multiple gene expression, dynamic model of gene expression data, singular value decomposition

Krzysztof SIMEK[*], Marek KIMMEL[**]

ANALYSIS OF DYNAMICS OF GENE EXSPRESSION USING SINGULAR VALUE DECOMPOSITION

Recently, data on multiple gene expression at sequential time points were analyzed, using Singular Value Decomposition (SVD) as a means to capture dominant trends, called characteristic modes, followed by fitting of a linear discrete-time dynamical model 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 standard MATLAB procedures. We use publicly available data to test the approach. Then, we investigate the sensitivity of the method to missing measurements and its possibilities to reconstruct 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.

1. INTRODUCTION

Multiple gene expression methods are gaining maturity as a tool to investigate dynamical changes in the genomes. The principal aim is to capture the dependencies between expressions of different genes. This latter was attempted even before introduction of DNA, in several different ways, depending on particular biological systems with corresponding different time scales. Time sequences of chromosomal aberrations were reconstructed in tumor systems, dating from the paper [14] on colon cancer and continuing with a recent series of papers on phylogenetic models of tumors (e.g.,[10]). The techniques employed there belong to the mainstream of phylogenetic reconstruction, with time flow represented by distances based on probabilistic models of evolution.

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 fitting of a linear time-discrete dynamical system of the form: $Y(t+\Delta t) = MY(t)$ to the dominant trend characteristics [7], [8]. This approach can be, arguably, employed for two purposes: First, the short-time changes in the components of vector $Y(t)$: $\Delta Y(t) = Y(t+\Delta t) - Y(t) = (M-I)Y(t)$ can be expressed using matrix M. Therefore, the off-diagonal entries of M reflect linear approximations of

[*] Department of Automatic Control, Silesian University of Technology, Gliwice, Poland; ksimek@ia.polsl.gliwice.pl

[**] Department of Statistics, Rice University, Houston TX, USA; kimmel@rice.edu

the influence of some components of vector $Y(t)$ on changes of 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 of a number of genes. Second, dynamical system representation may help reconstruct missing measurements at some time points, by providing interpolation between time points at which measurements exist.

In the paper, we attempt to address several aspects of the method presented in [7], [8]. We slightly reformulate the statement of the method to make it more amenable to mathematical analysis. Then, we investigate the sensitivity of the method to missing measurements and its possibilities to reconstruct missing data. We use the same data, for comparison purposes.

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.[6])

$$A = USV^T \qquad (1)$$

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 $n>m$ matrix S has the following structure

$$S = \begin{bmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_m \\ & 0 & \\ & 0 & \end{bmatrix};$$

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

Properties of the SVD matrices:

1. Singular values of rectangular matrix A are equal to square root of eigenvalues λ_1, λ_2,..., λ_m of matrix A^TA.
2. Rank of matrix A is equal to the number r of positive singular values: $rank(A)= r,\ r \leq m$.
3. Euclidean norm of A is equal to the largest singular value: $||A||_2 = s_1$.
4. First r columns of matrix U form an orthonormal basis for the space spanned by the columns of A.
5. First r columns of matrix V form an orthonormal basis for the space spanned by the rows of A.

2.2. BIOLOGICAL DATA

The SVD can be used to analyze time dynamics of gene expression data [8]. Each row of matrix of gene expression A corresponds to a different gene and each column corresponds to a different time point at which expression data were measured. The entries of the matrix A contain 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, then only the case $n>m$ is considered.

In the publication [7] by Holter et al., before applying the SVD, data were regularized using polishing. After polishing, the rows and columns of the matrix have mean values equal to 0. Because of polishing the rank of matrix A is equal to $r \leq m - 1$. Depending on circumstances, polishing can be desirable or not.

