Int. J. Appl. Math. Comput. Sci., 2003, Vol. 13, No. 3, 337–345

PROPERTIES OF A SINGULAR VALUE DECOMPOSITION BASED DYNAMICAL MODEL OF GENE EXPRESSION DATA

KRZYSZTOF SIMEK*

* Institute of Automatic Control, Silesian University of Technology
ul. Akademicka 16, 44–100 Gliwice, Poland
e-mail: ksimek@ia.polsl.gliwice.pl

Recently, data on multiple gene expression at sequential time points were analyzed using the Singular Value Decomposition (SVD) as a means to capture dominant trends, called characteristic modes, followed by the fitting of a linear discrete-time dynamical system in which the expression values at a given time point are linear combinations of the values at a previous time point. We attempt to address several aspects of the method. To obtain the model, we formulate a nonlinear optimization problem and present how to solve it numerically using the standard MATLAB procedures. We use freely available data to test the approach. We discuss the possible consequences of data regularization, called sometimes "polishing", on the outcome of the analysis, especially when the model is to be used for prediction purposes. Then, we investigate the sensitivity of the method to missing measurements and its abilities to reconstruct the missing data. Summarizing, we point out that approximation of multiple gene expression data preceded by SVD provides some insight into the dynamics, but may also lead to unexpected difficulties, like overfitting problems.

Keywords: multiple gene expression, singular value decomposition, dynamical model of gene expression data

1. Introduction

Multiple gene expression methods reach maturity as a tool to investigate dynamical changes in genomes. The principal aim is to capture the dependencies between expressions of different genes. Attempts to achieve it were made even before the introduction of DNA chips and SAGE (Sequential Analysis of Gene Expression, (Velculescu *et al.*, 1995)), in several different ways, depending on particular biological systems with the corresponding different time scales. Time sequences of chromosomal aberrations were reconstructed in a number of tumor systems, dating from the paper (Vogelstein *et al.*, 1988) on colon cancer and continuing with a recent series of papers on phylogenetic models of tumors (e.g., Radmacher *et al.*, 2001). The techniques employed in these papers belong to the mainstream of phylogenetic reconstruction, with the time flow represented by distances based on probabilistic models of evolution.

Another type of methods involves attempts to understand gene interaction by analyzing single-time snapshots representing equilibrium-state solutions under conditions in which particular genes are down- or up-regulated. These methods are based on various nonlinear models, including perceptrons and neural networks (Kim *et al.*, 2001).

Recently, data on multiple gene expression at sequential time points were analyzed using Singular Value Decomposition (SVD) as a means to capture dominant trends, followed by the fitting of a linear time-discrete dynamical system of the form

$$Y(t + \Delta t) = MY(t),$$

to the dominant trend characteristics (Holter *et al.*, 2000; 2001). This approach can be arguably employed for two purposes: First, the short-time changes in the components of vector $Y(t)$ can be expressed using matrix M,

$$\Delta Y(t) = Y(t + \Delta t) - Y(t) = (M - I)Y(t).$$

Therefore, the off-diagonal entries of M reflect linear approximations of the influence of some components of vector $Y(t)$ on the changes in other components. However, this kind of sensitivity analysis is not as straightforward as it might seem, since the components of $Y(t)$ are themselves combinations of expressions including the number of genes.

Second, dynamical system representation may help to reconstruct missing measurements at some time points, by providing an interpolation between the time points at which measurements exist.

In the present paper, we attempt to address several aspects of the method presented in (Holter *et al.*,

2000; 2001). We slightly reformulate the statement of the method to make it more amenable to mathematical analysis. Then we discuss the possible consequences of data regularization, called "polishing" by the original authors, on the outcome of the analysis. Also, we investigate the sensitivity of the method to missing measurements and its abilities to reconstruct the missing data. We use the same data as in (Holter *et al.*, 2000; 2001), for comparison purposes. The computer software applied was written in the Matlab programming language. The original m-files are available from the authors upon request.

Further comments, including possible new applications of the method, are included in the Discussion.

2. Algorithm Description

2.1. Singular Value Decomposition (SVD)

The singular value decomposition of any $n \times m$ matrix A has the form (e.g., (Golub and van Loan, 1996; Watkins, 1991)):

$$A = USV^T, \tag{1}$$

where U is an $n \times n$ orthonormal matrix whose columns are called the left singular vectors of A, and V is an $m \times m$ orthonormal matrix whose columns are called the right singular vectors of A. For $n > m$, the matrix S has the following structure:

$$S = \begin{bmatrix} s_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & s_m \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}.$$

The diagonal elements of the matrix S are customarily listed in descending order, $s_1 \geq s_2 \geq \cdots \geq s_m \geq 0$, and are called the singular values of A.

