Missing Value Estimation for Time Series Microarray Data Using Linear Dynamical Systems Modeling

Connie Phong
Department of Radiation Oncology
University of California, San Francisco
San Francisco, CA
cphong@cc.ucsf.edu

Raul Singh
Department of Computer Science
San Francisco State University
San Francisco, CA
rsingh@cs.sfsu.edu

Abstract

The analysis of gene expression time series obtained from microarray experiments can be effectively exploited to understand a wide range of biological phenomena from the homeostatic dynamics of cell cycle systems to the response of key genes to the onset of cancer or infectious disease. However, microarray data frequently contain a significant number of missing values making the application of common multivariate analysis methods, all of which require complete expression matrices, difficult. In order to preserve the experimentally expensive non-missing data points in time series gene expression data, methods are needed to estimate the missing values in such a way that preserves the latent interdependencies among time points within individual expression profiles. Thus we propose modeling gene expression profiles as simple linear and Gaussian dynamical systems and apply the Kalman filter to estimate missing values. While other current advanced estimation methods are either sensitive to parameters with no theoretical means of selection or attempt to learn statically from inherently dynamical data, our approach is advantageous exactly because it makes minimal assumptions that are consistent with the biology. We demonstrate the efficiency of our approach by evaluating its performance in estimating artificially introduced missing values in two different time series data sets, and compare it to a Bayesian approach dependent on the eigenvectors of the gene expression matrix as well as a gene wise average imputation for missing values.

1. Introduction

Genomes, or global gene expression, can be considered as a temporal process that self regulates on a feedback loop to maintain homeostasis. The complexity of this process is underscored by the fact that different normal cellular conditions require both the synthesis and repression of very specific sets of proteins. Normal development as well as disease are tied to the functionality or failure of genome systems. Thus elucidating the nature of the regulatory networks that control gene expression is motivated from many biological perspectives and has enormous potential for applications. It is in fact the central aim of current microarray experiments. The analysis of gene expression data obtained from microarray experiments has already proven successful in identifying genes that contribute to common functions and are therefore, at most, possibly coregulated. Efforts to infer explicit networks from microarray experiments, however, have been extensive but inconclusive.

At the most fundamental level it is known that genome systems contain sets of proteins called transcription factors that are required to initiate or repress transcription of particular genes by binding to their regulatory sequences. Furthermore post-translational modifications of proteins can in turn affect the activity of transcription factors. However, there also exists locally interacting processes between subsets of genes that may influence subsequent expression levels. Thus a *complete* description of gene expression systems includes not only the microarray measured gene expression profiles, but also physical yet unmeasurable factors such as protein degradation rates and mRNA levels, as well as intangible interaction and synergistic effects.

Extracting network information from gene expression profiles however requires the consideration of various preprocessing issues. Amongst the most pressing of these issues are the problems which arise from missing data values. Missing values in a gene expression matrix are common to both static and time series experiments alike. They are usually introduced when suspect values are removed from their respective genes' expression profile or by some technical error leading to a failed measurement. The occurrence and handling of missing time points is non-trivial. Of the 800 genes determined by Spellman *et al.* [9] to be cell cy-

978-0-7695-3096-3/08 $25.00 © 2008 IEEE
DOI 10.1109/WAINA.2008.23

cle regulated in *Saccharomyces cervisiae* more than 20% of the genes had incomplete profiles. Common statistical multivariate analyses such as principal component analysis (PCA) and singular value decomposition (SVD) can only be applied to complete expression matrices. On the other hand, hierarchical clustering methods that use the Euclidean distance between gene expression profiles as a metric, must also set work-arounds for dealing with missing time points or dimensions.

Clearly all exisiting analysis methods are extremely sensitive to missing values. Combined with the high costs of data collection, naive approaches such as removing an expression profile with missing data points from further analysis, setting missing values to 0, or setting a particular gene's or sample's missing values to its average expression levels are unsatisfactory if not misleading.

