

Reduced-Rank Multivariate Model for Time-Course Microarray Data

Rafal Kustra [†]

[†] Presenting author

Division of Biostatistics, Department of Public Health Sciences
Faculty of Medicine, University of Toronto
155 College, Toronto, ON M5T 3M7, Canada
r.kustra@utoronto.ca

Abstract: In this paper we present a novel, multi-gene approach to time course microarray experiments. One of the advantages of our approach is an explicit modeling of correlation structure among gene expression data. The approach proposed is computationally attractive. We apply the model to the well-known cell-cycle yeast microarray data and present results that compare favorably to the results of the previous studies.

Keywords: Cell-Cycle Expression Data, Penalized linear models, Fourier analysis, microarrays

I. INTRODUCTION

Microarray experiments have become one of the most popular genome-wide technologies. They are widely used in biological labs and in clinical settings, and a plethora of computational and statistical methods have been proposed for the analysis of subsequent data. Much of that methodological research focuses on a relatively simple design, where a set of few (often two) kinds of tissues or samples are compared to each other: for example cancer vs healthy tissue or treated vs control samples.

In 1998, Spellman and colleagues [1] published results on genome wide expression profiling in yeast to investigate cell-cycle development. He observed expression of most of the genes in yeast in different time points during the cell cycle development, and published 800 candidate genes whose expression profile was periodic. There are other microarray experiments besides cell-cycle designs, for example experiments on circadian rhythm in mice [2].

In this paper we propose a new method to analyse time-course expression data. Our method is an adaptation of a general multivariate non-linear model for very high-dimensional data [3]. It enjoys a number of advantages in the present context. For example it is well known that the gene expression values are highly correlated, mostly as a result of co-regulation. This suggest that gene effects should be estimated jointly to increase efficiency, and our multivariate model relies on penalized estimates of gene-gene covariance matrix to

“borrow-strength” across genes. We rely primarily on non-parametric (bootstrap) inference, which is made possible by our fast, dual-space algorithm.

Our model is applied to data from [1] and we compare our results with those obtained by Spellman et al., as well as by [4] in their recent study. Finally we suggest some possible extension for future research.

II. CELL-CYCLE YEAST EXPRESSION DATA

In this paper we develop a new model for cell-cycle and other time-dependent microarray data and apply it to previously analyzed experiment [1], [4]. In this section we briefly describe the data; for full description of the experiment please refer to [1]. A large amount of yeast solution is prepared and time-synchronized using bio-chemical agents. This insures that most of the cells in the solution are in the same developmental stage, although the synchronization will become progressively weaker as the experiment progresses. In [1] two synchronizing agents were used: α -factor and `cdc15`, and they also incorporated data from a previous, similar experiment [5] who used a third synchronizing agent, `cdc28`. In each of the three experiments, a portion of the solution was microarrayed at a number of time points. After pre-processing the data, there are 18 data samples for α -factor experiment (every 7 min for 119 min), 24 samples for `cdc15` experiment (every 10 min for 280 min with some time points omitted), and 17 samples for `cdc28` experiment of [5] (every 10 min for 160 min). Each such sample comprises 6178 log-normalized expression ratio numbers, indicating the relative expression level of each of 6178 yeast genes. (Some of these correspond to Open Reading Frames (ORF) that have not been proven yet to code for genes, but for simplicity of exposition we will refer to all of the as genes). From a purely analytic perspective we have 3 sets of 6178 individual time series, each with about two dozen time-points.

At the time of publication of [1] 104 out of 6178 genes were known to have been involved cell cycle regulation and the goal of this genome wide project is to determine the remaining genes that are involved. The basic premise is that involved genes will show a strong periodic expression pattern with a

period roughly equal to cell division time (about 66min for α -factor data).