The properties of SVD matrices are as follows:

1. The singular values of a rectangular matrix A are equal to the square roots of the eigenvalues λ_1, $\lambda_2, \ldots, \lambda_m$ of the matrix $A^T A$.
2. The rank of the matrix A is equal to the number of positive singular values
$$\mathrm{rank}(A) = r, \quad r \leq m.$$
3. The Euclidean norm of A is equal to the largest singular value,
$$\|A\|_2 = s_1.$$
4. The first r columns of the matrix U form an orthonormal basis for the space spanned by the columns of A.
5. The first r columns of the matrix V form an orthonormal basis for the space spanned by the rows of A.

2.2. Data

SVD can be used to analyze the time dynamics of gene expression data (Holter *et al.*, 2001). Each row of the matrix of gene expression A corresponds to a different gene, and each column corresponds to a different time point at which the expression data were measured. The entries of the matrix A contain the gene's relative logarithm expression ratios at discrete time points. For up-regulated genes the ratios are positive while for down-regulated genes they are negative. Since in most applications the number of samples or time points assayed is much smaller than the number of genes investigated, only the case of $n > m$ is considered.

In (Holter *et al.*, 2000), before applying SVD, the data were regularized using polishing. The polishing procedure includes replacing the original data matrix A with a new matrix of the form

$$[A_{ij} - \bar{A}_{\cdot j} - \bar{A}_{i\cdot} + \bar{\bar{A}}_{\cdot\cdot}],$$

where $\bar{A}_{\cdot j}$ is the average of the j-th column of A, $\bar{A}_{i\cdot}$ is the average of the i-th row of A, and $\bar{\bar{A}}$ is the average of all entries of A. This new matrix will also be called A, which involves no ambiguity. After polishing, the rows and columns of the matrix have zero mean values. Because of polishing the rank of A is equal to $r \leq m - 1$. Depending on circumstances, polishing may or may not be desirable.

2.3. Characteristic Modes

Let us denote by X_i, $i = 1, \ldots, r$ the upper r rows of matrix SV^T. The orthogonal vectors X_i are called the characteristic modes associated with matrix A:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_r \end{bmatrix} = \begin{bmatrix} s_1 v_1^T \\ \vdots \\ s_r v_r^T \end{bmatrix}.$$

The time changes of the j-th gene, included in the row A_j of matrix A, can be obtained as linear combinations of the characteristic modes. The coefficients of each combination are the corresponding entries of matrix U:

$$A_j = \sum_{i=1}^{r} U_{ji} X_i.$$

Usually, not all the characteristic modes are needed to reconstruct gene expression patterns with a reasonable accuracy (Alter *et al.*, 2000; Holter *et al.*, 2000; Raychaudhuri *et al.*, 2000). We may use a truncated expression

$$A_j = \sum_{i=1}^{l} U_{ji} X_i, \quad l < r.$$

The contribution of modes to the gene pattern decreases from the higher order to the lower order modes. The singular values which represent the magnitudes of the corresponding modes can be used as measures of the relative significance of each characteristic mode in terms of the fraction of the overall expression that it captures

$$p_i = \frac{s_i^2}{\sum\limits_{j=1}^{r} s_j^2}, \quad i = 1, \ldots, r.$$

A similar index can be defined for each gene:

$$c_k^i = \frac{(U_{ki} s_i)^2}{\sum\limits_{j=1}^{r} (U_{kj} s_j)^2}.$$

It defines the contribution of the i-th mode to the temporal pattern of the k-th gene. There are several heuristic methods to estimate the number l of the most significant characteristic modes (Everitt and Dunn, 2001; Jackson, 1991). One of the simplest techniques is to retain just enough modes to capture a large percentage of the overall expression. Usually, the values of 70–90% are proposed. Another procedure is to exclude characteristic modes such that the fraction of expression p_i they capture is less than $(70/r)\%$. Another method consists in examining the so-called scree plots for s_i^2 or $\log s_i^2$. Using this method, we can usually find a natural border between significant and insignificant singular values (the so-called elbow).

2.4. Dynamical Model for Characteristic Modes

Since the characteristic modes are functions of time, we can try to find a discrete-time dynamical model of changes in the modes following the approach from (Holter *et al.*, 2001). We assume the simplest linear model in which the expression values at a given time moment are linear combinations of the values at previous time instants.

Denote by $Y(t)$ the expression level of all characteristic modes at time t,

$$Y(t_j) = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{qj} \end{bmatrix}, \quad j = 1, \ldots, m, \tag{2}$$

where $q = r$ is the rank of matrix A. Assuming that the original data matrix has a full rank, the rank of A is equal to $r = m - 1$ or $r = m$, depending on whether or not polishing was carried out. The matrix of characteristic modes can be rewritten in the form

$$X = \big[Y(t_1), Y(t_2), \ldots, Y(t_m)\big],$$

where t_i are time points at which the gene expression was measured.

The model can be written in the form of the linear equation

$$Y(t + \Delta t) = MY(t), \tag{3}$$

where M is a $q \times q$ translation matrix ($q \leq m$) and Δt stands for the time step for the dynamical model. For evenly spaced measurements, Δt can be found from the expression $\Delta t = t_{i+1} - t_i$, where $t_i = i\Delta t$. Otherwise, Δt is defined as the maximal time interval such that each measurement time is an integer multiple of Δt, i.e., $t_i = n_i \Delta t$.