Furthermore, missing data points in time series gene expression data warrant special consideration and methods. As has already been discussed a gene expression system is in itself a complex, interactive temporal process. Thus not only does a strong autocorrelation exists within successive data points of a particular gene profile, but there also exists dependencies between the data points across profiles.

More advanced missing value estimation methods have taken on either global, local, or a combination of global and local approaches. Troyanskaya *et al.* [10] have proposed methods based on K-nearest neighbor (KNNimpute) and SVD (SVDimpute) approaches. They found that although both approaches outperformed imputing missing values to the gene-wise average, the SVD approach consistently outperformed the KNN approach in time series expression data. Both approaches were, however, sensitive to either the K-value of the number of eigenvectors used while there remains no theoretical way to choose these values.

Oba *et al.* [7] have proposed estimating missing values using Bayesian principal component analysis (BPCA). Briefly, missing values are imputed with respect to an estimated posterior distribution for a parameter set based on the principal components of the expression matrix and the eigenvectors of the expression covariance matrix. This method was shown to outperform KNNimpute and SVDimpute, but also relies heavily on eigenvectors.

Hu *et al.* [4] have proposed an "integrative" approach that incorporates information from multiple reference microarray data sets to improve missing value estimates. This approach however is dependent on the existence of valid reference data sets. Despite the authors' contention that the rapid accumulation of microarray data sets improves the performance of their method, it still is dependent on the tenuous assumption that microarray data sets can be easily compared. In fact there are a potential number of biological and experimental inconsistences, such as differences in sampling rate and variations in timing of biological processes, that hinder comparisons of microarray data sets (Bar-Joseph [1]).

Thus a method that estimates missing time series values by making effective use of all the information available– both explicit and latent, both global and local–is desirable. We thus propose a simple linear dynamical systems model for gene expression systems and apply Kalman filtering techniques to estimate values for missing time points. We evaluate our approach and compare it against the row-average method and BPCA. We also discuss the extension of linear dynamical systems modeling to overriding problem of network inference.

It has been recognized that effective models of gene expression are key to ultimately uncovering network information. We thus propose a method for the derivation of a linear dynamical model of genomes from microarray data. Many existing methods for analysis of gene expression data, however, do not take into consideration the dynamic nature of genomes. Clustering methods, for example, treat expression profiles independently. Thus not only are implied dependencies among consecutive points within the same gene's profile ignored but no attempt is made to account for necessary inherent dependencies among genes either.

A linear dynamical systems model has the advantage of fully utilizing the available microarray data, since it explicity considers the various contributing components of genomes and their complex relationships as well as the noise in the microarray data used to extract it. This model thus provides an effective conceptual framework from which to analyze gene expression data from microarray experiments.

2. Systems and methods

2.1. A linear dynamical system model for gene expression systems

A dynamical system can be described by a sequence of time-dependent variables $\mathbf{x}_t$=(x_1(t), x_2(t),...,x_n(t)) $\in \Re^n$, where $\mathbf{x}_t$ describes the complete state of the system and represents all the information that it has available to itself to propagate forward in time. In general, measurements on the system, which are of varying sensitivity, do not give the complete state. Rather a measurement $\mathbf{y}_t$=(y_1(t), y_2(t),...,y_m(t)) $\in \Re^m$ of the system only partially reveals its dynamics and the obscured components, or hidden variables, are lost.

We consider time series expression data obtained from microarray experiments in such a dynamical systems context. The system is taken to be complete gene expression following the conditions of interest in an organism, and the measurements are the gene expression levels as measured by the microarray time series experiment.