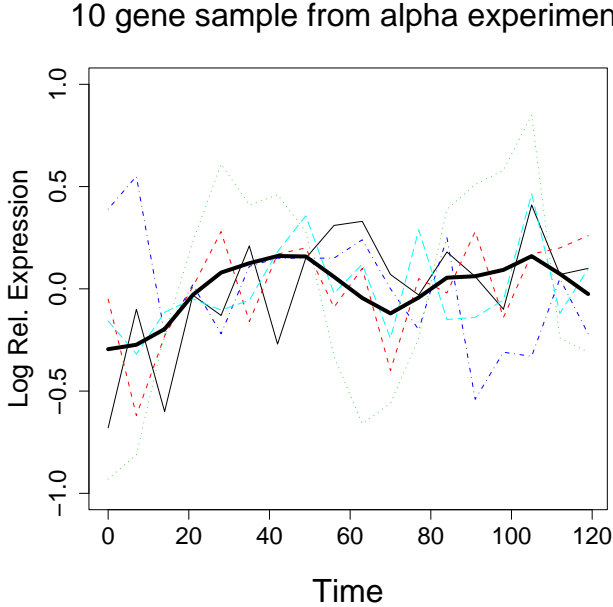


Fig. 1. A sample expression profile of 10 random genes from an α -factor experiment. A thick line is a smoothed average of the 10 curves.

Figure 1 shows 10 randomly chosen genes and their expression time series for the α -factor experiment. One can notice a fairly large range of amplitudes and shapes. Some individual curves and the mean curve (thick black) show a weak periodic pattern: for this experiment one expects two complete periods of expression for genes which are involved in cell-cycle regulation.

III. SOME PREVIOUS APPROACHES TO CELL-CYCLE DATA

In their original paper, [1] the authors employ a method based on gene-wise Fourier analysis. For each gene, they obtain two Fourier coefficients (for sin and cos basis) each an average of coefficients around a postulated cell cycle frequency. For example, for α -factor study they average the coefficients for periods in between 55 and 77 min. The phase parameters for other experiments are fitted by maximizing a total score (sum of squared Fourier coefficients) of known genes. Finally a CDC score is derived for each gene as a sum of all six Fourier coefficients squared, after some rescaling of *cdc15* and *cdc28* coefficients using weights derived from the known genes. The genes are ranked according to the CDC score and a threshold is chosen so that it includes 95 (91%) or known genes. This results in 800 overall, or 705 new candidate genes.

Luan and Li [4] proposed a B-spline based method for teasing periodicity. First a common periodic function is derived using B-spline basis (constrained to be periodic at the ends) with a shift parameter, and a fit obtained, using the 104

known (“guide”) genes. Then, using the common function, a separate phase and amplitude is fitted to each gene separately. The amplitude parameter is tested for being non-zero using a Likelihood Ratio test after Gaussian distribution with AR(1) covariance structure is proposed for the time series errors. They used p-value of this test and False Discovery Rate threshold of 0.5% to obtain 1010 candidate genes, including 89 (85%) of the known genes.

IV. MULTIVARIATE MODEL FOR CELL-CYCLE MICROARRAY DATA

In this section we develop a multivariate model for cell-cycle and other time-course microarray data, which is based on the mR3 statistical modelling framework described in [3]. This model has three advantages:

- It models the whole microarray expression vector as a single sampling unit. Most of the previous attempts use a single gene as a sampling unit even though the experimenters collect a whole microarray of observation in one experiment.
- It attempts to “borrow strength” when modelling gene effects by utilizing the estimated gene-to-gene covariance matrix. It is well known that expression observation across genes show consistent correlation patterns due to expression co-regulation in a cell.
- It offers very fast dual-space algorithm for estimation which makes non-parametric inference by bootstrap possible. The advantage of using bootstrap is that it employs minimal distributional assumptions.

Let \mathbf{y}_{t_i} denote an M -vector of expression values observed across M genes at time t_i , where $i = 1, 2, \dots, n$. In cDNA microarray data one can only observe relative expression values and these are typically log-transformed and perhaps processed further to make comparison across microarrays possible. We assume that \mathbf{y}_{t_i} denotes a vector of such processed and log-transformed data, as necessary, from the i -th microarray. We postulate a functional model for \mathbf{y}_{t_i} :

$$\mathbf{y}_{t_i} = \boldsymbol{\mu} + \mathbf{f}(t_i) + \boldsymbol{\epsilon}_i \quad (1)$$

where $\boldsymbol{\mu}$ is an overall mean and \mathbf{f}_i is an M -valued function of time. We will assume that the microarray data has been centered and will design the common function $\mathbf{f}(t)$ to be zero-mean and thus denote departures from the over-all mean. Therefore parameter $\boldsymbol{\mu}$ plays no role and is omitted from further derivations.