Since, as was mentioned earlier, time-series data can often be represented by the most significant modes only and a part of characteristic modes can be excluded, we can try to build a reduced-order model taking into account only a small number of variables. In this case the dimension of the vector (2) is $q = l$, but the form of the dynamical model (3) is not changed.

To obtain the model, we find matrix M based on the knowledge of temporal patterns of characteristic modes. The optimization problem as stated in (Holter *et al.*, 2001) consists in minimizing the performance index of the form

$$J = \frac{\sum\limits_{j=1}^{m} \|Y(t_j) - Z(t_j)\|^2}{\sum\limits_{j=1}^{m} \|Y(t_j)\|^2}, \tag{4}$$

where $Z(t)$ is a time variable described by the discrete linear equation

$$Z(t_1 + k\Delta t) = M^k Y(t_1),$$

with initial condition $Z(t_1) = Y(t_1)$. Since the measurements $Y(t_j)$ are given, the problem consists in finding the q^2 entries of matrix M which minimize J. In general, this minimization problem is nonlinear.

2.5. Steps of the Proposed Approach

1. Regularize gene expression data: rows and columns of A must have zero mean.

   Output:

   - polished data ready for SVD,

- the rank of the polished data matrix A which is decreased and equal to $r = m - 1$.

2. Perform SVD, i.e., find matrices U, S and V for the polished data.

   Output:
   - singular values s_i,
   - columns v_i of matrix V (defining characteristic modes),
   - the first r columns u_i of matrix U (needed for gene expression data reconstruction).

3. Find characteristic modes X_i.

   Output:
   - matrix X containing (in rows) temporal patterns of X_i.

4. Solve the resulting optimization problem (full or reduced order).

   Output:
   - translation matrix M.

3. Solution Methods

3.1. Evenly Spaced Measurements and $q=r$

For evenly spaced measurements and $q = r$, solving the problem leads to solving a system of linear algebraic equations. We have

$$Y(t_{k+1}) = Y_{k+1} = \begin{bmatrix} M_{1\cdot} \\ M_{2\cdot} \\ \vdots \\ M_{r\cdot} \end{bmatrix} Y_k$$

$$= \begin{bmatrix} Y_k^T & 0 & 0 & \cdots & 0 \\ 0 & Y_k^T & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & Y_k^T & 0 \\ 0 & 0 & \cdots & & Y_k^T \end{bmatrix} \begin{bmatrix} M_{1\cdot}^T \\ M_{2\cdot}^T \\ \vdots \\ M_{r\cdot}^T \end{bmatrix}, \quad (5)$$

where $M_{1\cdot}, \ldots, M_{r\cdot}$ are the rows of matrix M. Using the notation involving the Kronecker product (Bellman, 1960), we obtain

$$Y_{k+1} = \left(I_r \otimes Y_k^T\right) \begin{bmatrix} M_{1\cdot}^T \\ M_{2\cdot}^T \\ \vdots \\ M_{r\cdot}^T \end{bmatrix},$$

where I_r is the $r \times r$ identity matrix.

As the equality should hold for each $k = 1, 2, \ldots, r$, we obtain a combined equation of the form

$$\begin{bmatrix} Y_2 \\ Y_3 \\ \vdots \\ Y_{r+1} \end{bmatrix} = \begin{bmatrix} I_r \otimes Y_1^T \\ I_r \otimes Y_2^T \\ \vdots \\ I_r \otimes Y_r^T \end{bmatrix} \begin{bmatrix} M_{1\cdot}^T \\ M_{2\cdot}^T \\ \vdots \\ M_{r\cdot}^T \end{bmatrix},$$

or, more compactly,

$$Y = \tilde{Y}\tilde{M}. \quad (6)$$

where $\tilde{Y}$ is a square $r^2 \times r^2$ matrix.

Solving this equation, we obtain optimal elements of matrix M. Assuming that $\tilde{Y}$ is nonsingular, the equation has a unique solution and the value of the performance index is zero. The standard Matlab procedures are used to solve this equation.

3.2. Evenly Spaced Measurements and $q<r$

In this case the optimization problem may be reduced to the solution of an equation similar to (6), but now matrix $\tilde{Y}$ is an $rq \times q^2$ rectangular matrix. The resulting translation matrix M is the least-squares solution to the overdetermined system of equations of the type (6). The obtained fitting is not ideal. Again, the standard Matlab procedures can be applied.

