Modeling periodic DNA microarray data by multivariate support vector regression and non-linear curve fitting

Matthew D. Boardman [†]

[†] Computational Neuroscience Group, Dalhousie University, Halifax, NS, B3H 1W5, Canada

ABSTRACT

Motivation: Noise levels and cross-study variation present in gene-expression data from DNA microarray experiments create obstacles for genomic researchers. A reliable method for modeling such data is required in order to impute missing observations, to fuse multiple, similar experiments and to improve the signal to noise ratio in time-variant or cross-study experimental results. In this paper, we combine multivariate support vector regression and non-linear, periodic curve-fitting methods to model differential gene-expression in periodic microarray data. Support vector regression makes no assumptions about the distribution of the underlying data model and seeks a general, regularized solution for data sets with a low number of samples but with high dimensionality. The cleaned microarray data from this model may then be analysed further through other methods. In this paper we apply an additive, periodic model to detect regular periodicity, such as the transcription of mitotic genes during a cell's reproductive cycle.

Results: We apply these methods to a recent study of cell-cycle synchronization methods in fission yeast, *Schizosaccharomyces pombe*, and evaluate the models in comparison to univariate polynomial and linear regression approaches common for microarray data imputation. In this study, our goals are to impute missing data points in a periodically meaningful context, to determine which genes exhibit high periodicity that is strongly synchronized with the cell-cycle and to determine the statistically most-likely activation point of each gene within the cycle.

Availability: The methods implemented by the author in MATLAB (Mathworks) are freely available upon request. Supplementary material is available from the author's website at http://www.cs.dal.ca/∼boardman.

Contact: Matt.Boardman@dal.ca

1 INTRODUCTION

In this paper, we examine the modeling of differentially-expressed microarray data without regard to the underlying cause of errors. Our approach is based on two data models: first a multivariate, kernel-based, non-linear regression using Support Vector Machines (SVM); and second a univariate, maximum-likelihood model using non-linear curve-fitting, with both a linear and non-linear, periodic component. Our goal is to create models which can be used to impute missing observations, for data correction and normalization, or which can in themselves be used for further analysis.

Noise and errors arise in DNA microarray hybridization experiments from a variety of factors, including gene-specific dye bias (Martin-Magniette *et al*., 2005), probe and experiment design (Smyth *et al*., 2003), culture heterogeneity (Gilks *et al*., 2005), variations in slide quality and manufacturing processes which can create surface abnormalities or allow slide-movement within the microarray scanner (Agilent Technologies, 2005) and the normal deterioration of mRNA post transcription (Orengo *et al*., 2003). A comprehensive overview of many of these sources of experimental and analytical errors, and common statistical techniques used to overcome them, can be found in Smyth *et al*. (2003).

Many other statistical approaches have been applied to microarray data analysis, such as univariate or multivariate Analysis of Variances models (ANOVA or MANOVA) (Gilks *et al*., 2005; Kerr *et al*., 2000) commonly used for normalization prior to further analysis (Smyth *et al*., 2003). Independent Components Analysis (ICA) (Martoglio *et al*., 2002) appears to have significant potential for automatic artefact isolation and removal, by maximizing the statistical independence of the resulting signals.

Univariate or multivariate regression approaches are common in microarray analysis (Tsai *et al*., 2004; Wu, 2005). The majority of these are univariate, in which a single output variable is targeted (although there may be one or many input dimensions). In addition to reducing input noise and sources of error, univariate regression models have been used for classification and prediction (see e.g. Choi *et al*., 2003; Wu, 2005). More recently, multivariate regression approaches, in which the output is a vector rather than a single value, have also been suggested. Gilks *et al*. (2005) proposed a multivariate, linear regression technique based on a controlled design matrix to fuse data from multiple, similar experiments, in which the output is a fused, cleaned, time-variant microarray experiment for periodic data. Choi *et al*. (2003) also fuse multiple, time-variant data sets using a covariance measure for Bayesian meta-analysis and apply their algorithm to the problem of cancer profiling. Johansson *et al*. (2003) also used a multivariate approach, applying an algorithm based on Partial Least Squares (PLS) to fuse time-variant data sets for budding yeast, *Saccharomyces cerevisiae*: the authors of this study note that an advantage of PLS is to obtain models with high generalization performance for data sets with few observations but high dimensionality, which as we note below, is also a significant advantage for SVM as used in this paper. Tai and Speed (2004) have proposed a Bayesian approach of similar form, but which uses a Bayesian statistic to approximate this design matrix automatically and from which the goal is to create a vector of expression values for each individual gene probe, similar to the approach taken here.

Additive data models have also been proposed, such as Tsai *et al*. (2004) in which a linear and non-linear model are combined, an approach similar to the periodic model we apply in this paper, but using a sequential normalization algorithm rather than a non-linear maximum likelihood model. A non-linear maximum likelihood model was proposed in Huber *et al*. (2002), in order to normalize data prior to more complex analysis, however such non-linear transformations will negate the assumption of an additive error or residual component (Gilks *et al.*, 2005).

Data imputation methods attempt to find the most likely value of missing observations and minimize noise levels in the gene expression data to identify the most-likely underlying signals. However, in some cases it may not be necessary or warranted to identify each individual gene-expression signal in the data: we may simply wish to measure the goodness-of-fit to a simple, additive, periodic model in order to isolate a particular component of the underlying signal relevant for a particular biological analysis, or to impute missing pieces of source data in a periodically meaningful way rather than employing a statistical averaging technique such as K-Nearest Neighbors (KNN) (Gilks *et al.*, 2005) which is known to perform badly in data imputation in terms of RMSE (Troyanskaya *et al.*, 2001; Wang *et al.*, 2006). A comparison of several methods for data imputation is provided in Jörnsten *et al.* (2005), who provide a new method based on convex linear combination of several current methods, and in Troyanskaya *et al.* (2001), who compare SVD and KNN with a simple row-oriented mean for several real data sets.

Further analysis of the cleaned data set to identify periodically expressed genes during cell processes to identify mitotic genes has been performed via clustering (Rustici *et al.*, 2004) and Singular Value Decomposition (SVD) (Gilks *et al.*, 2005) or Principal Components Analysis (PCA) (Johansson *et al.*, 2003) which find signals with maximum variance, although a more common approach is to use a statistical ranking techniques such as the commonly used t-statistic (Smyth *et al.*, 2003). Here we will use the mean squared error (MSE), normalized by the variance, as a goodness-of-fit statistic.

1.1 Support Vector Regression

Although supervised machine-learning techniques, such as Artificial Neural Networks (ANN) and SVM, have been used for classification in microarray research (see e.g. Brown *et al.*, 2000), regression models based on ANN and SVM are less common for microarray regression models; in particular SVM, as the technology is comparatively recent (Burges, 1998; Smola and Schölkopf, 2004; Vapnik, 1999).

In binary classification, SVM are a supervised machine-learning technique using a statistical approach to maximize the margin between samples of opposite class, often using a non-linear kernel to provide a dot product between vectors in high-dimensional mapped space (Vapnik, 1999) in order to maximize the separation between non-separable sample data (Burges, 1998). In this paper we use the non-linear Radial Basis Function (RBF) kernel function throughout, as this has repeatedly been shown to work well with non-separable data sets for a wide variety of applications (see e.g. Boardman and Trappenberg, 2006; Chang and Lin, 2001). ϵ-tube Support Vector Regression (ϵ-SVR) extends the SVM statistical-learning philosophy to regression or function estimation problems, by assuming a constant noise threshold ϵ and penalizing samples outside this threshold using a cost parameter C (Smola and Schölkopf, 2004; Vapnik, 1999). A quick and painless introduction to SVM may be found in Bennet *et al.* (2000). More detailed and mathematically rigorous tutorials on practical application of SVM may be found in Burges (1998) for classification, or in Smola and Schölkopf (2004) for regression.