In this context, a state space model of complete organismal gene expression can be described by equations (1) and (2).

$$x_{t+1} = Ax_t + w_t \tag{1}$$

$$y_t = Cx_t + v_t \tag{2}$$

In a time course experiment the expression levels of *n* genes are each measured at *m* time points to track the effects of a set of conditions or events resulting in a *m* x *n* expression matrix **G**. Thus $\mathbf{y}_t$ is a *n*-dimensional vector representing the measured expression levels of all *n* genes at time *t* for *t* = 1, ..., *m*. The vectors x_t are *p*-dimensional complete state values from which the observed vectors y_t are generated. The *p* x *p* matrix **A** relates the transition of the state from the current time point to the next time point. (Note that the transitions follow a first-order Markov process.) The *p* x *n* matrix **C** relates the complete state at time *t* to the microarray measurement y_t.

w_t and v_t denote noise associated with the biological and measurement processes respectively. We make the assumption that w_t and v_t are independent (of each other and the initial values of *x* and *y*), white, and have Gaussian distributions

$$p(w) \sim N(0, Q) \tag{3}$$

$$p(v) \sim N(0, R) \tag{4}$$

Q denotes the process noise covariance matrix and **R** the measurement noise covariance matrix. Noise is therefore also hidden.

2.2. A linear dynamical system for gene expression profiles

We now formulate the missing time point problem in the context of linear dynamical systems. Consider gene *i*, or row *i* from the expression matrix **G**, with one or more lost measurements. We wish to estimate expression values for the missing time points i_k. A simplified linear dynamical system can then be derived, where the states $y_{i,t}$ to be estimated are exactly the expression levels for gene *i* at time *t*, and where the correlations have been simplified such that the expression state at one time point determines the expression state at the next time point.

$$y_{i,t+1} = Cy_{i,t} + w_t \tag{5}$$

$$z_{i,t} = Hy_{i,t} \tag{6}$$

In this case then the observations, $z_{i,t}$, give the complete state of the system (the observation transition matrix **H** is the identity matrix **I**). A missing time point in the expression profile of gene *i* is then the lack of corresponding observation $z_{i,t}$. w_t is again white with Gaussian distribution given by equation (3).

Since we have defined gene expression profiles in terms of a linear model subject to Gaussian and white noise processes, we can import the proven utility of Kalman filtering techniques from control and signal processing applications in physical systems.

2.3. The Kalman filter

The Kalman filter is an optimal, recursive algorithm used to estimate the state of a system while minimizing mean square error. It assumes that the system can be modeled linearly, and that all noise is white and follows a Gaussian distribution. The Kalman filter estimates the state of a system at any time *t* by propagating its probability density function conditioned on a set of measurements $\mathbf{z}^T$, p($\mathbf{y}_t|z^T$). Once the conditional probability density function is propagated, the mean or center of the probability mass is taken as the estimate of the state $\mathbf{y}_t$ given $\mathbf{z}^T$. Although the mean of the conditional probability density is the preffered estimate of the state it should be noted that under the assumptions of linearity, Gaussian densities, and whiteness the mean, median and mode of the density functions coincide. Thus the Kalman filter will always result in a *unique* best estimate of the state.

The estimation process is referred to as filtering if T = *t*, prediction if T > *t*, and smoothing if T > *t*.

A broad description of the Kalman filter is that of an estimator with feedback control. The filter begins by projecting forward the current state and error covariances estimates to obtain an *a priori* estimate for the state of the system at the next time step $\hat{y}_t^-$. Then the filter receives feedback in the form of a measurement, incorporating the measurement into the *a priori* estimate to give an improved *a posteriori* estimate of the next state $\hat{y}_t$. The Kalman filter equations for the simple linear system described in equations (5) and (6) are given by the following equations.

$$\hat{y}_t^- = C\hat{y}_{t-1} \tag{7}$$

$$P_t^- = CP_{t-1}C^T + Q \tag{8}$$

$$K_t = P_t^- H^T (HP_t^- H^T + R)^{-1} \tag{9}$$

$$\hat{y}_t = \hat{x}_t^- + K_t(z_t - H\hat{x}_t^-) \tag{10}$$

$$P_t = (I - K_t H)P_t^- \tag{11}$$

where $\mathbf{P}_t^-$ is the *a priori* estimate error covariance, and $\mathbf{P}_t$ is the *a posteriori* estimate covariance. Smoothing considers later measurements $\mathbf{z}_T$, T > *t* to improve estimates of state

