DETECTION OF CELL-CYCLIC ELEMENTS IN MIS-SAMPLED GENE EXPRESSION DATA USING A ROBUST CAPON ESTIMATOR

Thomas Bowles, Andreas Jakobsson and Jonathon Chambers

Centre for Digital Signal Processing Research
King's College London
Strand, London, WC2R 2LS, U.K.

ABSTRACT

We present a method for the estimation of possible cell cyclic elements in mis-sampled microarray data. Accurate assessment of the frequency content of microarray data gives insight into genes which could be cell-cycle regulated. Cell cycle regulation is one component of the complex network of genetic regulatory processes and is especially relevant to the study of cancer. As cDNA microarray experiments involve human sampling of cell populations, slight variations in the sampling times invariably occur. Here, we propose estimating the frequency content of microarray data using the recent robust Capon estimator, and formulate a suitable uncertainty region to minimize over. The estimator is shown to yield robust estimates with real microarray data and to identify cell-cyclic genes that elude both the traditional Periodogram and the Capon spectral estimator.

1. INTRODUCTION

Genes provide a blueprint for the manufacture of protein. Typically, one gene codes for a single protein (or a family of similar proteins). The actual process whereby a protein is created is known as gene expression. The large scale measurement of protein levels from specific genes is not yet possible. However, microarrays facilitate the measurement, on a genome-wide scale, of relative levels of messenger RNA (mRNA) from specific genes [1]. mRNA is a necessary intermediary of gene expression and is currently the best indicator of gene expression levels on a large scale. Microarrays are flexible devices; one common experimental procedure is to subject cells to a physical or chemical stimuli and then sample the cell population over time. At each time point, the relative mRNA levels are thereby measured with a microarray. Such an experiment gives a time series (time course) of mRNA ratios for each gene. Hence, if each microarray contains P gene sample spots and N microarrays are used, one for each time point, the results can be expressed in a $P \times N$ matrix of gene expressions over time. Several recent studies used data from such experiments to estimate the cell cycle frequency and then determine genes that were likely to be cell-cycle regulated [2, 3]. Regulation is the term given to the complex network of processes governing gene expression. The cell cycle is the natural process of cell growth and division; some genes are known to be regulated in a cyclic fashion, coincident with the cell cycle. One method of determining genes which could be cell cycle regulated is through spectral estimation. Dominant spectral peaks at the frequency of the cell cycle provide evidence that the regulation of a particular gene is linked to the cell cycle. In [4], the amplitude spectrum Capon spectral estimator was used to rank genes by their frequency content at the cell cycle. This approach is hampered by possible errors in the temporal sampling of the cell population; the sampling of the cell population is typically performed by human operatives in a lab. As a result, slight variations in the sampling times invariably occur. For example, in one major study [2] the sampling was performed every seven minutes for one experiment but possible errors were estimated to be up to twenty seconds [5]. The robust Capon beamformer (RCB) presented in [6, 7] is able to determine the power in a signal of interest given imprecise knowledge of the array steering vector. As the beamforming problem is directly analogous to spectral estimation, the steering vector uncertainty is equivalent to uncertainty in the Fourier vector in the case of spectral estimation. Here, we show that errors in temporal sampling can be represented as an uncertainty disc around the Fourier vector. Let the ideally sampled data be represented as

$$y(t) = \alpha_\omega e^{i\omega t} + n(t), \tag{1}$$

for $t = 1, \ldots, N$, where α_ω is the (complex) amplitude of a generic sinusoidal component at frequency ω, where $\omega \in [0, 2\pi]$, and $n(t)$ is an additive zero mean coloured noise process containing component power at frequencies other than ω (see, e.g., [8]). Introducing sampling errors, we rewrite (1) as

$$y(t) = \alpha_\omega e^{i\omega(t+\Delta_t)} + n(t), \tag{2}$$

Corresponding author: thomas.bowles@kcl.ac.uk.