The last term, $\boldsymbol{\epsilon}_i$, denotes zero-mean error term with an $M \times M$ variance-covariance matrix whose estimate is:

$$\text{Var}(\boldsymbol{\epsilon}) = \Sigma_{\kappa_1} = \hat{\Sigma} + \kappa_1 I \quad (2)$$

The covariance matrix is a penalized version of a standard sum-of-squares estimate:

$$\hat{\Sigma} = 1/n \sum_{i=1}^n \left(\mathbf{y}_i - \hat{\mathbf{f}}(t_i) \right) \left(\mathbf{y}_i - \hat{\mathbf{f}}(t_i) \right)^T \quad (3)$$

Here $\hat{\mathbf{f}}(\cdot)$ is an estimate of \mathbf{f} . Both \mathbf{f} and Σ are estimated jointly. The hyperparameter κ_1 , which controls the amount of fitting done to estimate the covariance structure, is considered fixed and can be estimated separately by cross-validation or bootstrap. In this report we have simply chosen a reasonable value for κ_1 , as explained in section IV-C.

A. Reduced-Rank Assumption and Eigen-structures

The mR3 framework has two mechanisms to deal with very high dimensional data. One of them is penalization of the covariance matrix (eq. (2)) and the other is the reduced-rank assumption on the mean function. Unconstrained, function $\mathbf{f}(\cdot)$ has a separate component for each gene. We are seeking to constrain this over-flexibility by R latent structures, $\boldsymbol{\theta}_r$:

$$\mathbf{f}(t) = \sum_{r=1}^R \gamma_r(t) \boldsymbol{\theta}_r \quad (4)$$

This constrains the rank of the fitted mean matrix as follows. In practice function $\mathbf{f}(\cdot)$ will be evaluated at the observed p time points. Let the $M \times p$ matrix B denote the values of the fitted function $\mathbf{f}(\cdot)$ at each time-point. Unconstrained, its rank is at most $\min(M, p)$ which equals to p in our case. The above constraint restricts the rank of matrix B to R since B can be now written as:

$$B = \Theta \Gamma \quad (5)$$

where Γ is an $R \times p$ matrix, with $[\gamma_r(t_1), \dots, \gamma_r(t_p)]$ as its r -th row, and Θ is an $M \times R$ latent structure matrix with $\boldsymbol{\theta}_r$ as its r -th *column*. This relationship can also be written in another, illustrative way. Consider a time series for gene j :

$$\mathbf{z}_j = [y_{t_1;j}, \dots, y_{t_p;j}] \quad (6)$$

Then we have that:

$$\mathbf{z}_j(t) = \sum_r \boldsymbol{\theta}_{r,j} \gamma_r(t) \quad (7)$$

so that elements of $\boldsymbol{\theta}_r$ are seen to be gene-specific amplitudes of an r -component of common periodic function, $\gamma_r(t)$.

B. Expansion of the common mean function

The mr3 framework described in [3] used cubic splines to model a smooth function \mathbf{f} . Since our present goal is to model periodicity, we have chosen a Fourier basis instead. We have data from multiple studies where, to a first order of approximation, cell-cycle genes can be assumed to have the same period but different phase. We use a common $\cos(\cdot)$ basis but separate $\sin(\cdot)$ basis for each study as an approximate way to model different phase (since phase equals to $\tan(-a/b)$, where a, b are frequency-paired coefficients of $\cos(\cdot)$ and $\sin(\cdot)$ functions, respectively). Thus modifying a coefficient on $\sin(\cdot)$ function allows for different phases, although it also changes the shape of the fitted function. A proper solution would be to fit the phase parameter directly, which was done in [4] using an grid minimization process, but this is computationally expensive. We will explore that option in future research.

With the reduced-rank assumption (eq. (5)) the mean function $\mathbf{f}(t)$ depend on time only through the periodic function $\gamma(t)$. We further expand $\gamma(\cdot)$ in a Fourier basis:

$$\begin{aligned} \gamma_r(t; s) = \sum_{\omega=1}^{p/2-1} \{ \alpha_{\omega,r} \cos(2\pi\omega t/p) + \beta_{\omega,r,s} \sin(2\pi\omega t/p) \} \\ + \beta_p \cos(\pi t) \end{aligned} \quad (8)$$

where s denotes a study, and p is a common number of points in each time series, which is assumed even. We will achieve common time-points by smoothing spline fit in next section.