3.3. Unevenly Spaced Measurements and $q\leq r$

For unevenly spaced measurements and the general case $q \leq r$, it is necessary to minimize the goodness-of-fit index J, as defined above. Holter *et al.* (2001) used simulated annealing, while we use a standard Gauss-Newton algorithm (for details, see (Branch and Grace, 1996) and references therein) as provided in Matlab, with very good results. The problem is strongly nonlinear and, in general, very hard to solve, especially for meaningful differences in measurement time intervals. Since the applied optimization algorithm is very sensitive to the choice of the initial guess for the solution, we propose a two-step optimization. In the first step we use a modified performance index (4). Instead of the variable $Z(t_1 + k\Delta t) = M^k Y(t_1)$, we use $Z_1(t_{i+1}) = M^{(n_{i+1}-n_i)}Y(t_i)$, where $t_i = n_i\Delta t$, which prevents raising M to a high power. For the new index we can apply any initial condition, i.e., a null, an identity, or a random matrix. In the second step we return to the original performance index and solve the problem with the initial condition resulting from Step 1. In most cases the appropriate tuning of the parameters of optimization procedures is required to obtain a precise solution.

4. Influence of Data Polishing

An important caveat takes place when using data polishing. This procedure may lead to dynamical models which do not necessarily reflect the intrinsic dynamics of the underlying systems. We demonstrate that when polished data are used, in the case of equally spaced measurements and a full-order model ($q = r = m - 1$), it is always possible to exactly fit the modes using a linear system with matrix M of rank $m - 1$ which has the spectrum composed of all complex m-th roots of unity, except for the root equal to 1. By the spectral mapping theorem, $M^m = I$, where I is the identity matrix. Therefore, this system necessarily leads to a prediction $Y(t_{m+1}) = M^m Y(t_1) = Y(t_1)$, i.e., to a periodic system with period $m\Delta t$.

Let us notice that if A has row sums equal to 0, i.e., $Ae = 0$, where e is the column vector of the appropriate dimension with all entries equal to 1, then we also obtain the same property for the matrix of modes X. Indeed,

$$SV^T e = U^T Ae = U^T 0 = 0. \tag{7}$$

On the other hand, rewriting the expression on the left-hand side of Eqn. (5), we obtain

$$M\,[\,Y_1 \mid Y_2 \mid \cdots \mid Y_{m-1}] = [Y_2 \mid Y_3 \mid \cdots \mid Y_m]\,. \tag{8}$$

However, based on the property (7), we obtain $Y_m = -\sum_{k=1}^{m-1} Y_k$ and, therefore, for matrix M we obtain

$$M = \Big[Y_2 \mid Y_3 \mid \cdots \mid Y_{m-1} \mid -\sum_{k=1}^{m-1} Y_k\Big] \times \big[Y_1 \mid Y_2 \mid \cdots \mid Y_{m-1}\big]^{-1}, \tag{9}$$

while assuming that $\det([Y_1 \mid Y_2 \mid \cdots \mid Y_{m-1}]) \neq 0$. For a complex number λ to be an eigenvalue of matrix M, it is necessary and sufficient to satisfy the equation $\det(M - \lambda I) = 0$, which, in view of the expression (9), is equivalent to

$$\det\left(\Big[Y_2 \mid Y_3 \mid \cdots \mid Y_{m-1} \mid -\sum_{k=1}^{m-1} Y_k\Big] - \lambda\,[Y_1 \mid Y_2 \mid \cdots \mid Y_{m-1}]\right) = 0.$$

This can be written down in the form

$$\det\big(\,[Y_1 \mid Y_2 \mid \cdots \mid Y_{m-1}]\,Q_{m-1}\big) = 0,$$

where

$$Q_{m-1} = \begin{bmatrix} -\lambda & & & & & -1 \\ 1 & -\lambda & & & \mathbf{0} & -1 \\ & 1 & -\lambda & & & -1 \\ & & 1 & \ddots & & -1 \\ & & & \ddots & \ddots & \vdots \\ & & & & -\lambda & \vdots \\ & \mathbf{0} & & & 1 & -1 \\ & & & & 1 & -(1+\lambda) \end{bmatrix}.$$

Expanding $\det(Q_{m-1})$ starting from the upper left element, we obtain

$$\det(Q_{m-1}) = (-1)^{m-1}(1 + \lambda + \lambda^2 + \cdots + \lambda^{m-1}).$$

The solutions of $\det(Q_{m-1}) = 0$, being identical with the eigenvalues of M, are therefore equal to all the complex m-th roots of unity, except for the root equal to 1, as claimed.

5. Results

To illustrate the ideas presented in the paper, we used freely available data on the yeast cdc-15 synchronized cell cycle described in (Spellman *et al.*, 1998). In a yeast culture synchronized by CDC15, over 6000 genes were monitored over approximately 2.5 cell cycle periods. Almost 800 of them were classified to be cell cycle regulated. We chose a data set consisting of 12 measurements at 20 minute intervals, starting at $t_1 = 10$ minutes. As in (Holter *et al.*, 2000), we disregarded the last 3 data columns corresponding to the beginning of the third cell cycle, where the data were becoming progressively less synchronized.

The analysis consists of three parts. In Part 1 we built a dynamical model for the original data. Since the measurements are evenly spaced in time, the analysis leads to the solution of a system of linear algebraic equations. We show the influence of data polishing in this case. In Parts 2 and 3 we deleted portions of the data to test the reconstruction properties of dynamical system fitting. In Part 2 we deleted two columns (times $t = 70, 150$), and in Part 3, 6 columns (times $t = 70,\ 110,\ 130,\ 170,\ 190,\ 210$) of the data matrix, obtaining two modified data sets with unevenly spaced measurements. The estimation of the translation matrix in these cases requires solving a nonlinear optimization problem as described earlier.