0-7803-8484-9/04/$20.00 ©2004 IEEE

where Δ_t is a random variable representing the sampling error at time t. Here, we make the natural assumption that $\{\Delta_t\}_{t=1}^{N}$ are independent identically distributed (IID) variables, with $\Delta_t \sim N\left(0,\sigma_\Delta^2\right)$, where σ_Δ^2 models the level of uncertainty in the sampling process. Let

$$\begin{aligned} \mathbf{y}_L(t) &= \left[\begin{array}{cccc} y(t) & y(t+1) & \ldots & y(t+L-1) \end{array}\right]^T \\ &= \alpha_\omega \tilde{\mathbf{a}}_L e^{i\omega t} + \mathbf{e}_L(t), \end{aligned} \tag{3}$$

for $t = 1,\ldots,N-L+1$, where $(\cdot)^T$ denotes the transpose operator,

$$\begin{aligned} \tilde{\mathbf{a}}_L &= \mathbf{a}_L \odot \mathbf{a}_\Delta \\ \mathbf{a}_L &= \left[\begin{array}{cccc} 1 & e^{i\omega} & \ldots & e^{i\omega(L-1)} \end{array}\right]^T \\ \mathbf{a}_\Delta &= \left[\begin{array}{cccc} e^{i\omega\Delta_t} & e^{i\omega\Delta_{t+1}} & \ldots & e^{i\omega\Delta_{t+L-1}} \end{array}\right]^T \end{aligned}$$

with $\odot$ denoting the Schur-Hadamard (elementwise) product. To form the uncertainty region created by the sampling uncertainty, we proceed to evaluate the expected value and the covariance matrix of $\tilde{\mathbf{a}}_L$. The expectation of $\tilde{\mathbf{a}}_L$ is

$$\bar{\tilde{\mathbf{a}}}_L = E\left(\tilde{\mathbf{a}}_L\right) = \mathbf{a}_L \odot E\left(\mathbf{a}_\Delta\right) = \mathbf{a}_L E\left(e^{i\omega\Delta_t}\right) \tag{4}$$

where we exploited the assumption that $\{\Delta_t\}_{t=1}^{N}$ are IID. Noting that $E\left(e^{i\omega\Delta_t}\right)$ is the characteristic function of a zero-mean Gaussian random variable yields

$$\bar{\tilde{\mathbf{a}}}_L = e^{\frac{-\omega^2\sigma_\Delta^2}{2}} \mathbf{a}_L \tag{5}$$

Similarly,

$$\begin{aligned} \mathbf{C}_{\tilde{\mathbf{a}}} &= E\left((\tilde{\mathbf{a}}_L - \bar{\tilde{\mathbf{a}}}_L)(\tilde{\mathbf{a}}_L - \bar{\tilde{\mathbf{a}}}_L)^*\right) \\ &= \left(1 - e^{-\omega^2\sigma_\Delta^2}\right)\mathbf{I}_L \end{aligned} \tag{6}$$

where $(\cdot)^*$ denotes the conjugate transpose. This covariance model for the sampling uncertainties could be easily enhanced with additional prior knowledge from the laboratory experiments.

2. ROBUST CAPON SPECTRAL ESTIMATION

Based on the above derivation, we assume that $\tilde{\mathbf{a}}_L$ belongs to the uncertainty ellipsoid

$$\left(\tilde{\mathbf{a}}_L - \bar{\tilde{\mathbf{a}}}_L\right)^* \mathbf{C}_{\tilde{\mathbf{a}}}^{-1} \left(\tilde{\mathbf{a}}_L - \bar{\tilde{\mathbf{a}}}_L\right) \leq 1 \tag{7}$$

where $\mathbf{C}_{\tilde{\mathbf{a}}}$ is given by (6) and $(\cdot)^*$ denotes the Hermitian transpose operator. Using (6), the hyperspherical uncertainty region is given by

$$\left\| \tilde{\mathbf{a}}_L - e^{\frac{-\omega^2\sigma_\Delta^2}{2}} \mathbf{a}_L \right\| \leq \varepsilon \tag{8}$$