In Wang *et al.* (2006), Support Vector Regression was shown to be superior to K-Nearest Neighbors (KNN), Bayesian Principal Components Analysis (BPCA) and Local Least Squares (LLS) for data imputation on microarray data sets, in terms of mean squared error normalized by individual gene variance. The free parameters ϵ, γ, C for the ϵ-SVR implementation in this study were determined through a grid search, whereas here we use the simulated annealing method of Boardman and Trappenberg (2006) which enhances generalization performance by using a model complexity penalty. These optimum parameters were determined individually for each column (time point) in Wang *et al.* (2006), whereas here we use a representative sample of genes in order to further enhance generalization and reduce the computational burden. Finally, the column-wise orthogonal input coding scheme to flag missing values in Wang *et al.* (2006) was used to alleviate a restriction that only a single missing value could be estimated for each row (gene) of the data in their implementation, whereas the multivariate ϵ-SVR approach used in this paper allows any number of missing values so long as at least one observation is present. Some extreme examples of this are presented in Fig. 5(c-d).

In terms of a true representation of the data rather than purely mean squared error, we would expect a model based on ϵ-SVR to create the best possible representation of noisy, inconsistent microarray data, even in comparison to neural network approaches, since the advanced regularization capability of these kernel-based methods allows for a highly-accurate model with a relatively small number of observations. The efficiency of the ϵ-SVR algorithm and a small model representation allows for much higher dimensionality in the input observations and output targets than neural network approaches. The importance of generalization performance, as a tradeoff to mean squared error, is illustrated in Fig. 1.

In this paper, LIBSVM (Chang and Lin, 2001) was used as a typical ϵ-SVR implementation, in which high computational efficiency in the quadratic optimization is achieved through Sequential Minimum Optimization (SMO).

2 DATA AND METHODS

In this paper, we will primarily follow the notation of Gilks *et al.* (2005). We define $\mathbf{D}$ as an $N \times m$ observed data matrix of m gene probes taken at N time points. As is generally the case, here $m \gg N$ since for the *elutriation2* data set, $m = 5038$ and $N = 20$. We refer to the individual elements of $\mathbf{D}$ such that for a particular gene i we have we have observations $\mathbf{x_i}$ and targets y_i as

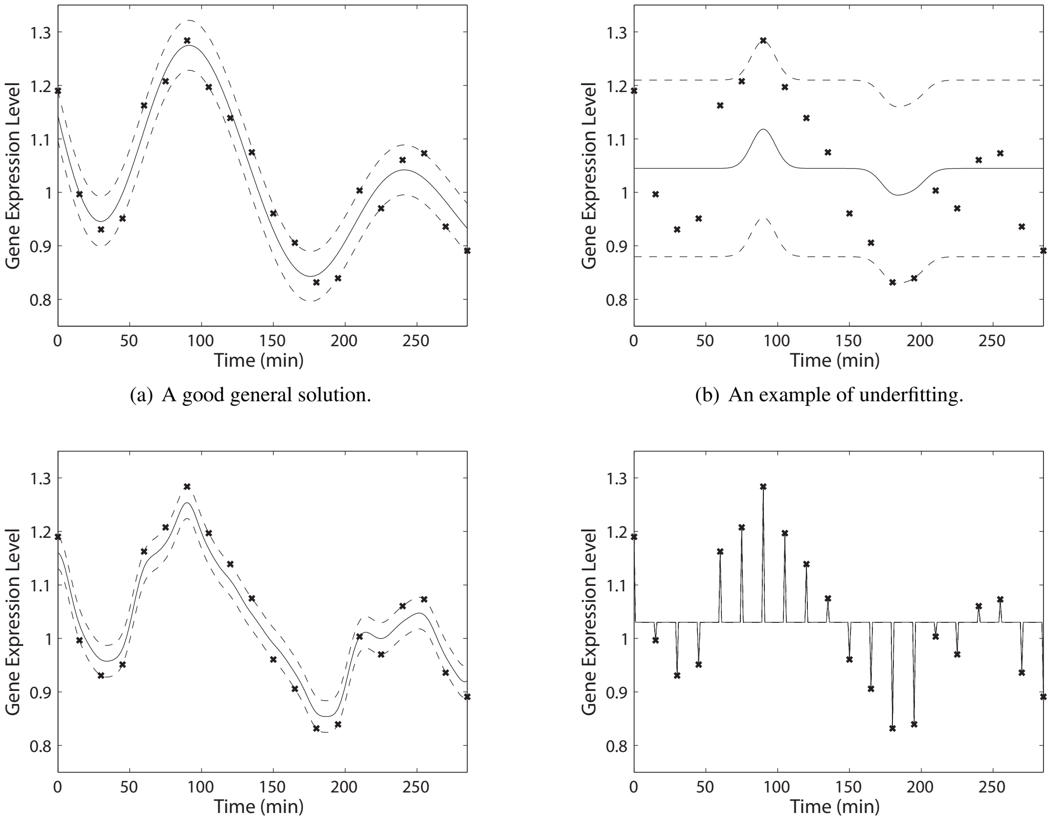
$$(\mathbf{x}_{i1}, y_{i1}), \ldots, (\mathbf{x}_{i\ell}, y_{i\ell}) \qquad (1)$$

corresponding to row i of $\mathbf{D}$, where $N = \ell$.

2.1 Experimental Data

In Rustici *et al.* (2004), nine different cell-cycle synchronization techniques were applied to *S. pombe* in comparison to unsynchronized cell cultures, including elutriation to isolate fine cells from heavier cells in the culture, or selective blocking and releasing of particular proteins known to control the cell cycle through temperature variation. The raw data from these experiments were made available through ArrayExpress (http://www.ebi.ac.uk/arrayexpress) under accession numbers E-MEXP-54 through E-MEXP-64.

Each experiment in this data set shows the normalized, unitless signal ratio of the experimental culture at each timepoint for each gene, in comparison to that of an unsynchronized, control culture of the same organism, for $> 99.5\%$ (Rustici *et al.*, 2004) of all



(a) A good general solution.

(b) An example of underfitting.

(c) An example of overfitting.

(d) Severe overfitting.

Fig. 1. Inappropriate selection of the free model parameters, or hyperparameters, in ϵ-SVR is likely to lead to improperly modelled data. Here we see four regression models for the C222.06 gene which, incidentally, was not identified as periodic in previous analyses (Gilks *et al.*, 2005; Rustici *et al.*, 2004); this may be due to it's small levels of expression, however our model does find statistically significant periodic activity. (a) An ϵ-SVR model trained using the heuristic in Boardman and Trappenberg (2006) adapted here for regression, which includes a model complexity penalty as described in Sec. 2.2. (b) An example of underfitting: in this case, the ϵ-tube width, corresponding to the expected noise level, is too high. Such a model overgeneralizes, ignoring all but the most extreme values in the training data. (c) An example of overfitting: in this case, the C cost parameter is too high, disallowing outliers such that every observed point must be within the ϵ-tube bounds. A model such as this contains many support vectors and is not likely to generalize well. (d) An example of severe overfitting: in this case the mean squared error is near zero, as the model fits all observed points. However, it is quite obvious that every observed point is a support vector, and so this model will certainly not generalize well to any points between the observed points: such a model would be useless for imputing missing observations.

identified genes in the *S. pombe* genome through several successive hybridizations taken every 15 minutes. A value of one would indicate that the gene has equal levels of expression in the synchronized and unsynchronized cultures. A value higher than one indicates that the synchronized culture exhibits a proportionately higher gene expression level than the unsynchronized control culture. In this study, through a clustering algorithm, 407 genes were identified as periodic and 136 of these were identified as strongly periodic, defined as those whose maximum difference (from peak to trough) was greater than 2. The data in this study was later analyzed using a multivariate linear regression and SVD in Gilks *et al.* (2005).

There are many missing data points in these data sets. For example, in the *elutriation2* data set, 286 of the 5038 genes evaluated on each microarray slide had no data in any of the observed hybridizations, and a further 819 genes had fewer than 75% of the data points available. A total of 16.9% of the observations are missing. It is self-evident that the fewer observations available for any particular gene probe, the less accurate any attempt to model the underlying data will be. However, from the analysis performed in Gilks *et al.* (2005), the *elutriation2* data set appears to exhibit excellent periodicity through nearly two cell cycles, so in this preliminary work, we initially concentrate on this one particular experiment.