2.3. CHARACTERISTIC MODES

Let us denote by X_i the upper r rows of matrix SV^T. The orthogonal vectors $X_i = s_i v_i^T$, $i=1,...,r$ are called the characteristic modes associated with matrix A. The time changes of the jth gene, included in the row A_j of matrix A, can be obtained as linear combinations of the characteristic modes $A_j = \sum_{i=1}^{r} U_{ji} X_i$. The coefficients of the combination are the corresponding entries of matrix U. Usually not all characteristic modes are needed to reconstruct gene expression patterns with a reasonable accuracy ([11], [2], [7]). We may use a truncated expression: $A_j \approx \sum_{i=1}^{l} U_{ji} X_i, \quad l \leq r$

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 relative significance of each characteristic mode in terms of the fraction of overall expression that it captures

$$p_i = \frac{s_i^2}{\sum_{j=1}^{r} s_j^2}; \qquad i = 1,...r$$

There are several heuristic methods to estimate the number l of the most significant characteristic modes ([5],[9]). One of the simplest is to retain just enough modes to capture large percentage of overall expression. Usually values of 70-90% are proposed. The other procedure is to exclude characteristic modes such that the fraction of expression p_i they capture is less than $(70/r)\%$. Different method is examination of so-called scree plots for s_i^2 or $\log s_i^2$.

2.4. DYNAMIC MODEL FOR CHARACTERISTIC MODES

Since characteristic modes are function of time, we can try, following approach from [8], to find a discrete-time dynamical model of changes of the modes. We assume the simplest linear model in which the expression values at a given time are linear combinations of the values at a pre-

vious time. Let us denote by $Y(t_j)$ the expression level of all characteristic modes at time points t_j, when gene expression was measured. Matrix of characteristic modes can now be rewritten as: $X = [Y(t_1), Y(t_2), ..., Y(t_m)]$ and the dynamical model can be written in the form of a linear equation:

$$Y(t + \Delta t) = MY(t) \qquad (3)$$

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

Since, as mentioned earlier, time-series data often can be represented by the most significant modes only and a part of characteristic modes can be excluded, we can try to build reduced order model taking into account only small number of variables. In this case the dimension of vector $Y(t)$ 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 [8] consists of minimization of the performance index of the form:

$$J = \frac{\sum_j \left\| Y(t_j) - Z(t_j) \right\|^2}{\sum_j \left\| Y(t_j) \right\|^2} \qquad (4)$$

where $Z(t)$ is a time variable described by linear discrete equation: $Z(t_1 + k\Delta t)) = M^k Y(t_1)$, $k = 1,2,...n_m$ with initial condition $Z(t_1)=Y(t_1)$. Since the measurements $Y(t_j)$ are given, the problem consists of finding the q^2 entries of matrix M, which minimize J. In general this minimization problem is nonlinear.

2.5. METHODS OF SOLUTION

For equally spaced measurements and $q=r$, the solution of the problem leads to the solution of a linear system of algebraic equations.

$$Y = \tilde{Y}\tilde{M} \qquad (5)$$

where $\tilde{Y}$ is a square $r^2 \times r^2$ matrix and $\tilde{M}$ is a vector containing transposed rows of matrix M. Solving the equation one obtains optimal elements of matrix M. Assuming that matrix $\tilde{Y}$ is nonsingular, the equation has one unique solution and the value of the index (4) is equal to 0.

In the case of equally spaced measurements and $q < r$ optimization problem may be brought to the solution of the equation similar to (5), but now matrix $\tilde{Y}$ is a $rq \times q^2$ rectangular matrix. The resulting translation matrix M is the solution in the least squares sense to the overdetermined system of equations of type (5). Obtained fitting is not ideal.

For unequally spaced measurements and the general case $q \leq r$, it is necessary to minimize the goodness of fit index J, as defined above. In [8] the authors used simulated annealing, while we use a standard Gauss-Newton algorithm (see [4] and references therein for details) 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 measurements time intervals. Since the applied optimization algorithm is very sensitive to the choice of the initial guess of the solution, we use a two-step optimization. In most cases appropriate tuning of parameters of optimization procedures is required to obtain a precise solution.

3. RESULTS

To illustrate the considerations we used publicly available data on yeast cdc-15 synchronized cell cycle, described in [13]. In a yeast culture synchronized by cdc-15 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 data set consisting of 12 measurements at 20 minute intervals, beginning at t_1=10 minutes. The analysis consists of three parts. In Part 1 we built a dynamical model for the original data. Since the measurements are equally spaced in time the analysis leads to solution of a linear system of algebraic equations. 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 unequally spaced measurements. Estimation of the translation matrix in these cases requires solving a nonlinear optimization problem as described earlier.