This is a full Fourier basis set that fits all available degrees of freedom for each dimension r . Instead of truncating it at some artificial value we penalize this basis using a roughness penalty. We are fitting coefficients α, β , and hence $\gamma(t)$ by Penalized Least Squares with penalty functional:

$$\mathcal{P}_{\kappa_2}(\gamma) = \kappa_2 \sum_s \int_0^{t_{\max}} (\gamma''(\tau, s))^2 d\tau \quad (9)$$

which is a familiar second-derivative penalty used widely in cubic spline smoothing. This translates ([6]) into a diagonal penalty matrix on coefficients $\alpha_\omega, \beta_\omega$ with diagonal terms equal to $\kappa_2(2\pi\omega)^4$. Here κ_2 is another hyper-parameter which controls the bias-variance tradeoff and can be expressed as an Equivalent Degrees of Freedom of the regression fit ([3]).

Penalizing a Fourier basis is a convenient approach for three reasons. First it allows flexible functional form (no need to truncate basis) but gives a larger preference to a smoother fit: this is evident by weights which are increasing with frequency. Second, the diagonal penalty matrix allows for very fast computation times. Third, our interest is in genes whose expression is periodic and we expect the second harmonic ($\omega = 2$) to be the most indicative of cell-cycle involvement since all three studies cover about two cell-division times. Having the diagonal penalty weight first two harmonic most heavily, we can expect to recover interesting behaviour without fully imposing it on the data but without diluting the results too much with uninteresting (i.e., high-frequency) patterns.

C. Eigen-structures and Degrees of Freedom

It is more convenient to parametrize the model in terms of *eigen-structures* ([3]), $\boldsymbol{\xi}_r$ (see eqs. (2), (4)):

$$\boldsymbol{\xi}_r = \Sigma_{\kappa_1}^{-1} \boldsymbol{\theta}_r \quad (10)$$

This makes interpretation easier since:

$$\begin{aligned} \text{Var} \mathbf{Y}^T \boldsymbol{\xi}_r &= 1, & \text{and} & & (11) \\ \text{Cov}(\mathbf{Y}^T \boldsymbol{\xi}_r, \mathbf{Y}^T \boldsymbol{\xi}_q) &= 0, & r &\neq q & \end{aligned}$$

where \mathbf{Y} is a random vector representing one microarray experiment. The first part of the equation shows that the elements of eigen-structures are rescaled to compensate for varying variances of individual genes. The second part indicates that different eigen-structures can be hoped to discover distinct phenomena.

It is worth noting that the eigen-structures are obtained by the generalized-eigendecomposition and are thus determined up to a sign. Further the eigen-structure/periodic curve pairs are ordered. The first pair is the one that explains the most variability in the data, the second one explains the most variability subject to the orthogonality constraints (11), etc. In fact the pairs are the generalization of Canonical Correlation vectors [3], and can be seen as pairs of functions (one, the eigen-structure, linear in expression data; the other, periodic curve, non-linear periodic in time) that best correlate with each other, subject again to orthogonality constraints to previous pairs.

Reparameterizing the model with eigen-structures has one more benefit in terms of interpreting the results. One can show [3] that the estimation can be cast as a penalized double regression problem:

$$\operatorname{argmin}_{\alpha_{\omega,r}, \beta_{\omega,r}; \xi_r} \sum_{i=1}^n (\mathbf{y}_{t_i}^T \xi_r - \gamma_r(t))^2 \quad (12)$$

subject to constraints (11) and (9). This also has a benefit of providing us with a well-understood framework for penalty hyper-parameters: κ_1, κ_2 . When viewed as a regression of $\mathbf{y}^T \xi_r$ on Fourier-expanded $\gamma_r(t)$, the κ_2 parameters controls the smoothness of the resulting regression function, $\gamma_r(t)$. When viewed as a regression of $\gamma_r(t_i)$ on y_{t_i} , κ_1 acts as a ridge-regression parameter controlling the variability of coefficient vector, ξ_r . This allows us to use a standard smoothing practice ([7], [3]) and express both hyper-parameters as *Equivalent Degrees of Freedom*, or EDF, which are more intuitive than unscaled smoothing parameters like κ_1, κ_2 . In this research we have not optimized them numerically, but have subjectively picked 7 degrees of freedom to model both the covariance structure (and hence eigen-structures) and time association function, $\gamma_r(t)$, for each r .