2.2 Non-linear Multivariate Regression

We first aim to impute these missing observations, and reduce noise levels throughout the data set, using SVM. Since both the input and output for the SVM is the full, time-course experiment, in this case 5038 genes over 20 time points, our SVM model naturally falls into the category of multivariate regression. However, in the implementation of this model, we use a series of univariate regressions: an ϵ-SVR model for each gene with multidimensional input (the values of the all remaining genes at a specific time point) and unidimensional output (the value of the target gene for a specific time point). 3515 of the 5038 genes were identified as each having 100% of the available data points available. These were used as the training input for the model, and were normalized such that

$$\mathbf{x}_{ij} \in [-1, 1] \;\; \forall \; i \in \{1, \ldots, m\}, j \in \{1, \ldots, N\} \qquad (2)$$

Note that the time vector itself was *not* provided as an input vector for this model; only the gene expression data is used as input. This allows the SVM model to seek a generalized solution in 3515-dimensional input space purely on the basis of exploring the relationships between individual genes. This removes any assumption of the time-variant nature of the gene expression data, and allows us to impute missing observations even if a significant number of observations is missing: for example, see Fig. 5. Of course, the target gene is also excluded from the training data for each model.

An important consideration in the practical application of SVM is the determination of appropriate values for the free parameters, or hyperparameters, of the SVM during training: we wish to determine optimum values for the ϵ-tube width, RBF γ-radius and C cost free parameters. Here we use the heuristic suggested in Boardman and Trappenberg (2006), adapted to ϵ-SVR by adjusting the scoring function as follows.

The goal of this heuristic is to obtain a low mean squared error while still providing adequate generalization performance. Through the course of the three-dimensional, stochastic parameter search in this heuristic, we minimize the cost functional $\mathscr{E}(\epsilon, \gamma, C, \mathbf{D})$ defined as

$$\mathscr{E} = \mathscr{E}_s + \lambda \, \mathscr{E}_c \qquad (3)$$

following the notation of Boardman and Trappenberg (2006), where $\mathscr{E}_s(\epsilon, \gamma, C, \mathbf{D})$ is the 10-fold cross-validation mean squared error (MSE) resulting from a model trained using parameters defined by this point in three-dimensional parameter space, normalized by the inverse of the standard deviation of the target observations, as

$$\mathscr{E}_s = \frac{1}{N\sigma_{y_i}} \sum_{j=1}^{N} \left| y_{ij} - \mathscr{F}(\mathbf{x}_{ij}) \right|^2 \qquad (4)$$

where $\mathscr{F}(\mathbf{x}_{ij})$ is the predicted value of y_{ij} from the regression model. Normalizing in this manner allows those genes with small expression to more significantly affect the mean squared error in comparison to those with relatively high expression. $\mathscr{E}_c(\epsilon, \gamma, C, \mathbf{D})$ is a model-complexity penalty defined by

$$\mathscr{E}_c = \left(\frac{n_{sv}}{\ell} \right)^2 \qquad (5)$$

where $n_{sv}(\epsilon, \gamma, C, \mathbf{D})$ is the number of support vectors in the resulting model representation and ℓ is the total number of training observations. The importance of including a model complexity measure in regression problems is illustrated in Fig. 1.

The regularization parameter λ balances the tradeoff between $\mathscr{E}_s$ and $\mathscr{E}_c$, and the square introduces a non-linearity to more sharply penalize those models which obtain a low mean squared error at the expense of high model complexity. We found that a value of $\lambda = 10$ brought the two terms to the same orders of magnitude, giving roughly equal weight to each.

To balance computational complexity with model accuracy, rather than calculating the optimum free parameters for each of the 5038 models, we minimize the sum of $\mathscr{E}$ taken over ten representative genes at each evaluated point in three-dimensional free parameter space. Each of these genes were identified as periodic by the clustering methodology in Rustici *et al.* (2004). Some had small expression throughout the time course of the data, others had large expression, indicating the need for normalization in the MSE term as shown above. The selected genes were *slp1*, *cdc20*, *h4.2:hhf2*, *sly1*, *h3.1:hht1*, *h3.3:hht3*, *h4.3:hhf3*, *h3.2:hht2*, *h4.1:hhf1* and *bgs4*. The optimum parameters found from the analysis of these ten genes were then used to train the ϵ-SVR for all genes.

The simulated annealing heuristic was employed with a moderate cooling schedule, such that the objective function was evaluated at a total of 4590 points in parameter space. This is contrast to the classification approach in Boardman and Trappenberg (2006) which searched a two-dimensional parameter space for the RBF γ and C regularization parameters for the purpose of binary classification, using either a fast-cooling schedule with 440 evaluations, or a slow-cooling schedule with 6880 evaluations.

We define the matrix of observations cleaned by this multivariate regression as $\mathbf{S}$, of the same dimensions as $\mathbf{D}$.

2.3 Periodic Additive Model

The output from this multivariate analysis is then used as an input to a periodic, additive model. For each gene i, we define the true underlying signal to be the time-variant function $C_i(t)$. The estimate of this signal obtained through the periodic model is defined to be $\hat{C}_i(t)$. The difference between these signals is then the residual $\xi_i(t)$:

$$C_i(t) = \hat{C}_i(t) + \xi_i(t) \qquad (6)$$

We propose that the model of each gene contain both a periodic component and a linear component:

$$\hat{C}_i(t) = \rho_i \sin\left(\frac{2\pi t}{\Lambda} + \phi_i \right) + (t\alpha_i + \beta_i) \qquad (7)$$

where the parameters α_i and β_i define the linear component of the model for each gene i, the magnitude and phase of the periodic component is defined by ρ_i and ϕ_i for each gene i, and the cell-cycle length is expressed by Λ as a constant for all genes within the particular experiment (not to be confused with the λ regularization parameter in Eqn. 3).

In order to determine the model parameters for each gene, we use a non-linear, least-squares curve-fitting procedure provided by the MATLAB Curve Fitting Toolbox (Mathworks). Specifically, we use the robust implementation of the trust-region reflective Newtonian algorithm, which allows us to impose logical bounds on each model parameter. To shake the optimization procedure from local minima, two starting points were used for opposite phase $\phi = \{0, \pi\}$. A purely linear model was also applied. Of these three resulting models, we take that with the lowest mean squared error. For additional detail on the specific implementation of similar curve-fitting procedures, see e.g. Press *et al.* (1992) §15.5–15.7.

We first apply the proposed periodic model (above) to each gene for a particular set of hybridizations, using non-linear, least-squares curve fitting, to find the most likely model parameters for each gene allowing Λ to vary independently for each gene. For this step, we consider only the 407 genes identified as periodic by Rustici *et al.*

Table 1. Comparing mean squared error (MSE) ×1000 resulting from periodic expression models for several genes, from *elutriation2* data set.

Method	slp1	hhf2	bgs4	cdc20
Polynomial Regression (Degree 6)	35.0	42.4	11.0	17.3
Univariate Support Vector Regression†	3.3	13.6	4.3	20.9
Linear Regression	385.1	658.1	37.1	66.5
Periodic Model	24.9	53.6	14.3	35.4

†Note that the data from gene C222.06 was used to find optimum ϵ-SVR model parameters for this univariate analysis.

Table 2. Comparing model parameters found from proposed periodic maximum likelihood model for several genes, from the *elutriation2* data set cleaned through multivariate ϵ-SVR, with fixed cell-cycle length $\Lambda = 153.9$ minutes.

Model Parameter	slp1	hhf2	bgs4	cdc20
ϕ_i : Phase (radians)	0.068	-1.269	-0.198	-0.080
ρ_i : Periodic Amplitude	0.865	1.055	0.197	0.249
α_i : Linear Decay (×1000)	0.064	-0.144	0.039	0.246
β_i : Linear Offset	1.170	1.297	1.098	1.016
$\mathcal{GF}_{SVM}$: NRMSE of SVM Model	0.100	0.089	0.277	0.335
$\mathcal{GF}_{Model}$: NRMSE of Periodic Model	0.202	0.269	0.464	0.469

(2004). Some of these genes were identified as linear by our model, and these were naturally excluded. The median Λ of the remaining genes was taken to be the "true" cell-cycle length.

