statistical Computing*, vol. 12, no. 2, pp. 362–382, 1991.

[17] H. Ouibrahim, "Prony, Pisarenko, and the matrix pencil: a unified presentation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, pp. 133–134, 1989.

[18] D. Kundu and A. Mitra, "Fitting a sum of exponentials to equispaced data," *Sankhya, Series B*, vol. 60, no. 3, pp. 448–463, 1998.

[19] K. Holmstrom and J. Petersson, "A review of the parameter estimation problem of fitting positive exponential sums to empirical data," *Applied Mathematics and Computation*, vol. 126, no. 1, pp. 31–61, 2002.

[20] M. Al-Baali and R. Fletcher, "An efficient line search for nonlinear least squares," *Journal of Optimization Theory and Applications*, vol. 48, no. 3, pp. 359–377, 1986.

[21] J. Petersson and K. Holmstrom, "Initial values for the exponential sum least squares fitting problem," Tech. Rep. IMa-TOM-1998-01, Department of Mathematics and physics, Mälardalen University, Sweden, http://www.mdh.se/ima/personal/khm01/tom/tom-prints/print-technical_reports.htm, June 1998.

[22] A. R. Gallant, *Nonlinear Statistical Models*, John Wiley & Sons, New York, NY, USA, 1987.

[23] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[24] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1997.

[25] P. Stoica and R. L. Moses, "On biased estimators and the unbiased Cramer-Rao lower bound," *Signal Processing*, vol. 21, pp. 349–350, 1990.

[26] P. Stoica and B. Ottersten, "The evil of superefficiency," *Signal Processing*, vol. 55, no. 1, pp. 133–136, 1996.

[27] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[28] J. Rissanen, "MDL denoising," *IEEE Trans. Inform. Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[29] J. Rissanen, "Lectures on statistical modeling theory," Tampere University of Technology, Tampere, Finland, August 2004.

[30] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models," *Pattern Recognition*, vol. 28, pp. 781–793, 1995.

[31] J. W. Evans, W. B. Gragg, and R. J. LeVeque, "On least squares exponential sum approximation with positive coefficients," *Mathematics of Computation*, vol. 34, no. 149, pp. 203–211, 1980.

[32] B. S. Everitt, *Cluster Analysis*, Edward Arnold, London, UK, 3rd edition, 1993.

[33] S. Draghici and S. A. Krawetz, "Global functional profiling of gene expression data," in *A Practical Approach to Microarray Data Analysis*, D. P. Berrar, W. Dubitzky, and M. Granzow, Eds., pp. 306–325, Kluwer Academic Publishers, Boston, Mass, USA, 2002.

[34] S. S. Wilks, *Mathematical Statistics*, John Wiley & Sons, New York, NY, USA, 1962.

Ciprian Doru Giurcăneanu was born in Birlad, Romania, in 1968. He received the M.S. degree from the Department of Control and Computers, "Politehnica" University of Bucharest, Romania, in 1993, and the Ph.D. degree (with honors) from the Department of Information Technology, Tampere University of Technology, Finland, in 2001. From 1993 to 1997, he was a Junior Assistant at "Politehnica" University of Bucharest. Since 1997, he has been with Institute of Signal Processing, Tampere University of Technology, where he is currently a Senior Researcher. From September 2002 to August 2003, he was a Research Fellow with the Academy of Finland. His current research interests include genomics and lossless signal compression.

Ioan Tăbuş received the M.S. degree in electrical engineering in 1982, the Ph.D. degree from the "Politehnica" University of Bucharest, Romania, in 1993, and the Ph.D. degree (with honors) from Tampere University of Technology (TUT), Finland, in 1995. He was a Teaching Assistant from 1984 to 1990, Lecturer from 1990 to 1993, and Associate Professor from 1994 to 1995 with the Department of Control and Computers, "Politehnica" University of Bucharest. From 1996 to 1999, he was a Senior Researcher at TUT. Since January 2000, he has been a Professor with the Institute of Signal Processing at TUT. His research interests include genomic signal processing, speech, audio, image and data compression, joint source and channel coding, nonlinear signal processing, and image processing. He is a coauthor of two books and more than 90 publications in the fields of signal compression, image processing, and system identification. He is a Senior Member of IEEE and Associate Editor for IEEE Transactions on Signal Processing. He was a Chair of IEEE SP/CAS Chapter of Finland Section. Dr. Tăbuş is a corecipient of 1991 "Traian Vuia" Award of the Romanian Academy and corecipient of the NSIP 2001 Best Paper Award and Norsig 2004 Best Paper Award.

Jaakko Astola (Fellow of IEEE) received the Ph.D. degree in mathematics from Turku University, Finland, in 1978. From 1976 to 1977, he was with the Research Institute for Mathematical Sciences, Kyoto University, Kyoto, Japan. Between 1979 and 1987, he was with the Department of Information Technology, Lappeenranta University of Technology, Lappeenranta, Finland. In 1984, he worked as a Visiting Scientist in Eindhoven University of Technology, The Netherlands. From 1987 to 1992, he was an Associate Professor in Applied Mathematics at Tampere University, Tampere, Finland. From 1993, he has been a Professor of Signal Processing and Director of Tampere International Center for Signal Processing leading a group of about 60 scientists and was nominated as Academy Professor by Academy of Finland (2001–2006). His research interests include signal processing, coding theory, and spectral techniques and statistics.