Analysis of singular values (s_i) and coefficients of relative significance (p_i) of characteristic modes in each case reveals that two first characteristic modes capture roughly 70% of the overall variability of the expression. It means that the temporal pattern of gene expression can be described by the use of two characteristic modes with reasonable accuracy.

For original data the solution matrix M is unique The dynamical model provides an exact reconstruction of the characteristic modes. In Figure 1 characteristic modes of first two data sets are presented. It easy to notice that the small distortion of the data, i.e., deleting two columns, equivalent to 16% missing data, did not change shapes of the original characteristic modes. Using similar procedures as in [15] and [1] the dynamical model can be used to recover missing data with reasonable fidelity. Figure 3 and Figure 5 show reconstruction of 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. It means that optimization procedure provides accurate solutions. However, inspection of Figure 5 reveals that for the strongly distorted data set the values of the characteristic modes at retained time points are conserved (compared to the original data), but this time the dynamical model can not be used to reconstruct the characteristic modes at deleted time points. The obtained dynamical model is unstable, i.e., 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 Figures 2, 4 and 6, which present reconstruction of the first two characteristic modes for reduced dynamical models in all three cases the main features of expression patterns are

reproduced quite well. It means that influence of the high order modes on dominant ones is weak and dominant modes could be reconstructed basing on a reduced order model.

4. 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 from [7], [8] and using the same data that were originally used by these authors. SVD allows to reconstruct the data exactly, using the complete set of characteristic modes, or approximately, using a subset of dominant modes, whether the data are provided at equally or unequally spaced time points. In this way, the time pattern in the data can be represented by a small number of principal constituents.

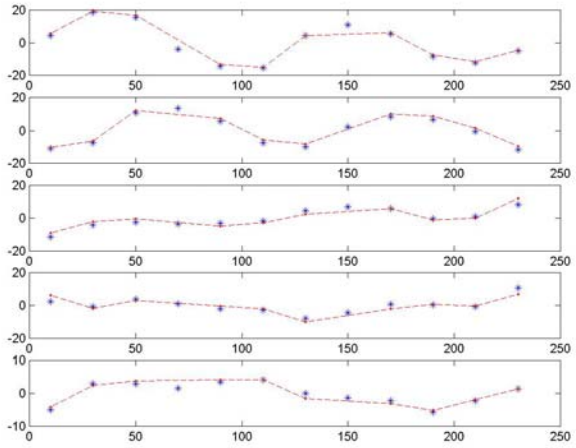

Fig.1. Characteristic modes for the original data (stars) and for the first modified data set

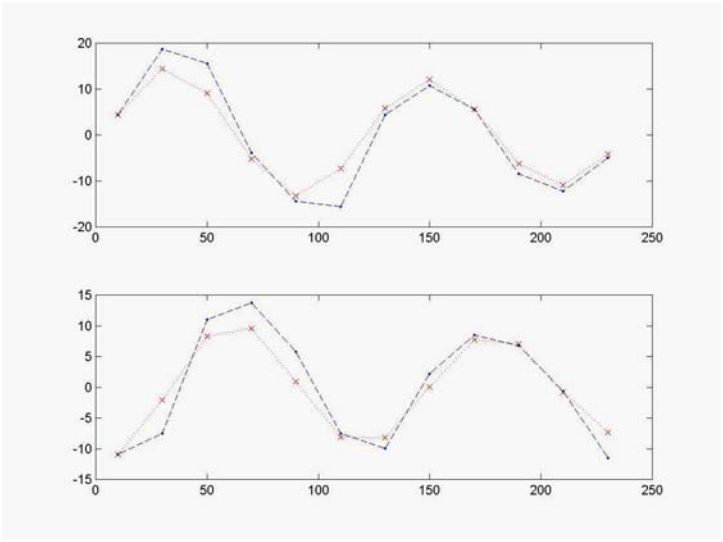

Fig.2. Reconstruction of the first two characteristic modes for original data. The dots correspond to characteristic modes, the crosses correspond to reconstructed variables basing on the reduced dynamical model of order 2.