We then reapply the curve-fitting procedure for all genes, taking this Λ value as fixed.

2.4 Goodness-of-Fit Statistics

As a goodness-of-fit statistic, here we normalize the commonly-used mean squared error by the variance (see e.g. Wang *et al.*, 2006), to define the normalized root mean squared error (NRMSE) of the observations compared to the SVM model as

$$\mathcal{GF}_{SVM}(\mathbf{D}_i, \mathbf{S}_i) = \sqrt{\frac{1}{N\sigma^2(\mathbf{D}_i)} \sum_{j=1}^{N} |d_{ij} - s_{ij}|^2} \qquad (8)$$

where d_{ij} is the j-th observation of the i-th gene, s_{ij} is the j-th prediction of the i-th gene by the ϵ-SVR model, and $\sigma^2(\mathbf{D}_i)$ is the variance of the observations for gene i.

Similarly, to determine how well the periodic, additive model fits the SVM model, we define the NRMSE of the SVM model compared to the periodic model as

$$\mathcal{GF}_{Model}(\mathbf{S}_i, \hat{C}_i) = \sqrt{\frac{1}{N\sigma^2(\mathbf{S}_i)} \sum_{j=1}^{N} |s_{ij} - \hat{c}_{ij}|^2} \qquad (9)$$

where $\hat{c}_{ij}$ is the periodic model's expected value for the j-th observation of the i-th gene. and $\sigma^2(\mathbf{S}_i)$ is the variance of the ϵ-SVR model's predictions for gene i.

3 RESULTS

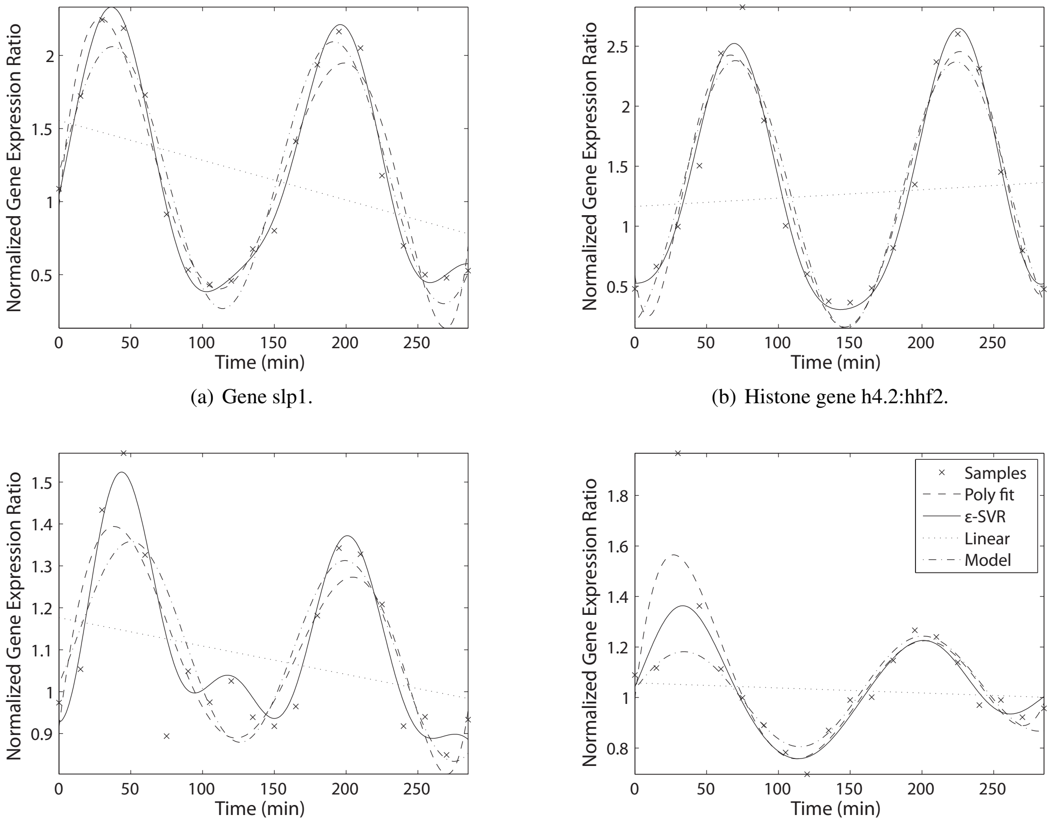
We first obtained some preliminary results based on applying a univariate ϵ-SVR (considering only the time vector as input) and the periodic model to each gene independently in the *elutriation2* data set. Fig. 2 shows a comparison of the observed data for four specific genes, identified as typical examples in Gilks *et al.* (2005), to four different signal estimation techniques: linear regression, least-squares polynomial regression with six degrees of freedom, ϵ-SVR and the proposed additive, periodic model. We find that although the ϵ-SVR curves precisely approximate all four genes with the least mean squared error, the proposed periodic model also aligns well these four genes and appears to be a good fit in all four cases. Table 1 shows the mean squared error for each of the four genes obtained from each estimation technique.

We then extend the technique to multivariate ϵ-SVR and estimate the cell-cycle length for the *elutriation2* data set. Only the 407 genes found to be periodic using a clustering algorithm in Rustici *et al.* (2004) were included. We find that the most likely cell-cycle length for this data set is 153.9 minutes, indicating that the 285 minutes in this experiment cover ∼1.85 cell-cycles. This appears to match well with the result found in Gilks *et al.* (2005) obtained through SVD. Fig. 4(a) shows a histogram of the most likely cell-cycle length obtained from the model for each gene in this data set: the estimated cell-cycle length is taken as the median value. Fig. 4(c–d) show histograms of the NRMSE of the models. A museum of interesting genes, showing examples of what is possible with this multivariate approach, is presented in Fig. 5.

Finally, we run the curve-fitting procedure again, holding the cell-cycle length $\Lambda = 153.9$. The results for the same four genes are compared in Fig. 3. Table 2 shows the periodic model parameters obtained for each of these four genes. Models for the full set of 407 genes identified as periodic by Rustici *et al.* (2004) is available online at the author's website, for both fixed and unfixed Λ.

In contrast to the conclusion in Rustici *et al.* (2004), in which a clustering algorithm determined that there were 407 periodic genes and 136 were strongly expressed, our analysis shows that 1252 of the 5038 genes show some statistically significant level of periodic activity higher than the estimated noise level within the *elutriation2* experiment, and 332 of these show periodic activity with a magnitude twice as high or greater than the estimated noise level. The noise level, which the multivariate ϵ-SVR model determines through the ϵ-tube width free parameter, was used as a significance measure in comparison to the periodic amplitude ρ_i determined by the periodic model's curve-fitting algorithm, in order to remove any potentially biased or arbitrary assumption. The complete list of 332 strongly expressed periodic genes may be found on the author's website.

As a visualization of the result, we then compare the magnitude and phase model parameters of the periodic component of the model obtained for each gene to determine the relative periodicity in relation to cell-cycle length for each active gene. This results in a plot similar to the "peppered fried egg plot" in Gilks *et al.* (2005)(Fig. 4) which was obtained from the first two eigenvectors of an SVD for each gene. Rather than reproduce the actual "egg yolk" (loess curve of radius of gene expressions through the cell cycle) and "egg white"



(a) Gene slp1.

(b) Histone gene h4.2:hhf2.

(c) Gene bgs4.

(d) Gene cdc20.