Table 1 shows the singular values (s_i) and the coefficients of relative significance (p_i) of each characteristic

Table 1. Singular values (s_i) and coefficients of relative significance (p_i) of each characteristic mode, based on the data considered.

Part 1				Part 2		Part 3	
unpolished		*polished*					
s_i	p_i	s_i	p_i	s_i	p_i	s_i	p_i
39.26	0.40	38.59	0.43	36.96	0.46	29.06	0.45
30.75	0.25	30.50	0.27	27.02	0.25	22.48	0.27
22.62	0.13	18.75	0.10	17.41	0.10	16.98	0.15
16.24	0.07	15.41	0.07	14.38	0.07	11.65	0.07
11.10	0.03	10.97	0.03	10.80	0.04	9.23	0.05
10.75	0.03	9.97	0.03	9.27	0.03		
9.81	0.03	8.48	0.02	7.56	0.02		
8.48	0.02	7.62	0.02	6.84	0.02		
7.48	0.01	7.22	0.01	6.10	0.01		
7.09	0.01	6.56	0.01				
6.51	0.01	5.65	0.01				
5.71	0.01						

mode. In each case two first characteristic modes capture roughly 70% of the overall variability of the expression. This means that the temporal pattern of the gene expression can be described by the use of two characteristic modes with reasonable accuracy.

For original data the solution matrix M is unique and has properties resulting from data polishing as described in Section 4. Figure 1(a) shows the eigenvalues of the translation matrix M for a full-order dynamical model. They all lie on the unite circle. The resulting model is stable and provides the exact reconstruction of the characteristic modes presented in Fig. 2. For unpolished data (Fig. 1(b)) the obtained model is unstable but it also provides the exact reconstruction of the data. Spectral properties of the reduced second-order model in both cases are very similar.

In Fig. 4(a) the characteristic modes of the first two data sets are presented. It is easy to notice that the small distortion of the data, i.e., deleting two columns, which is equivalent to 16% missing data, did not change the shapes of the original characteristic modes. Using procedures similar to those in (Alter *et al.*, 2001; Wall *et al.*, 2001), we can use the dynamical model to recover the missing data with reasonable fidelity.

Figures 4(b) and 6 show the reconstruction of the characteristic modes with the use of the full dynamical model ($q = r$). For both distorted data sets the reconstruction at the retained measurement points is very precise. This means that the optimization procedure provides accurate solutions.

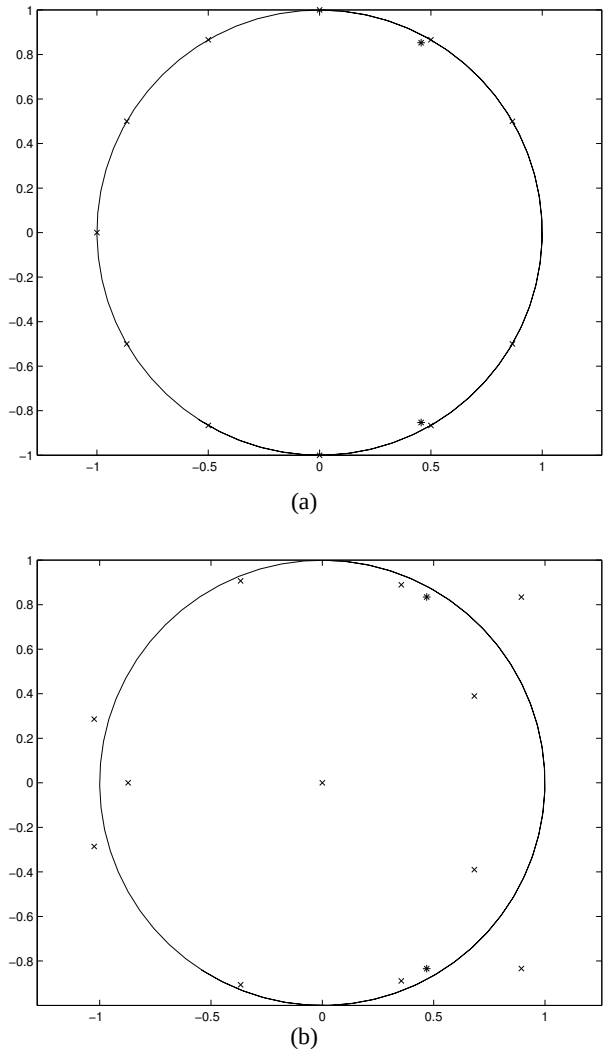


Fig. 1. Eigenvalues of the translation matrix M for full (crosses) and reduced second-order (stars) models for (a) polished, (b) non-polished original data.