$\mathbf{y}_t$ by reducing noise. Smoothing involves an initial forward recursion followed by a backward recursion. In the forward step, the Kalman filter equations are applied through *T* and the values $\hat{y}_t^t$, $\hat{y}_t^{t-1}$, $\mathbf{P}_t^t$, and $\mathbf{P}_t^{t-1}$ for $t = 1, \ldots, T$ are stored. In the backward step, these values are then used to initialize the Kalman smoother equations given by equations (12) - (14).

$$J_{t-1} = P_{t-1}^{t-1} C^T [P_t^{t-1}]^{-1} \tag{12}$$

$$\hat{y}_{t-1}^T = \hat{y}_{t-1}^{t-1} + J_{t-1}(\hat{y}_t^T - \hat{y}_t^{t-1}) \tag{13}$$

$$P_{t-1}^T = P_{t-1}^{t-1} + J_{t-1}(P_t^T - P_t^{t-1})J_{t-1}^T \tag{14}$$

2.4. Estimating missing values using Kalman smoothing

Using the state space model given by Equations (5) and (6) with the state transition matrix **C** set to the identity matrix, Kalman smoothing is then applied to each individual profile to estimate the missing time points.

In actuality the state transition matrices may change from step to step, however, Kalman filtering is robust to the simplification that they are constant. Thus we do not attempt to use computationally intensive methods, such as the expectation maximization algorithm, to learn the transition matrix **C** and argue that setting **C = I** is adequate. This simplification will be investigated in further work. It should also be noted that in learning a transition matrix an inherently dynamical process is treated as static and that such a transition matrix would not have readily translatable biological meaning.

In contrast, the assumptions of our approach are biologically sound. A linear system model for gene expression profiles is appropriate given that temporal changes in the transcriptome are relatively smooth and continuous (Rifkin and Kim [8]). In general, when non-linearities do exist it is often possible to linearize about some nominal point. Furthermore, the Kalman filter can be extended to a nonlinear setting. The assumption that both the system and measurement process noises are Gaussian can also be justified by considering that noise is typically caused by a number of small sources. By the central limit theorem it is known that when a number of independent random variables are added together the summed effect can be described very closely by a Gaussian distribution regardless of the shape of individual densities.

3. Results and Discussion

We tested our Kalman smoother approach to estimating missing data points on two time series microarray data sets. Spellman *et al.* [9] used various methods of cell synchronization to determine cell cycle regulated genes in the yeast *Saccharomyces cerevisiae*. Expression was tracked for 6177 open reading frames over the course of two and a half cell cycles or 290 minutes. We used in particular the cdc15 synchronized expression data. It should be noted that the sampling rate was not constant rather varying from 10 to 20 minutes. We consider only the measurements made at 20 minute intervals starting from the first measurement. Laub *et al.* [5] conducted an analysis of gene transcription over a single cell cycle in the bacterium *Caulobacter crescentus*, collecting expression levels for 2966 predicted reading frames at 15 minute intervals for a total of 11 time points. The data sets "cdc15" and "caulobacter" were formed by removing all expression profiles with at least one missing time point from the published data sets. The data sets were not normalized. The characteristics of the test data sets are summarized in Table 1.

To test the performance of each method, missing values were introduced into the cdc15 and caulobacter data sets by randomly selecting and removing expression values such that a specific percentage of entries over the complete expression matrix are rendered missing. The Kalman smoother method, BPCA, and gene average imputation are then used to recover these artificial missing values. Each test was performed 10 times to reduce randomness. The relative performance of each method can be assessed in part by considering the root mean squared error between the estimated matrix and the true matrix normalized over the variance in the true matrix (NRMSE) as defined in equation (15) and the maximum error between the estimated and true values.

$$NRMSE = \sqrt{\frac{mean[(y_{guess} - y_{true})^2]}{variance[y_{true}]}} \tag{15}$$

The NRMSE of all three methods were high and relatively constant over the percentage of artifically introduced missing entires. The NRMSE of the Kalman smoothing estimates are always bounded above by that of the gene average and below by that of BPCA as shown in Figure 1. Given the assumptions of the model we used, specifically a crude estimation of the state transition matrix and no fine tuning of any other parameters, the performance can be considered comparable to that of BPCA. The results also suggests that

	cdc15-synchronized yeast	caulobacter
ORFs in published data	6178	2966
Percent missing	23.7%	48.1%
ORFs in test set	4712	1538
Time points	15	11

Table 1. Characterisitcs of the data sets comprising the test sets used in evaluations of estimator performance.