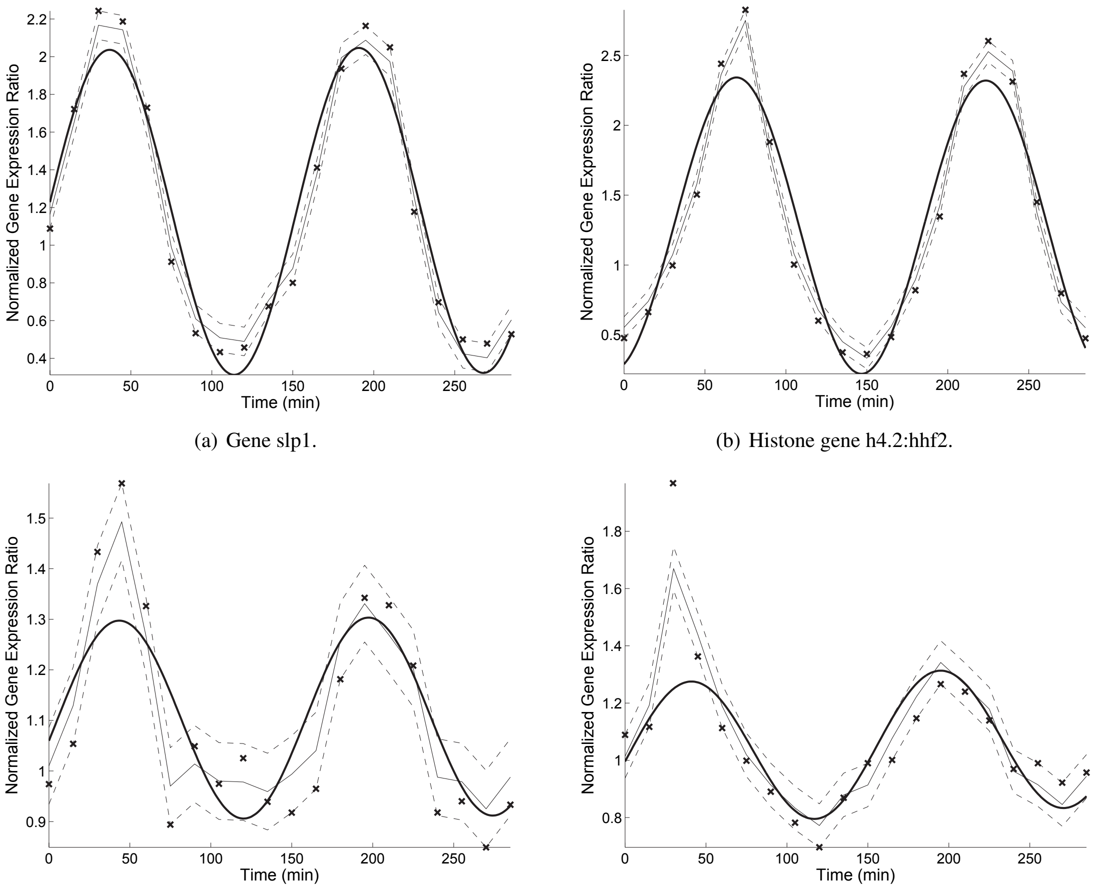
Fig. 2. Modeling periodic gene expression levels for several genes (see Gilks *et al.*, 2005, Fig. 5), comparing sixth-degree polynomial regression, ϵ-SVR and linear regression to the periodic model proposed in this paper, from the *elutriation2* data set. The legend for all four figures is as (d) (not shown in other figures for clarity). These results agree with the findings in Gilks *et al.* (2005), with slp1 and hhf2 exhibiting highly periodic behaviour but finding much less periodicity with bgs4 and cdc20. However, although the ϵ-SVR model finds some evidence of a second peak in bgs4, found to be slightly biperiodic in Gilks *et al.* (2005), the other models ignore this secondary peak as experimental noise. Also, cdc20 was thought to be non-periodic in Gilks *et al.* (2005), however these results do find some evidence of weak cell-cycle periodicity.

(loess curve of the density of genes through the cell cycle) individually, we combine these into a single curve as an indication of the average total gene expression through the cell cycle, and fit the curve using univariate ϵ-SVR rather than a loess curve: since the free parameters for ϵ-SVR are obtained through the same simulated annealing heuristic in Boardman and Trappenberg (2006), no smoothing assumption is necessary (a “span” is needed for the loess curve method, determining a breadth of values for smoothing). This appears more intuitive and gives a better picture of the activity of the mitotic genes. This visualization is presented in Fig. 6 for the 407 genes identified as periodic by Rustici *et al.* (2004). In this plot, all genes identified as periodic by the model are shown, regardless of the NRMSE statistic.

However, we may not wish to trust that these 407 genes truly are periodic: if we apply logical thresholds on the full set of all 5038 genes to determine which are strongly periodic, can we obtain the same result?

In Fig. 7, we show the rotational plot of the cell-cycle for all genes identified as significantly periodic by the additive model (i.e. those with a periodic amplitude $\rho_i > 2\,\epsilon$) and for which the NRMSE statistics are below $\mathcal{G}F_{SVM} < 0.4$ and $\mathcal{G}F_{Model} < 0.6$. Those genes with a large decay constant $\alpha_i \times 1000 > 3$ are also removed from the plot, as a large decay constant indicates that the curve-fitting procedure considers the linear component of the additive model to be too highly significant and was not able to properly converge. It seems likely in these cases that the slowly changing gene expression levels described by the linear component are due to some unobserved, external factor (such as changing ambient light levels, for example) rather than simply sensor drift in the microarray reader. This resulted in the selection of a total of 274 genes, shown in Fig. 7.

These thresholds were selected arbitrarily, based on the empirical distributions of each parameter, however it would be better to set these thresholds based on sound biological reasoning: for the moment, we leave this as future work. Note that the average total gene expression curve strongly resembles that of Fig. 6. The full set of genes identified as strongly periodic in this way can be found on the author’s website, with rotational plots for both the 407 periodic genes identified by Rustici *et al.* (2004) and the 274 identified here, including both the average total gene expression curve and the



(a) Gene slp1.

(b) Histone gene h4.2:hhf2.

(c) Gene bgs4.

(d) Gene cdc20.

Fig. 3. Modeling periodic gene expression levels for several genes (see Gilks *et al.*, 2005, Fig. 5), comparing multivariate ϵ-SVR (thin line) within ϵ error bounds (thin dashed lines) and the periodic model with cell-cycle length fixed at $\Lambda = 153.9$ minutes (thick line) with the original observed data ($\times$). The ϵ-SVR appears more jagged than in the univariate comparison (Fig. 2) since with the multivariate model only those points defined by the remaining genes may be used, whereas in the univariate model (which considers only the time vector as input) the curve may be evaluated at much finer resolution. It is important to note that in the multivariate case, the time vector itself is *not* used as an input: only those 3515 genes which have 100% of the data points (and which are not the target gene!) are used.

separated radius and density curves for comparison.

4 DISCUSSION

Comparing the visualization of the 407 periodic genes in Fig. 6 to those of Gilks *et al.* (2005)(Fig. 4), it is clear that the combined multivariate ϵ-SVR and non-linear curve-fitting approach has been successfully applied to this data set. There are many similarities between these plots, for example we see nine of the histone genes (those starting with *h*) occurring at the same point in the cycle in both plots, and the *slp1*, *plo1* and *spd1* genes appear to match as well.

There are also some differences, most likely since we only consider the *elutriation2* experimental data in this analysis, rather than the fusion of all nine experiments. For example, the *rds1* gene appears to be active earlier in the cycle. Several genes have stronger expression in this plot, for example the *meu19*, *exg1* and *etd1* genes appear in Fig. 6 but are not labelled as outliers in Gilks *et al.* (2005)(Fig. 4). There also appears to be a greater deviation between those genes with small expression and those with large expression, this may be partly due to the loss of information in plotting only two dimensions of the SVD, which in Gilks *et al.* (2005) were estimated to contain 83% of the SVD information.

In the visualization plot in Fig. 7, of all genes found to be strongly periodic by our model, we see more examples of genes with strong enough expression to be labelled as outliers. For example, the *C191.09c*, *C15D4.08c* and *C359.04c* genes appear to be strongly periodic. Fusion of this data with the remaining synchronization experiments could determine if these are truly additional periodic that were not originally discovered in the analyses of Gilks *et al.* (2005); Rustici *et al.* (2004). It is clear that overall, however, the genes identified in these two plots match fairly closely: the curve describing the average total gene expression throughout the cell-cycle is nearly identical in both plots.