D. Obtaining Candidate Genes

The exposition so far has centered on deriving an efficient model for massively parallel and noisy times series data, $y_j(t_i)$ with a set of observation for each gene j at timepoints, t_i . This section highlight our proposed method for choosing candidate genes that exhibit significant periodic behaviour and may be investigated further as putative cell-cycle regulatory agents. It is important to underline that our method as described so far does not make a uses of “guide” status of genes which are known to be involve in cell-cycle regulation. This is in contrast to the previously proposed methods that we summarized in section III and will allow us to evaluate the performance of our method in unbiased way.

We propose to pick candidate genes by examining the loading factors of “interesting” eigen-structures, ξ_r . If an j -th component of ξ_r , which we denote by $\xi_r(j)$ is significantly large, in absolute value, this means that it has a significant association with time variable, an association modelled by the r -th periodic curve, $\gamma_r(t)$. This is since it can be viewed as a rescaled regression coefficient of:

$$E\gamma_r(t)|\mathbf{y}_t \sim \xi_r^T \mathbf{y}_t \quad (13)$$

from equation (12), subject to:

$$\xi_r^T \Sigma_{\kappa_1} \xi_q = \delta_0^1 \quad (14)$$

where δ_0^1 is a Kronecker delta equal to 1 for $p = q$, and zero, otherwise. Thus if for a moment we think of $\gamma_r(t)$ as a fixed reparameterization of time, we can think of ξ_r as regression coefficients of fitting this reparameterization of time onto relative gene expression levels. The genes most associated with this time pattern will show larger coefficients, $\xi_r(j)$.

Now, the periodic curves $\gamma_r(t)$ are jointly estimated with ξ_r to maximized variability of the data explained by the model. Hence we can subjectively examine each periodic curve and concentrate our attention on these which seem most connected to the cell-cycle behaviour. In our context, this means examining the Fourier power and looking for larger concentration in the second harmonic.

Once interesting eigenstructures are chosen, we simply rank genes, for each r , which have the highest absolute loadings in ξ_r . This is plausible since the eigen-structures are rescaled by the gene covariance matrix, although a more statistically sound method is probably worth researching.

E. Computational Algorithm for Fitting mR3 models

Our computational strategy closely follows that outlined in [3] paper. Here we just briefly mention few notational equivalents and spell out main points of the strategy.

In the notation of that paper, we have one predictor, time. As mentioned before, instead of cubic splines used in the original paper we are using Fourier basis. Further since we are dealing with multiple studies where period can be assumed same, but where phases may differ, we expand the periodic functions, $\gamma_r(t)$ (called *association curves* in [3]) into a basis with common $\cos()$ but separate $\sin()$ components. If there are p unique time points and s studies, the design matrix, X , in the notation of [3], will have $(p/2) + s(p/2 - 1)$ columns where the first set will be comprised of $\cos()$ functions and the last s sets will be $\sin()$ functions evaluated only for the arrays of the respective study, and zero elsewhere.

The computational strategy outlined in [3] avoid the explicit construction of the covariance matrix, Σ_{κ_1} which would otherwise be a great computational burden (with over 36 million elements for our data). This is accomplished by an innovative dual-space algorithm. The mR3 algorithm requires instead that the smoother matrix, $X(X^T X + \kappa_2 \Omega)^{-1} X^T$ be directly evaluated, where here the columns of X matrix are evaluated Fourier basis functions, as mentioned above, and Ω is a penalty matrix. Since our basis functions are orthogonal and penalty matrix is diagonal (see eq. (9) the following discussion) this provides an additional simplification to the algorithm.