However, the inspection of Fig. 6 reveals that for the strongly distorted data set the values of the characteristic modes at retained time points are conserved (compared with the original data), but this time the dynamical model cannot be used to reconstruct the characteristic modes at deleted time points. The obtained dynamical model is unstable, i.e., the model variables are oscillatory with growing amplitudes, although at the measurement points the values are very close to the values of characteristic modes.

As shown in Figs. 3, 5 and 7, which present the reconstruction of the first two characteristic modes for reduced dynamical models in all three cases, the main features of the expression patterns are reproduced quite well. This means that the influence of the high order modes on the dominant ones is weak and the dominant modes could be reconstructed based on a reduced-order model.

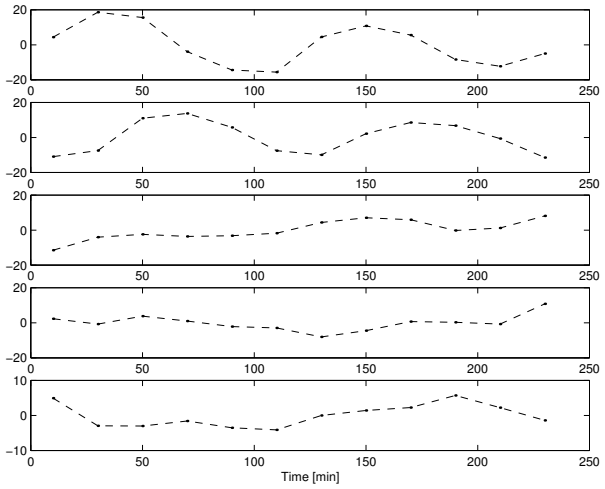


Fig. 2. Six out of 11 characteristic modes for the gene expression data.

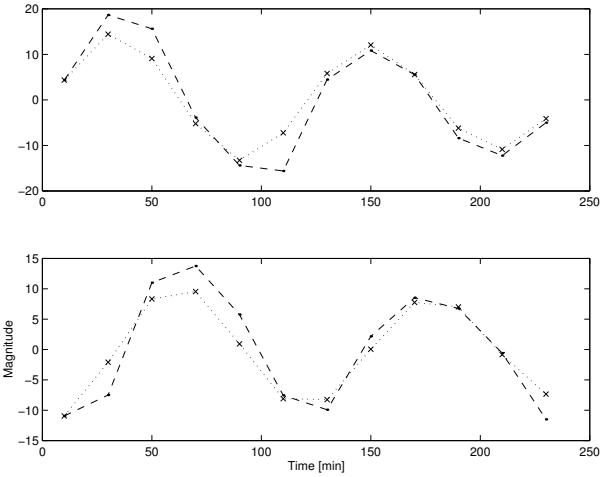
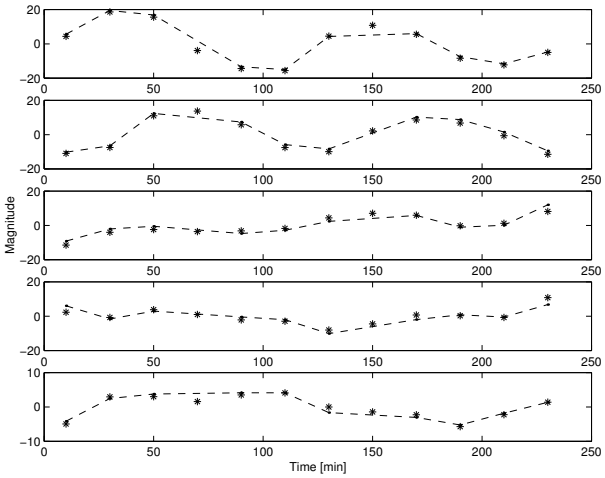
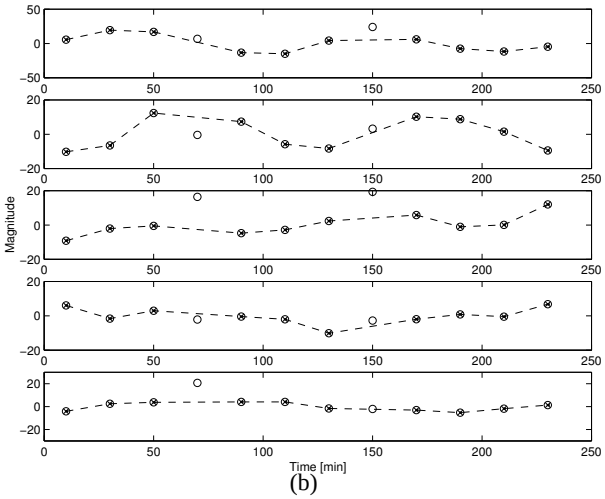


Fig. 3. Reconstruction of the first two characteristic modes for the original data. The dots correspond to characteristic modes, the crosses correspond to the reconstructed variables based on the reduced second-order dynamical model.