This data fusion would be the next logical step in this analysis, and could be implemented as follows. We would first perform the same cleaning and data imputation procedure for each of the

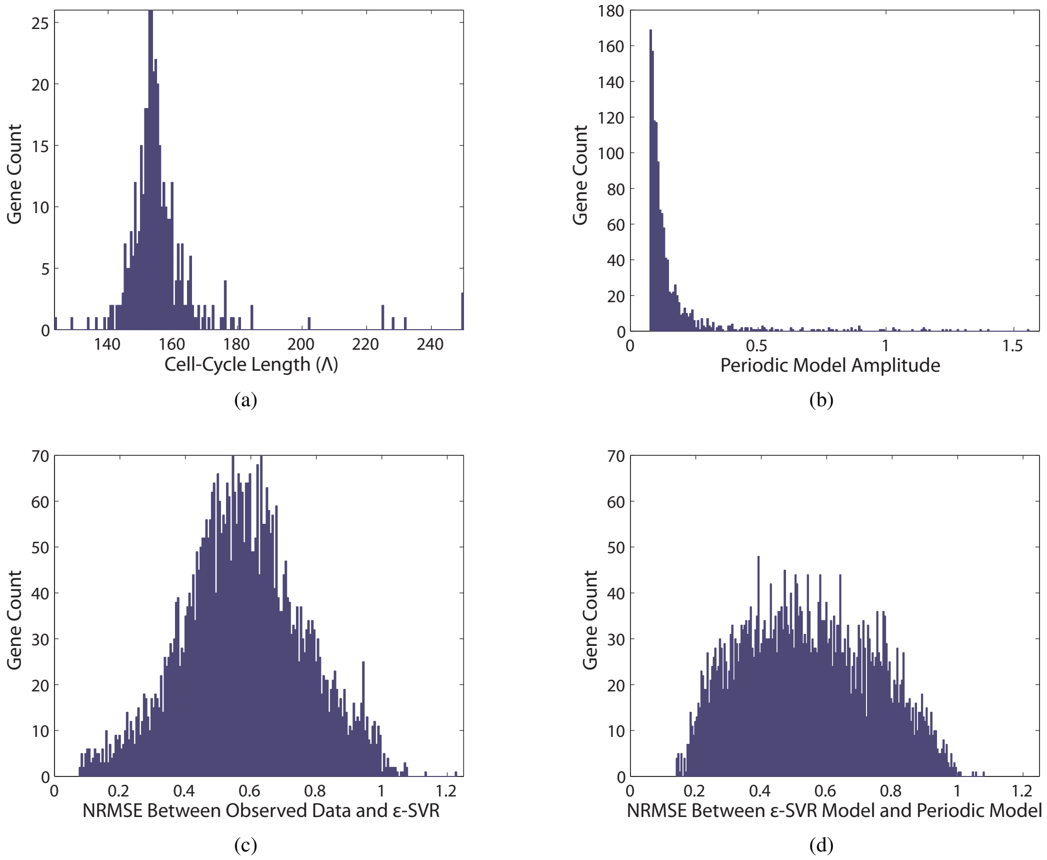
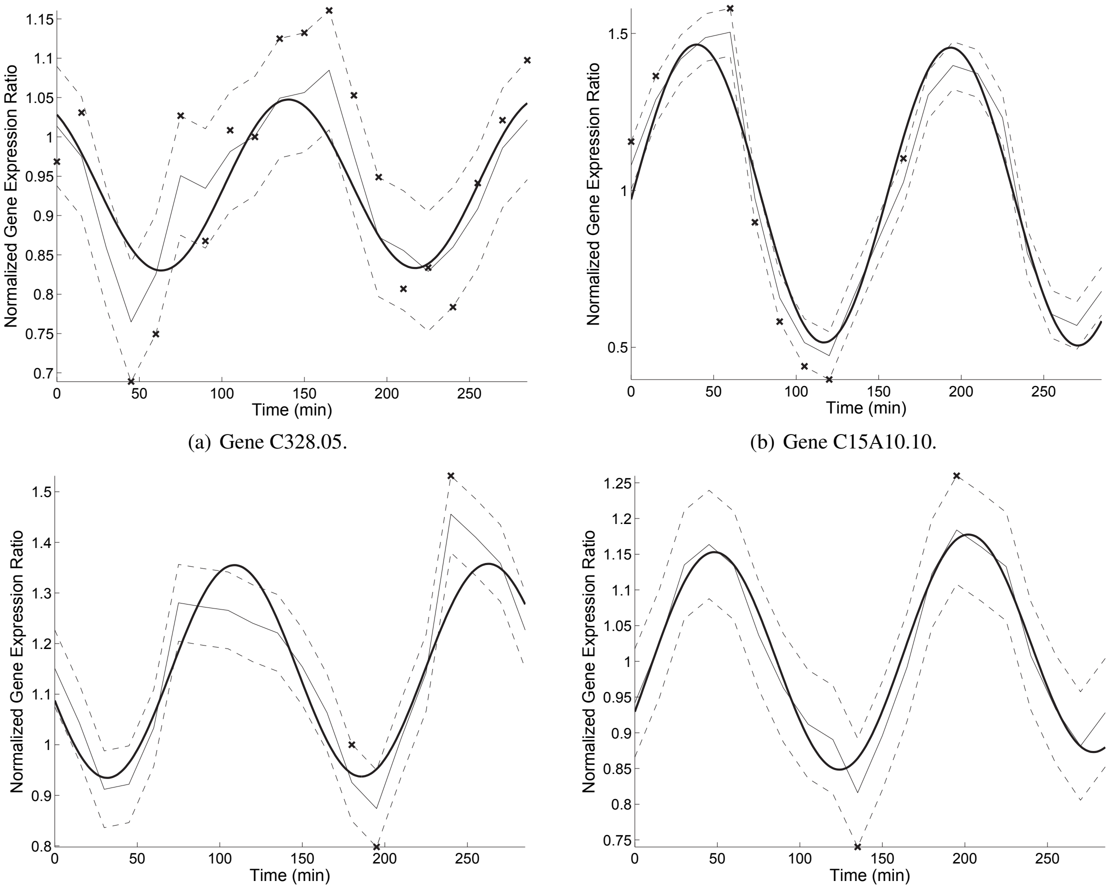


Fig. 4. (a) Distribution of cell-cycle length from the *elutriation2* data set for each gene, found by the periodic model using the cleaned multivariate ϵ-SVR data set as input. The 407 genes identified as periodic by Rustici *et al.* (2004) were used to find the median cell-cycle length of 153.9 minutes. Note that the sharp distribution allows high confidence in the resulting cell-cycle estimate. (b) Distribution of the amplitude of the periodic component of the additive model, for those genes with a statistically significant periodic amplitude, ie. greater than the estimated observed noise ϵ determined by the multivariate analysis. The majority of the genes exhibit low periodicity, but a significant number are strongly expressed with amplitude $\rho > 2\,\epsilon = 0.152$. (c) Distribution of the $\mathcal{G}F_{SVM}$ NRMSE statistic for all genes, between the original data and the data cleaned by multivariate ϵ-SVR. The median NRMSE is 0.5725. (d) Distribution of the $\mathcal{G}F_{Model}$ NRMSE statistic for all genes, between the data cleaned by multivariate ϵ-SVR and the predictions of the periodic model. The median NRMSE is 0.5322. Note the difference in the overall shape of the distributions in (c) and (d), although both have similar median values.

nine cell-cycle synchronization experiments performed by Rustici *et al.* (2004). From the analysis in Gilks *et al.* (2005), we know that the cell-cycle length and phase both will significantly vary between experiments due to the differences in synchronization procedures: the use of a periodic model allows us to estimate the cell-cycle length for each experiment individually. The phase for each experiment could then be estimated using a representative set of genes, such as the nine histone genes, using a best-fit optimization. The resulting cell-cycle length and phase for each experiment could then be used to create a time vector for each observation, for each gene, for each experiment, using an idealized cell-cycle stretching from 0 to $2\,\pi$. Univariate ϵ-SVR could then be applied to each gene individually to determine the most likely curve within this idealized cycle, and our curve fitting procedure could then determine the magnitude and phase of each gene as performed above. The results of this non-linear fusion might be interesting to compare with the above data, for example to determine how accurately the single *elutriation2* data set shows the periodic expression of each gene compared to the fusion of all sets, however, unfortunately, this is beyond the scope of the analysis performed in this paper. We therefore leave this for future work.