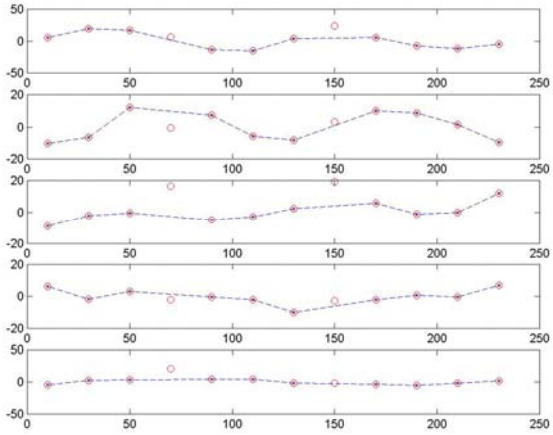

Fig.3. Reconstruction of the characteristic modes for the first modified data set. Dots correspond to the characteristic modes, crosses correspond to the reconstructed characteristic modes (full dynamical model), circles show approximation of the temporal pattern resulting from the dynamical model.

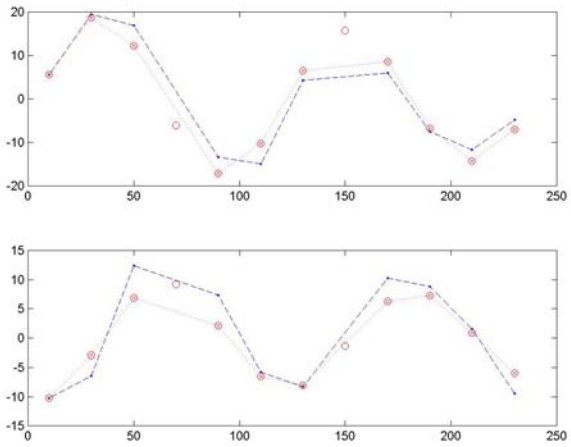

Fig.4. Reconstruction of the first two characteristic modes for the first modified data set. Dots correspond to the characteristic modes, crosses correspond to the reconstructed modes (the second order dynamical model), circles show approximation of the temporal pattern resulting from running the model for each time $t=n*\Delta t$.

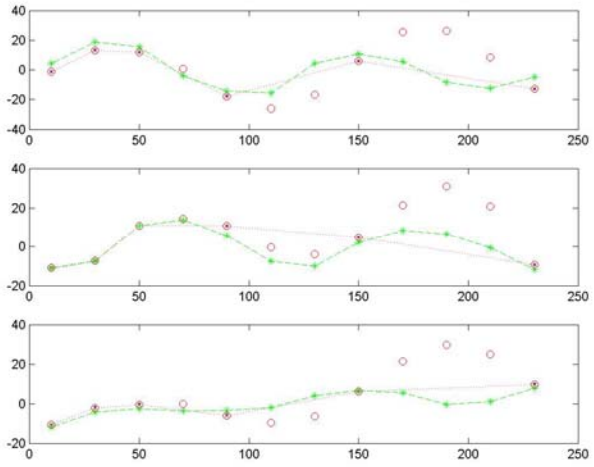

Fig.5. Reconstruction of characteristic modes for the second modified data set. Dots correspond to the characteristic modes for the data set, crosses correspond to the reconstructed characteristic modes based on the full dynamical model, circles show approximation of the temporal pattern resulting from the dynamical model for each time $t = n*\Delta t$, stars represent temporal pattern of characteristic modes of original data set.

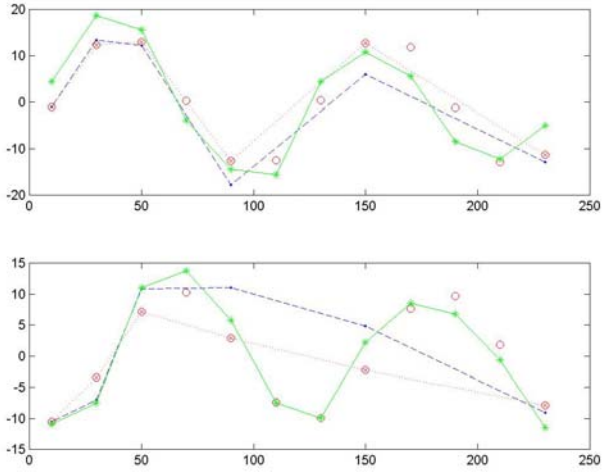

Fig.6. Reconstruction of the first two characteristic modes for the second modified data set. Dots correspond to characteristic modes, crosses correspond to the reconstructed variables based on the second order dynamical model, circles show approximation of the temporal pattern resulting from the model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

However, this decomposition does not allow prediction of future trends or reconstruction of gene expression at 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 by running it for times between the data points.