(a)



(b)

Fig. 4. (a) Characteristic modes for the original data (stars) and for the first modified data set (dots); (b) Reconstruction of the characteristic modes for the first modified data set. The dots correspond to the characteristic modes for the data set, the crosses correspond to the reconstructed characteristic modes based on the full dynamical model, the circles show the approximation of the temporal pattern resulting from the dynamical model for each time moment $t = n\Delta t$.

6. Discussion

The present note is concerned with the SVD representation of time-dependent multiple gene expression data and their approximation by linear dynamical systems, following the approach of Holter *et al.* (2000; 2001) and using the same data that were originally used by them. SVD allows us to reconstruct the data exactly, using the complete set of characteristic modes, or approximately, using a subset of dominant modes, if the data are provided at evenly or unevenly spaced time points. In this way, the time pattern in the data can be represented by a small number of principal constituents.

However, this decomposition does allow neither the prediction of future trends, nor the reconstruction of the gene expression at the time points at which measurements were not carried out. These tasks can be accomplished using an approximation of the modes by a dynamical system and then either extrapolating the data by running the system for times beyond the existing data points, or interpolating them by running the system for times between the data points.

The extrapolation of data should be approached with caution, particularly if preprocessing is carried out. In-

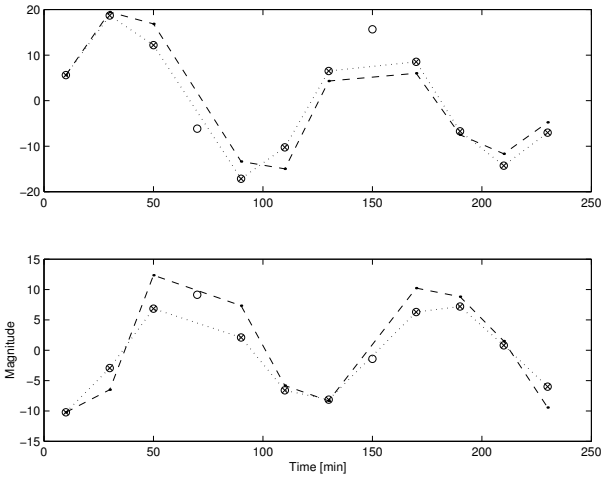


Fig. 5. Reconstruction of the first two characteristic modes for the first modified data set. The dots correspond to the characteristic modes, the crosses correspond to the reconstructed modes based on the second-order dynamical model, the circles show the approximation of the temporal pattern resulting from running the model for each time moment $t = n\Delta t$.

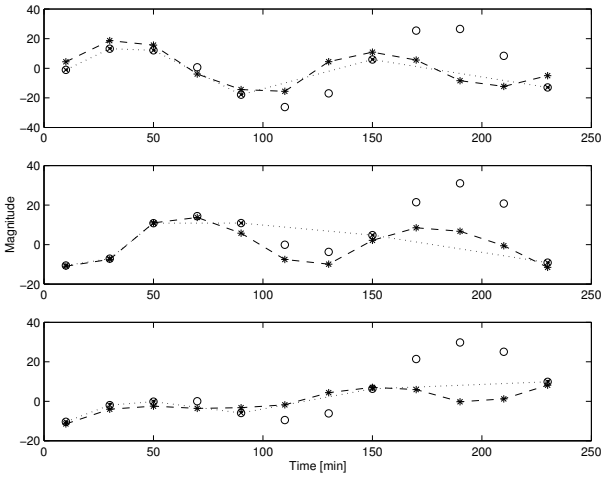


Fig. 6. Reconstruction of characteristic modes for the second modified data set. The dots correspond to the characteristic modes for the data set, the crosses correspond to the reconstructed characteristic modes based on the full dynamical model, the circles show the approximation of the temporal pattern resulting from the dynamical model for each time moment $t = n\Delta t$, the stars represent the temporal pattern of the characteristic modes of original data set.

deed, we demonstrated that using the procedure called polishing in (Holter *et al.*, 2000; 2001) may lead to serious distortions in the dynamics of a linear model of the characteristic modes. Polishing makes the powers of the estimated translation matrix M periodic with period m, yielding a dynamical system with period $m\Delta t$. This

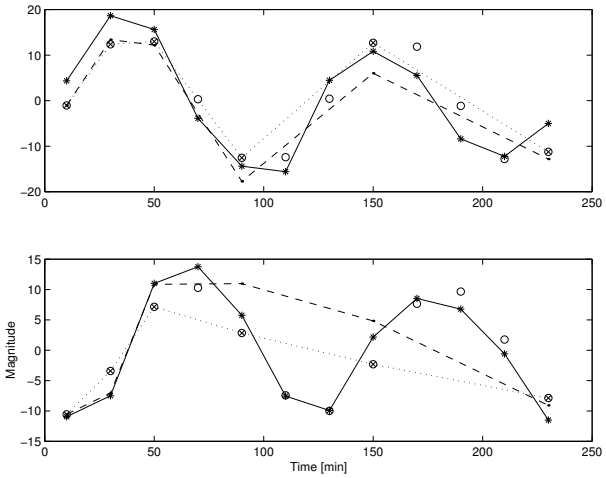


Fig. 7. Reconstruction of the first two characteristic modes for the second modified data set. The dots correspond to characteristic modes, the crosses correspond to the reconstructed variables based on the second-order dynamical model, the circles show approximation of the temporal pattern resulting from the model for each time moment $t = n\Delta t$, the stars represent the temporal pattern of the characteristic modes of original data set.

might be justified if the underlying biological process is by its nature periodic, as in the data we used in this paper. However, in general, polishing is not advisable when the model is to be used for prediction purposes.