Another avenue for future work may be to determine the sensitivity of each gene to the values of the remaining genes using a modified Monte Carlo approach. For example, the formation of the SVM allows us to set all input genes to zero, the average expression level of the normalized input genes, then vary the values of each independently to determine which genes affect the outcome of the regression, and to quantify the extent of this effect and rank the input genes accordingly. This in turn would show us which groups of genes are closely related, and could perhaps lead to an advanced clustering technique, progressively forming a dendrogram based solely on these sensitivity relationships.

A simple web interface is available online at the author's website, allowing users to browse through the original observations, cleaned data set and generated models for the 407 periodic genes used in these experiments, for both fixed and unfixed cell-cycle length



(a) Gene C328.05.

(b) Gene C15A10.10.

(c) Gene C1105.13c.

(d) Gene C8E4.06.

Fig. 5. A museum of interesting genes found through the course of this analysis, showing the potential power of the multivariate ϵ-SVR algorithm. Since one of the strengths of SVM is to generalize the solution to a problem from a low number of observed samples, we would expect superior imputation performance for a data set such as this. It is important to note that the time vector is not used as an input for the ϵ-SVR: when imputing observations for a target gene, only that gene's relations to the remaining genes are considered. (a) Imputation of a single observation (at $t = 30$ min) with a higher than normal level of noise and small expression levels. (b) Data for the first cycle is present, but the second cycle is entirely imputed. KNN or other linear methods would simply assume a flat curve for the second cycle, but the relationships between genes in this multivariate approach allows the true expression of this gene to be seen. (c) Imputing periodicity from only three observations, and (d) from only two observations: these are possible since the observations are not closely spaced, so that the differences in other genes allow for prediction even in these seemingly impossible scenarios. Of course in practice, it would be unwise to rely on data modelled on only two or three observations; fusing the results from this experiment with the other experiments with different cell-cycle synchronization techniques, or simply repeating the particular experiment, would allow us to determine if these predictions are true.

and to see the genes identified here as strongly periodic. This comparison for each gene is available in both PNG and PDF formats, suitable for on-screen viewing and printing respectively.

From our analysis of the full *elutriation2* data set visualized in Fig. 7, we conclude that our combined technique of data imputation and noise reduction by multivariate ϵ-SVR, and non-linear curve-fitting by a periodic, additive model, is well-suited for the identification of periodic genes from DNA microarray data.

Conflict of Interest: none declared.

ACKNOWLEDGEMENT

The author is grateful for a discussion with Thomas Trappenberg, Dalhousie University, to discuss the methods used in this paper; and to the genomic researchers at the Wellcome Trust Sanger Institute, Hinxton, Cambridge, for making their experimental data publicly available.

REFERENCES

Agilent Technologies (2005) Agilent SureScan technology, *Publication No. 5988-7365EN*, http://www.chem.agilent.com.

Bennett,J.C. *et al.* (2000) Support vector machines: hype or hallelujah?, *SIGKDD Explorations*, **2**, 1–13.

Boardman,M.D. and Trappenberg,T. (2006) An heuristic for free parameter optimisation with support vector machines, *Proc. 2006 WCCI* (to be printed), Vancouver, BC, July 2006.

Brown,M.P.S. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines, *PNAS*, **97**, 262–267.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, **2**, 121–167.

Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

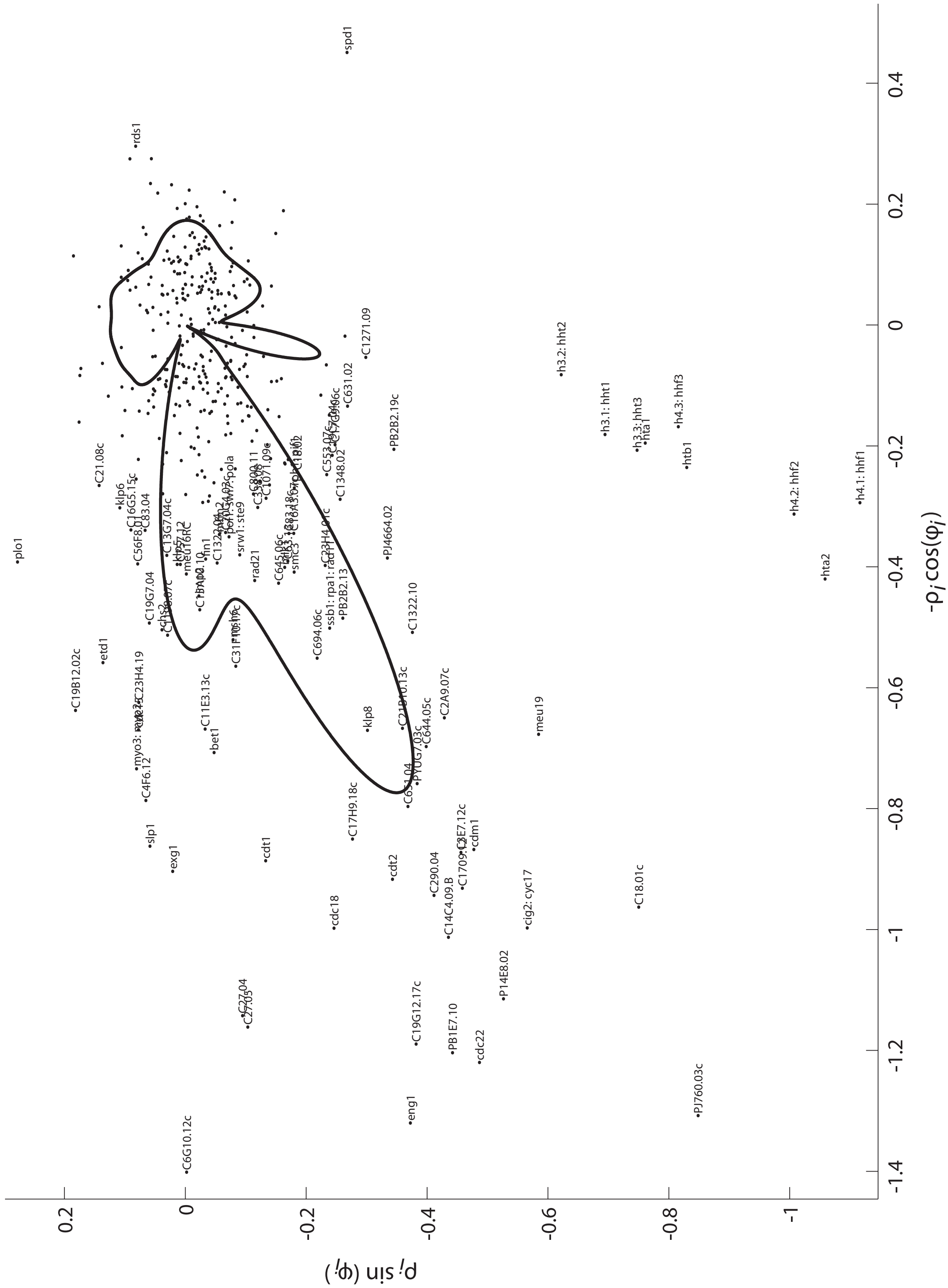


Fig. 6. A rotational plot of the periodic magnitude and phase, as captured by the periodic additive model created by the data set cleaned by multivariate ϵ-SVR, of each of the 407 genes identified as periodic by Rustici *et al.* (2004). The black dots indicate the periodic expression of individual genes, with some outliers labeled. The curve represents the average total gene expression throughout the cycle, smoothed by a univariate ϵ-SVR based on a histogram with 100 bins. Note the activation of the histone genes (those that start with *h*), clustered together near the bottom of the cycle, with a corresponding bump in the gene expression curve. This figure may be compared with Gilks *et al.* (2005) Fig. 4, which was obtained through singular value decomposition (SVD) rather than a periodic, additive model.