where $\varepsilon = \beta\left(1 - e^{-\omega^2\sigma_\Delta^2}\right)$. Note that the radius of the hypersphere is a function of ω and σ_Δ. The reliance on σ_Δ is, of course, expected, but the presence of ω is also intuitive as the estimation of the spectral content at low frequency should be less affected by sampling errors than at higher frequencies. The extra scalar parameter β allows the uncertainty disc to be extended to give a more conservative estimate, which is useful for allowing extra unstructured uncertainty due to short data lengths and unknown noise characteristics. The robust Capon estimator [6, 7] is then obtained using the solution to the constrained minimization

$$\min_{\tilde{\mathbf{a}}_L} \tilde{\mathbf{a}}_L^* \hat{\mathbf{R}}_y^{-1} \tilde{\mathbf{a}}_L \quad \text{subject to} \quad \left\| \tilde{\mathbf{a}}_L - \bar{\tilde{\mathbf{a}}}_L \right\| \leq \varepsilon \tag{9}$$

where $\hat{\mathbf{R}}_y$ is the (estimated) covariance matrix of the measured data, and $\|\cdot\|$ denotes the Euclidean norm. To eliminate the trivial solution $\tilde{\mathbf{a}}_L = \mathbf{0}$, it is assumed that $\|\bar{\tilde{\mathbf{a}}}_L\|^2 > \varepsilon$. In this case, the solution will lie on the boundary of the constraint, simplifying the problem to a minimization with equality constraint

$$\min_{\tilde{\mathbf{a}}_L} \tilde{\mathbf{a}}_L^* \hat{\mathbf{R}}_y^{-1} \tilde{\mathbf{a}}_L \quad \text{subject to} \quad \left\| \tilde{\mathbf{a}}_L - \bar{\tilde{\mathbf{a}}}_L \right\| = \varepsilon \tag{10}$$

The solution to (10) is obtained using a Lagrange multiplier [6]

$$f = \tilde{\mathbf{a}}_L^* \hat{\mathbf{R}}_y^{-1} \tilde{\mathbf{a}}_L + \lambda\left(\left\| \tilde{\mathbf{a}}_L - \bar{\tilde{\mathbf{a}}}_L \right\|^2 - \varepsilon \right) \tag{11}$$

The optimal solution $\hat{\tilde{\mathbf{a}}}_L$ is given by differentiation of (11) with respect to $\tilde{\mathbf{a}}_L$, yielding the solution:

$$\hat{\tilde{\mathbf{a}}}_L = \bar{\tilde{\mathbf{a}}}_L - \left(\mathbf{I} + \lambda\hat{\mathbf{R}}_y\right)^{-1} \bar{\tilde{\mathbf{a}}}_L \tag{12}$$

The Lagrange multiplier λ is obtained by the solution of the constraint equation:

$$g(\lambda) \equiv \left\| \left(\mathbf{I} + \lambda\hat{\mathbf{R}}_y\right)^{-1} \bar{\tilde{\mathbf{a}}}_L \right\|^2 = \varepsilon \tag{13}$$

A unique solution to (13) is obtained through gradient descent (see [6] for details and the formulation of upper and lower bounds). With the Lagrange multiplier determined, $\hat{\tilde{\mathbf{a}}}_L$ is given by (12). The robust Capon spectral estimate is given by using $\hat{\tilde{\mathbf{a}}}_L$ in place of $\mathbf{a}_L$ in the classical power spectrum Capon, i.e., the estimated power spectral density is obtained as

$$\hat{\sigma}_\omega^2 = \frac{1}{\hat{\tilde{\mathbf{a}}}_L^* \hat{\mathbf{R}}_y^{-1} \hat{\tilde{\mathbf{a}}}_L} \tag{14}$$

In the following, $\hat{\mathbf{R}}_y$ is estimated by the forward-backward method to avoid sensitivity to phase errors (see [9] for a more detailed discussion on the benefits of this estimator as compared to the forward-only estimator).