V. RESULTS WITH YEAST CELL-CYCLE DATA

We have used the original dataset analyzed in [1] which is available online. The authors have already pre-processed the data by taking log-ratios, normalizing the data from individual microarrays to common values and centering the time-series. For the purposes of this paper, we have restricted our attention

to two out of the three studies: α -factor and *cdc28*, since they had the longest time series. A few further pre-processing steps were applied to the data. First we have imputed missing values using an `knnimpute` method described by [8]. The resulting data was centered to avoid the need to estimate the common mean in equation (1). Finally we have put the two sets of time series on a common basis by the following process:

- 1) fit generous 10 EDF spline to each gene time series in each study
- 2) interpolate at 0-119min (7min int): 18 points

Hence we have $p = 18$ and $s = 2$ in the notation of the previous sections.

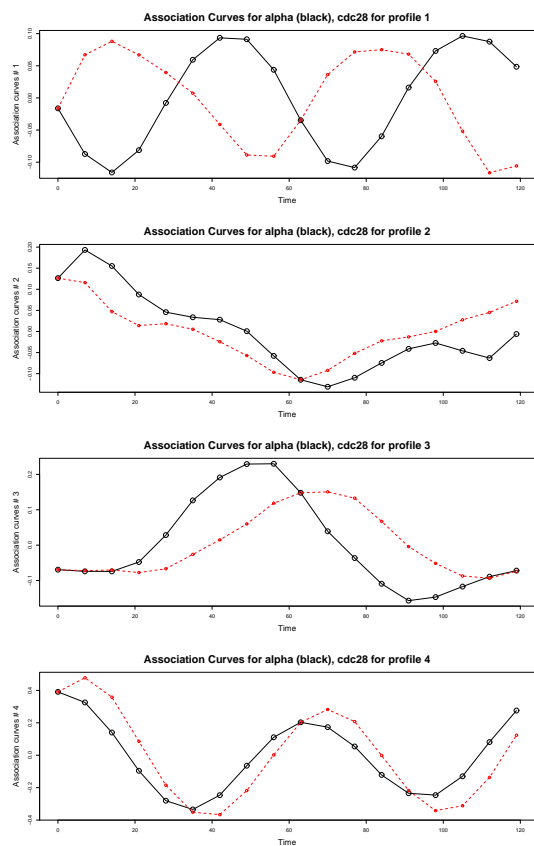


Fig. 2. Periodic curves for $r = 1 - 4$, left to right, top-to-bottom. The black curve is for α -factor study, and the red one for *cdc28* study.

Figure 2 shows the first four periodic curves, $\gamma_r(t)$ for each of the two studies. Re-reassuringly the first curve, which explains the largest amount of variation in the data, shows a clear two-period pattern that was expected for this study. This is notable since neither the information on two expected periods, nor the indicators for “guide” genes were available to the method, so this discover is purely data-driven. The two versions of this curve, corresponding to the two studies, seem in complete anti-phase and with slightly different periods. This could be an artifact of the model and warrants further research.

Perhaps more revealing are the power bars, which show a contribution of each harmonic of $\gamma_r(t)$ to the overall power,

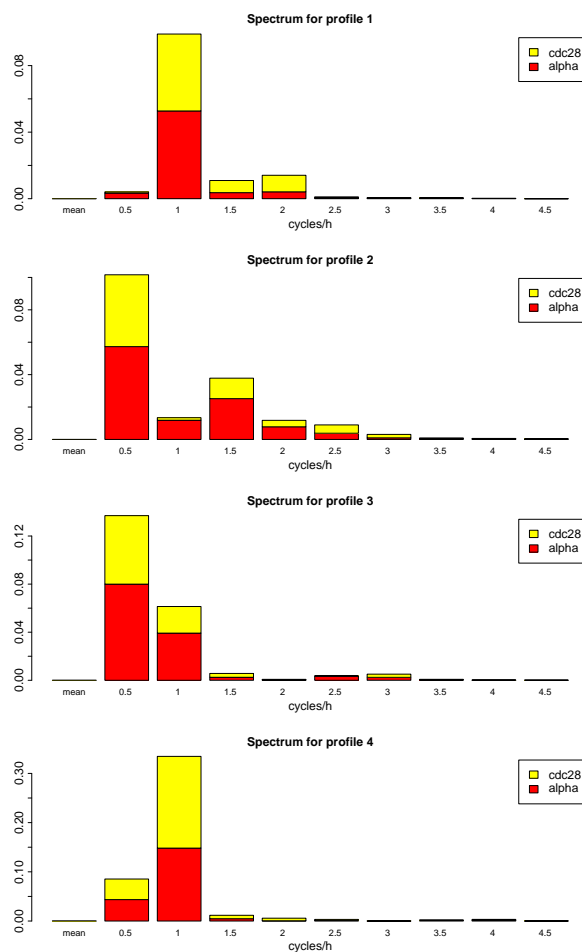


Fig. 3. Components of frequency contribution to overall power for the first four periodic curves. Red is for α -factor study, and yellow for *cdc28* study.