Extrapolation of the data should be approached with much caution, particularly if preprocessing is carried out. Indeed, we demonstrated in [12] that using the procedure called polishing in [7], [8] may lead to serious distortions of dynamics of a linear model of the

characteristic modes. Polishing causes the powers of the estimated translation matrix M to become periodic with period m, yielding a dynamical system with period $m\Delta t$. This might have a justification 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 correct reconstruction of missing data. In our experiment, we deliberately removed data points and found that the linear dynamical system, while accurately fitting the existing observations, may provide very inadequate interpolations at the missing data points. If the data are polished, then system of the maximum order (equal to the number of measurements minus 1), becomes unstable, overshooting between the existing data point (Figure 5). It is interesting that this phenomenon can be alleviated by reducing the order of the system (Figure 6).

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

BIBLIOGRAPHY

[1] ALTER O., P.O. BROWN, D. BOTSTEIN, Processing and modeling genome-wide expression data using singular value decomposition, Proc. of SPIE 4266 (2001) 171-186.

[2] ALTER O., P.O. BROWN, D. BOTSTEIN, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Natl. Acad. Sci. USA 97 (2000) 10101-10106.

[3] BELLMAN R., Introduction to Matrix Analysis, McGraw-Hill, New York, 1960.

[4] BRANCH M.A., A.GRACE, Matlab Optimization Toolbox. User's Guide, MathWorks, 1996.

[5] EVERITT B.S., G. DUNN, Applied Multivariate Data Analysis, Oxford University Press, New York, 2001.

[6] GOLUB G.H.,C.F. VAN LOAN, Matrix Computations, Johns Hopkins University Press, Baltimore, 1996.

[7] HOLTER N.S., M. MITRA, A. MARITAN, M. CIEPLAK, J.R. BANAVAR, N.V. FEDOROFF, Fundamental patterns underlying gene expression profiles: Simplicity from complexity, Proc. Natl. Acad. Sci. USA 97 (2000) 8409-8414.

[8] HOLTER N.S., M. MITRA, A. MARITAN, M. CIEPLAK, N.V. FEDOROFF, J.R. BANAVAR, Dynamic modeling of gene expression data, Proc. Natl. Acad. Sci. USA 98 (2001) 1693-1698.

[9] JACKSON J.E., A User's guide to principal components, Wiley, New York, 1991.

[10] RADMACHER M. D., R. SIMON, R.DESPER, R. TAETLE, A. A. SCHAFFER, M. A. NELSON, Graph models of oncogenesis with an application to melanoma , J. Theor. Biol. 212 (2001) 535-548.

[11] RAYCHAUDHURI S., J.M. STUART, R. ALTMAN, Principal components analysis to summarize microarray experiments:application to sporulation time series, in: R.B.ALTMAN, K.LAUDERDALE, DUNKER A.K., L.HUNTER, T.E.KLEIN, (Eds.), Proc.Pac.Symp.Biocomput.2000, World Scientific, Singapore, 2000, 455-466.

[12] SIMEK K., M. KIMMEL, A note on estimation of dynamic of multiple gene expression based on singular value decomposition, Mathematical Biosciences [in press].

[13] SPELLMAN P.T., G. SHERLOCK, M.Q. ZHANG,V.R. IYER, K.ANDERS, M.B.EISEN, P.O.BROWN, D.BOTSTEIN, B.FUTCHER, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Mol.Biol.Cell 9 (1998) 3273-3297.

[14] VOGELSTEIN B., E. R. FEARON, S. R. HAMILTON, S. E. KERN, A. C. PREISINGER, M. LEPPERT, Y. NAKAMURA, R. WHITE, A. M. SMITS, J. L. BOS, Genetic alterations during colorectal-tumor development, N. Engl. J. Med. 319 (1988) 525-532.

[15] WALL M.E., P.A. DYCK, T.S. BRETTIN, SVDMAN-singular value decomposition analysis of microarray data, Bioinformatics 17 (2001) 566-568.

This work has been supported by a NIH grant CA 84978, by a KBN (Polish Scientific Committee) grant PBZ/KBN/040/P04/2001 and by a NATO grant LST CLG 977845.