3. NUMERICAL EXAMPLES

To demonstrate the advantages of the robust Capon spectral estimator in the application of microarray data we use data from [2]. The data consists of 6075 gene expression time series from the genes of *saccharomyces cerevisiae* (bakers' yeast) subjected to α-factor arrest. The data consists of only $N = 18$ time points. The value of the estimated cell cycle frequency was estimated from the peak in the ensemble average of the spectral estimates for all genes. Figure 1 shows the estimated spectrum of selected genes with the Periodogram, classical Capon and robust Capon[1], both the latter with filter length $L = 10$. Here, we use $\beta = 8$.

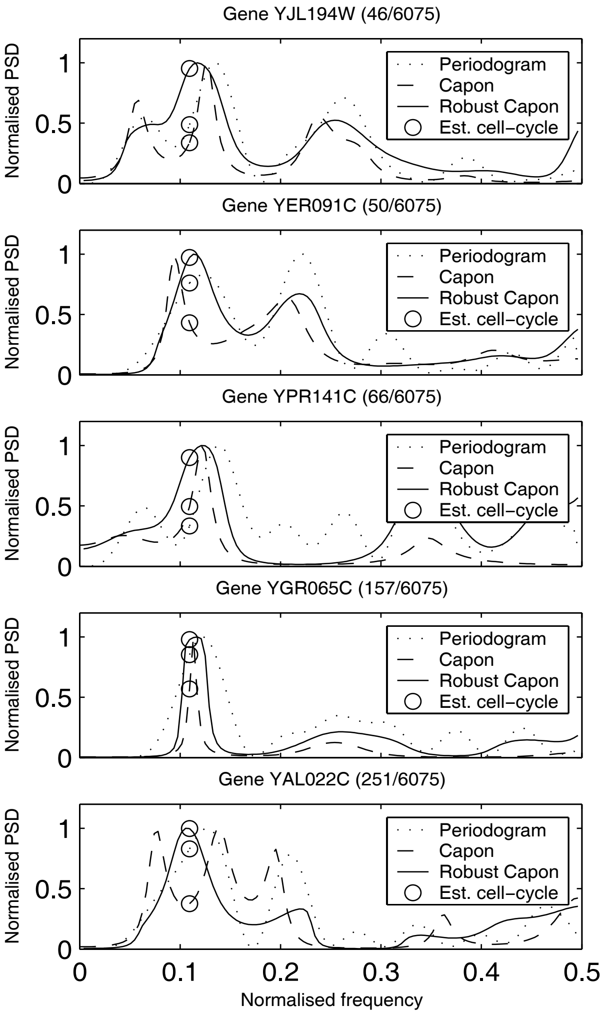

Fig. 1. Spectrum estimates of selected genes by robust Capon and classical Capon and periodogram methods . The estimated cell cycle frequency is circled. Both axes are normalised.

The estimates given in Figure 1 show typical examples of genes which in [2] were decided to be cell-cycle regulated. The marked cell-cycle frequency was obtained as the dominant peak of the ensemble average of the entire data set. In all cases, the robust Capon places a definite peak at the location of the estimated cell cycle frequency. The classical Capon tends to place a very sharp peak in the vicinity of the cell cycle frequency but the amplitude value at the cell cycle frequency can be relatively low. It is likely that the sharp peak is misplaced because of the significant uncertainty in the data. The Periodogram, has a broader peak but, this too is often misplaced and, as expected, suffers from spurious peaks due to the large sidelobes. The Periodogram and classical Capon both show more variation than the robust Capon in the spectrum outside the region containing the estimated cell cycle. Figure 2 shows a scatter plot of the values of the robust Capon against the classical Capon at the estimated cell-cycle frequency.