for $r = 1, \dots, 8$. The first bar is for overall mean which was constrained to be zero in our expansion. The third bar represents power in the second harmonic, the most interesting for us. It shows that the first, fourth and seventh curves have the highest concentration of two-cycle periodic component, and that the second and perhaps fifth curves have noticeable contributions as well. While a more formal approach to testing these contributions is possible, for this report we have chosen to concentrate on exploratory measures. Figure 4 attempts, in an informal way, to validate our approach. Since, contrary to other studies, we have not used the status of the guide genes, we can examine the concentration of the guide genes in candidate genes picked by eigen-structures as an unbiased way to guide the success of our method, as well as to understand various trade-offs necessary for this high-dimensional noisy datasets. One complication is that, while we are reasonably sure that the guide genes are involved in cell-cycle regulation, since their involvement has largely been confirmed by less variable biological experiments, we have no way of knowing which if the remaining genes are true positives or negatives. Figure 4 attempts to quantify the results by showing how

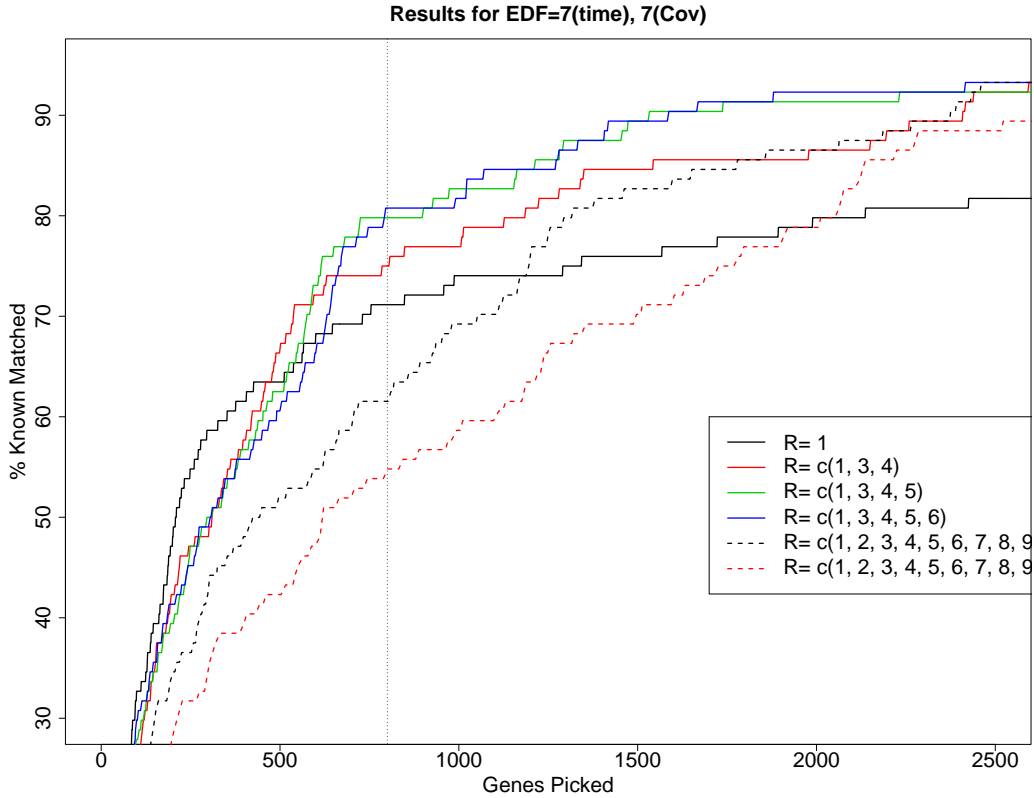


Fig. 4. The percentage of guide genes above given percentile threshold on the loadings of eigen-structures for various thresholds and subsets of “interesting” eigen-structures