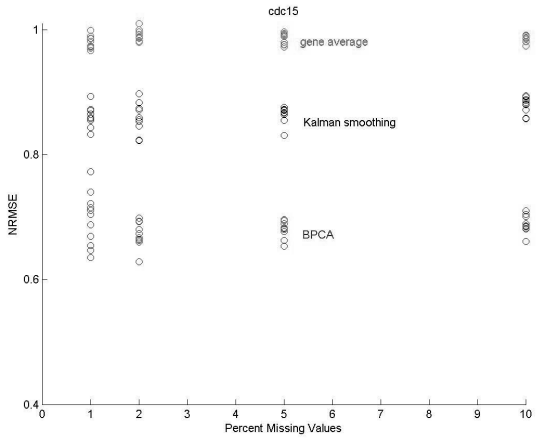

Figure 1. Estimation ability as evaulated by NRMSE for various percentages of missing entries over 10 repetitions.

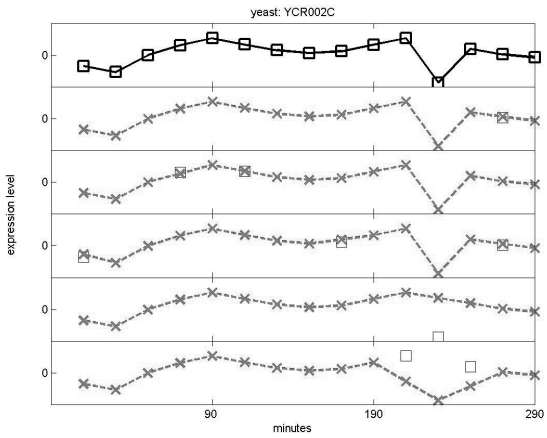

Figure 2. Kalman smoother estimated values for simulated missing time points on a complete gene expression profile for the published yeast cdc15 data. The top panel shows the complete profile. In the subsequent panels, the time points rendered missing are indicated by black squares. The complete estimated profile is shown as a red dotted line.

	cdc15	caulobacter
gene wise average	3.9627	6.8209
BPCA	3.5617	4.2167
Kalman smoother	3.4150	6.3475

Table 2. A comparison of the maximum errors on the estimation methods over 10 repetitions for 5% missing entries.

with refinements our approach may outperform BPCA. In fact for relatively large data sets, such as the cdc15 test set, the Kalman smoother already results in smaller maximum errors than BPCA(Table 2).

To better assess the performance of the Kalman smoother, a variety of missing time point scenarios were simulated on complete profiles. Figures 2 and 3 show that the smoother is robust to the number of missing time points so long as the assumption that smoothness of transcriptome holds. For example, the Kalman smoother was unable to find good estimates around 210 - 270 minutes because of the relatively sharp dip in the profile. This interval, however, corresponds to the yeast cells entering the third cell cycle where the degree of synchronization with cdc15 was noted to progressively decline. Thus the failure of the Kalman smoother to match these unsynchronized time points is not at all indicative of its lack of rigor or applicability. In actuality it suggests that the Kalman filtering techniques can be exploited to assess the validity of gene expression measurements which is also an open problem in microarray gene expression analysis.

On incomplete profiles, the Kalman smoother matched spot on the actual non-missing data while giving estimates for the missing values that generally followed the trends in the profile. Figure 4 shows the estimates for a representative group of genes from the published cdc15 data set that contained missing time points. As expected the profiles become flatter when the points to be estimated are consecutive. It is expected that as the interval between consecutive time points decreases the Kalman smoother can more finely resolve the actual expression dynamics.