As evident from our numerical experiments, approximation by a linear model does not necessarily lead to a correct reconstruction of the missing data. In our experiment, we deliberately removed data points and found that, while accurately fitting the existing observations, the linear dynamical system may provide a very inadequate interpolation at the missing data points. If the data are polished, then the system of the maximum order (equal to the number of measurements decremented by 1) becomes unstable, overshooting between the existing data point (Fig. 6). It is interesting that this phenomenon can be alleviated by reducing the system order (Fig. 7). However, in this case the modelling accuracy decreases.

Summarizing, the approximation of multiple gene expression data preceded by SVD provides some insight into the dynamics but it may also lead to unexpected difficulties. Substantial numerical and mathematical effort will be needed to understand these problems in a satisfactory manner.

Acknowledgements

The work has been supported partly by the grant of the State Committee for Scientific Research (KBN) in Poland No. 4T11F01824 in 2003 and partly by the NIH grant CA

84978 during the author's long-term visit at the Department of Statistics, Rice University, Houston, TX. The author would like to thank Professor Marek Kimmel for considerable help and thoughtful remarks.

References

Alter O., Brown P.O., and Botstein D. (2000): *Singular value decomposition for genome-wide expression data processing and modeling*. — Proc. Natl. Acad. Sci., Vol. 97, No. 18, pp. 10101–10106.

Alter O., Brown P.O. and Botstein D. (2001): *Processing and modeling genome-wide expression data using singular value decomposition.* — Proc. SPIE , Vol. 4266, No. 2, pp. 171–186.

Bellman R. (1960): *Introduction to Matrix Analysis*. — New York: McGraw-Hill.

Branch M.A. and Grace A. (1996): *Matlab Optimization Toolbox. User's Guide*. — Natick, MA: MathWorks.

Everitt B.S. and Dunn G. (2001): *Applied Multivariate Data Analysis*. — New York: Oxford University Press.

Golub G.H. and van Loan C.F. (1996): *Matrix Computations*. — Baltimore: Johns Hopkins University Press.

Holter N.S., Mitra M., Maritan A., Cieplak M., Banavar J.R. and Fedoroff N.V. (2000): *Fundamental patterns underlying gene expression profiles: Simplicity from complexity*. — Proc. Natl. Acad. Sci., Vol. 97, No. 15, pp. 8409–8414.

Holter N.S., Mitra M., Maritan A., Cieplak M., Fedoroff N.V. and Banavar J.R. (2001): *Dynamic modeling of gene expression data*. — Proc. Natl. Acad. Sci, Vol. 98, No. 4, pp. 1693–1698.

Jackson J.E. (1991): *A User's Guide to Principal Components*. — New York: Wiley.

Kim S., Dougherty E.R., Bittner M.L., Chen Y., Krishnamoorthy S., Meltzer P. and Trent J.M. (2001): *General nonlinear framework for the analysis of gene interaction via multivariate expression arrays*. — J. Biomed. Optics, Vol. 5, No. 4, pp. 411–424.

Radmacher M.D., Simon R., Desper R., Taetle R., Schaffer A.A. and Nelson M.A. (2001): *Graph models of oncogenesis with an application to melanoma*. — J. Theor. Biol., Vol. 212, No. 4, pp. 535–548.

Raychaudhuri S., Stuart J.M. and Altman R. (2000): *Principal components analysis to summarize microarray experiments: Application to sporulation time series*. — Proc. Pac. Symp. *Biocomput'2000*, Singapore: World Scientific, pp. 455–466.

Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D. and Futcher B. (1998): *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization*. — Mol. Biol. Cell, Vol. 9, No. 12, pp. 3273–3297.

Velculescu V.E., Zhang L., Vogelstein B. and Kinzler K.W. (1995): *Serial analysis of gene expression*. — Science, Vol. 270, No. 5235, pp. 484–487.

Vogelstein B., Fearon E.R., Hamilton S.R., Kern S.E., Preisinger A.C., Leppert M., Nakamura Y., White R., Smits A.M. and Bos J.L. (1988): *Genetic alterations during colorectal-tumor development*. — N. Engl. J. Med., Vol. 319, No. 9, pp. 525–532.

Wall M.E., Dyck P.A. and Brettin T.S. (2001): *SVDMAN-singular value decomposition analysis of microarray data*. — Bioinformatics, Vol. 17, No. 6, pp. 566–568.

Watkins D.S. (1991): *Fundamentals of Matrix Computations*. — New York: Wiley.