many of the guide genes are picked versus how many candidates are picked with varying thresholds on the loadings of eigen-structures (Section IV-D) and various combinations of “interesting” eigen-structures chosen. The candidate genes are picked from various eigen-structures by first setting a common percentile threshold, α , then taking the top $M \cdot \alpha/2$ (for positive loadings) and bottom $M \cdot \alpha/2$ (for negative loadings) genes, and the taking a union over all “interesting” eigen-structures. While we expect to pick up a significant number of new candidate genes even while missing some guide genes, we hope that we will pick a vast majority of guide genes for reasonably small thresholds. For example, for threshold $\alpha = 1$ we are guaranteed to pick up 100% of guide genes but we do it at the expense of any specificity since all other genes are new candidate genes. This kind of curves are sometimes called hit-curves in rare-target discovery research, such as for drug discovery experiments.

Comparing the results of figure 4 to the results of the two previous study described, we note that for a choice of threshold which results in 800 candidate genes, same as a corresponding number in [1], we match 81% (or 84) of the guide genes, compared to 91% for Spellman et al study. This is notable since their method made an extensive use of the “guide” status, thus likely leading to an over-estimate of their success rate. Comparing with their results, we also note that while we

agree with Spellman et al results in about a half (384) of the 800 candidates, we tend to agree much better for candidates that have higher CDC scores according to their methods. For instance, we match 165 of their top 200 candidates, but only 50 of their bottom 200.

Figure 4 also shows the advantage of reduced-rank approach to this data. We see that strategically choosing the eigen-structures provides a benefit over choosing all of them or cutting the dimensionality at an arbitrary point. Thus the best results are for sets of eigen-structures that include first, third, fourth and fifth, and perhaps sixth eigen-structures. Choosing all the first 10 structures (black-dotted line) or the first 25 eigen-structures (red-dotted line) leads to significantly worse performance.

VI. DISCUSSION AND CONCLUSIONS

We have presented a multivariate model for time-course microarray expression data. Our model has an advantage of incorporating gene-gene correlation structure into the estimation, which may result in more efficient estimates of time function and gene effects. We have briefly summarized the model, which is a special application of a general framework presented in [3]. This paper outlines the specific choices needed to model the time course data and to test which genes show a strong periodic behaviour that may be indicative of an important role in cell’s developmental cycle. We have

presented an initial result of applying our model to the well-studied cell-cycle yeast expression data. While our model was not heavily optimized, it still shows very promising results as compared with previous studies. Another advantage of our approach is that the status of the “guide” genes, whose role in cell cycle has previously be ascertained, is not used in the analysis which enables an unbiased validation of our approach. This is in contrast to two other studies who utilize the guide genes extensively and are thus unable to validate their results without external data.

REFERENCES

- [1] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher, “Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization,” *Molecular Biology of the Cell*, vol. 9, pp. 3273–97, 1998.
- [2] K. Storch, O. Lipan, I. Leykin, *et al.*, “Extensive and divergent circadian gene expression in liver and heart,” *Nature*, vol. 417, pp. 78–83, 2002.
- [3] R. Kustra, “Reduced-rank regularized multivariate model for high-dimensional data,” *Journal of Computational and Graphical Statistics*, 2006, in press.
- [4] Y. Luan and H. Li, “Model-based methods for indentifying periodically expressed genes based on time course microarray gene expression data,” *Bioinformatics*, vol. 20, no. 3, pp. 332–9, 2004.
- [5] R. Cho *et al.*, “A genome wide transcriptional analysis of the mitotic cell cycle,” *Molecular Cell*, vol. 2, pp. 65–73, 1998.
- [6] G. Wahba, “Bayesian “confidence intervals” for the cross-validated smoothing spline,” *J. Royal Stat Soc*, vol. 45, no. 1, pp. 133–150, 1983.
- [7] T. Hastie and R. Tibshirani, *Generalized Additive Models*. Chapman and Hall, 1990.
- [8] O. Troyanskaya, M. Cantor, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, “Imputing missing data for gene expression arrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–25, 2001.