4. Conclusion

It has been shown that a very simple linear dynamical systems model for gene expression profiles can be used to effectively estimate missing values in a time series gene expression matrix. In fact these estimates are consistently better than simple gene wise average imputation, and comparable to an advanced estimation method based on Bayesian principal component analysis that has previously been shown to outperform other commonly used methods.

Our method has the distinct advantage in that a minimal number of assumptions are made and that these assumptions have valid biological meaning. In contrast, current estimation methods are based on parameters, such as the eigenvectors of expression matrices, that hold no clear biological meaning. Furthermore, the simplified linear dynamical model proposed here is ripe for further study and refine-

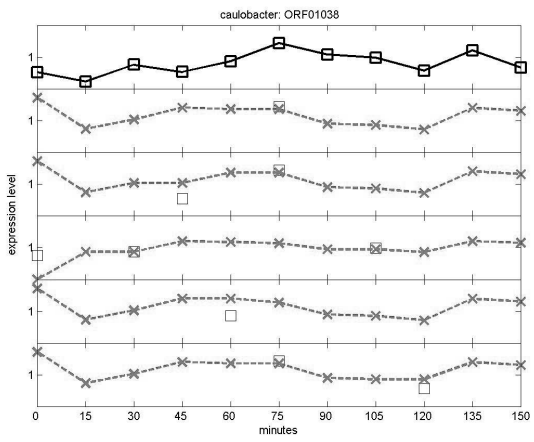

Figure 3. Kalman smoother estimated values for simulated missing time points on a complete gene expression profile from the published caulobacter data. The top panel shows the complete profile. In the subsequent panels, the time points rendered missing are indicated by black squares. The complete estimated profile is shown as a red dotted line.

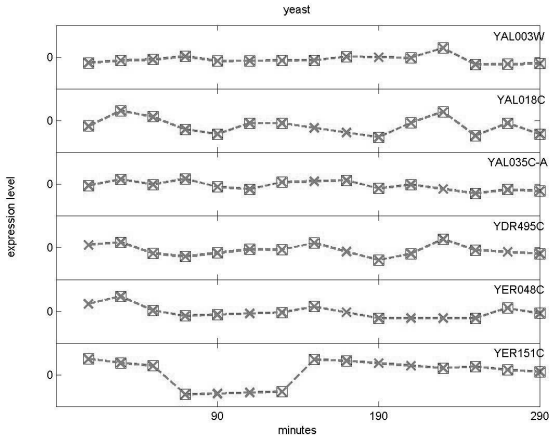

Figure 4. Kalman smoother estimated values for representative incomplete gene expression profiles in the published cdc15 data set. The raw data are represented by black squares and the complete estimated profiles are shown as red dotted lines.

ment that may not only lead to better estimates of missing values but more importantly be exploited to explore underlying network information.

References

[1] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20:2493–2503, 2004.

[2] N. S. Holter, A. Maritan, M. Cieplak, N. Fedoroff, and J. Banavar. Dynamic modeling of gene expression data. *PNAS*, 98:1693–1698, 2001.

[3] N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. Banavar, and N. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *PNAS*, 97:8409–8414, 2000.

[4] J. Hu, H. Li, M. S. Waterman, and X. Zhou. Integrative missing value estimation for microarray data. *BMC Bioinformatics*, 7:449–462, 2006.

[5] M. T. Laub, H. H. McAdamns, T. Feldblyum, C. Fraser, and L. Shapiro. Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, 290:2144–2148, 2000.

[6] P. Maybeck. *Stochastic models, estimation, and control, volume 1*. Academic Press, New York, NY, 1979.

[7] S. Oba, M. Sato, M. Takemasa, K. Matsubara, and S. Ishii. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19:2088–2096, 2003.

[8] J. K. S. A. Rifkin. Geometry of gene expression dynamics. *Bioinformatics*, 18:1176–1183, 2002.

[9] P. Spellman, G. Sherlock, M. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycleregulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Bio. Cell*, 9:3273–3297, 1998.

[10] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17:520–525, 2001.

[11] G. Welch and G. Bishop. An introduction to the kalman filter.