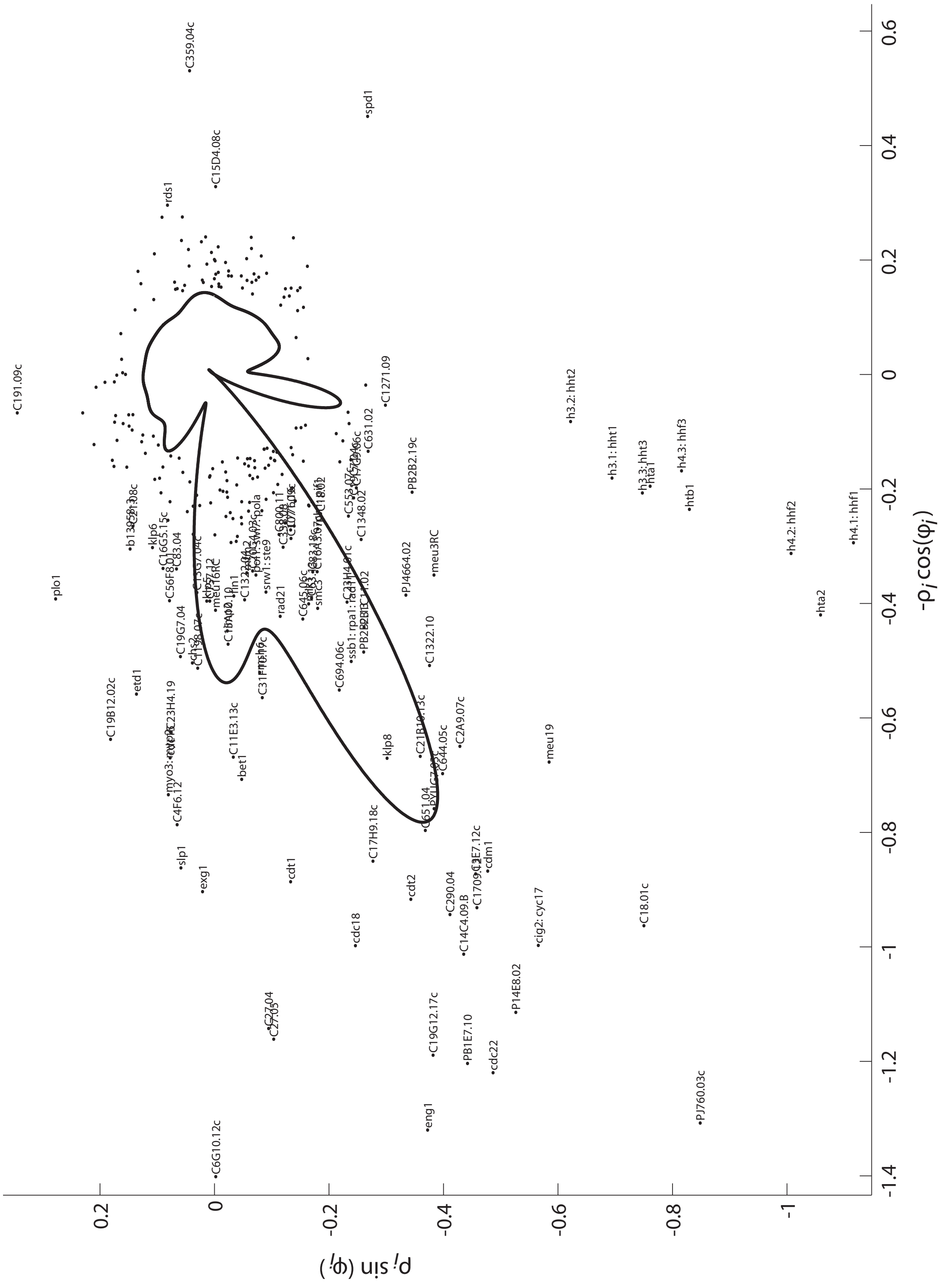


Fig. 7. A rotational plot as Fig. 6, but including only the 274 genes determine to be strongly periodic by our analysis. Note that the average total gene expression curve strongly resembles that of Fig. 6, indicating that differences between the plots will most likely be found in genes with small expression. Statistical significance was used to limit the genes from the 5038 original genes to the 332 found to have a periodic component of the additive model with an amplitude $\rho_i > 2\,\epsilon$, where ϵ is the estimated experimental noise provided by the free parameter estimation technique. Genes were further reduced by including only those with a goodness-of-fit values $\mathcal{G}F_{SVM} < 0.4$ and $\mathcal{G}F_{Model} < 0.6$. Those genes with a large decay constant $\alpha_i \times 1000 > 3$ were also removed as the high linear component appeared to prevent convergence of the non-linear curve-fitting.

Choi,J.K. *et al*. (2003) Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, **19**, i84–i90.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes, *Mol. Bio. Cell*, **11**, 4241–4257.

Gilks,W.R. *et al*. (2005) Fusing microarray experiments with multivariate regression, *Bioinformatics*, **21**, ii137–ii143.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, **18**, S96–S104.

Kerr,M.K. *et al*. (2000) Analysis of variance for gene expression microarray data, *J. Comput. Biol.*, **7**, 819–837.

Johansson,D. *et al*. (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription, *Bioinformatics*, **19**, 467–473.

Jörnsten,R. *et al*. (2005) DNA microarray data imputation and significance analysis of differential expression, *Bioinformatics*, **21**, 4155–4161.

Martin-Magniette,M.-L. *et al*. (2005) Evaluation of the gene-specific dye bias in cDNA microarray experiments, *Bioinformatics*, **21**, 1995–2000.

Martoglio,A.M. *et al*. (2002) A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer, *Bioinformatics*, **18**, 1617–1624.

Orengo,C.A. *et al*. (2003) *Bioinformatics: Genes, Proteins & Computers*, Springer-Verlag, NY, pp. 218–228.

Press,W.H. *et al*. (1992) *Numerical Recipes in C: The Art of Scientific Computing (2nd ed.)*, Cambridge University Press, UK, pp. 681–688.

Rustici,G. *et al*. (2004) Periodic gene expression program of the fission yeast cell cycle, *Nat. Genet.*, **36**, 809–817.

Segal,M. *et al*. (2003) Regression approaches for microarray data analysis, *J. Comput. Biol.*, **10**, 961–980.

Smola,A.J. and Schölkopf,B. (2004) A tutorial on support vector regression, *Statistics and Computing*, **14**, 199–222.

Smyth,G.K. *et al*. (2003) Statistical issues in cDNA microarray data analysis, In Brownstein,M.J. and Khodursky,A.B. (eds.), *Functional Genomics: Methods and Protocols*, Methods in Molecular Biology **224**, Humana Press, Totowa, NJ, pp. 111–136.

Thomas,J.G. *et al*. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles, *Genome Res.*, **11**, 1227–1236.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520–525.

Tai,Y.C. and Speed,T.P. (2004) A multivariate empirical Bayes statistic for replicated microarray time course data, *Technical Report No. 667*, University of California, Berkeley.

Tsai,C.-A. *et al*. (2004) A generalized additive model for microarray gene expression data analysis, *J. Biopharm. Stat.*, **14**, 553–573.

Vapnik, V.N. (1999) *The Nature of Statistical Learning Theory (2nd ed.)*, Springer-Verlag, NY.

Wang,X. *et al*. (2006) Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme, *BMC Bioinformatics*, **7**, (to be printed).

Wu,B. (2005) Differential gene expression detection using penalized linear regression models: the improved SAM statistics, *Bioinformatics*, **21**, 1565–1571.