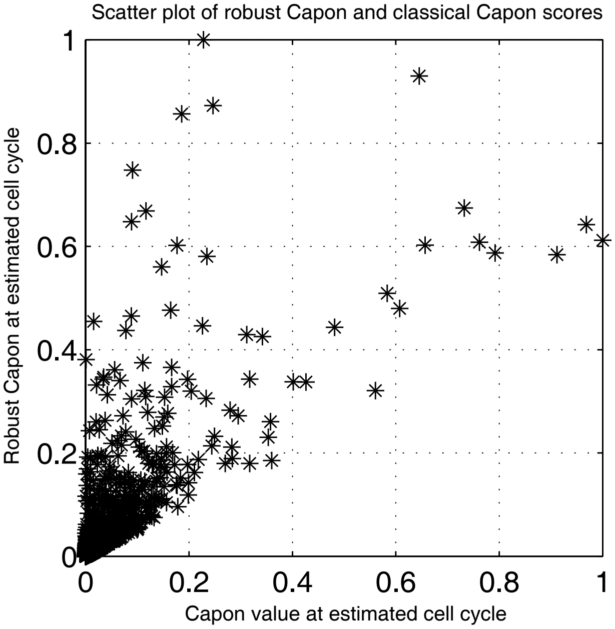

Fig. 2. Scatter plot showing the correlation between the classical and robust Capon values at the estimated cell cycle frequency. Note the group of genes in the top left of the plot showing expression profiles scoring highly for the robust Capon but not for the classical Capon.

In general, the results show the expected positive correlation. However, as seen in the top-left of the figure, there is a subset of genes where the robust Capon estimate is significantly higher than the classical Capon. These are candidates for genes with cell-cyclic components which have not been identified by the classical Capon. Our results support the findings of [2], despite this there remains no absolute

[1] The authors are grateful to Prof J. Li, Prof P. Stoica and Mr Z. Wang for providing us with their efficient implementation of the RCB.

knowledge of exactly which genes are cell cycle regulated and to what extent. Because of this our method was validated on a synthetic data set. A set of 200 synthetic gene expression profiles of length $N = 18$ were created with 10% containing a single sinusoid with a randomly distributed phase and a frequency approximately matching that of the estimated cell-cycle frequency in [2]. Further, the data was corrupted by an additive white Gaussian noise with a signal to noise ratio of 12 dB. Finally, the signal was mis-sampled by a Gaussian distributed random variable with variance σ_{Δ}^2 being 5% of the sampling time; this corresponding well to the reported mis-sampling in [2, 5]. As the frequency of the sinusoid is now known, the hyperspherical constraint was kept constant at $\varepsilon = 6$. Figure 3 shows a scatter plot of the correlation between the classical and the robust Capon estimates at the (known) frequency of the sinusoidal component for the synthetic data. The profiles containing the sinusoidal component, which represents a cell-cyclic component, are distinguished from the expression profiles which are purely noise. Both methods give low scores to the synthetic gene profiles with no sinusoidal component. However, whilst the robust Capon assigns generally high scores for the genes with a sinusoidal component, the classical Capon's scores are widely distributed over the entire range. This supports the conclusion from the real data that the robust Capon is able to pick up cell-cyclic elements which the classical Capon does not.

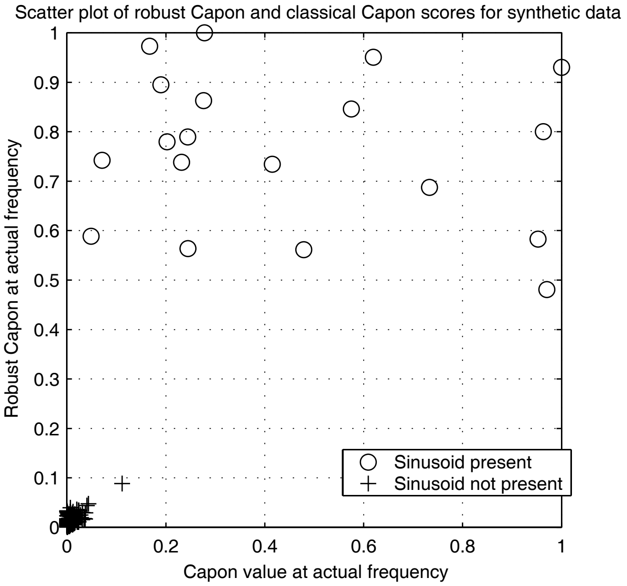

Fig. 3. Scatter plot showing the correlation between the classical and robust Capon values for synthetic data. Scores for the sinusoidal profiles are high for the robust Capon but spread throughout the range for the classical Capon.

4. CONCLUSIONS

We have presented a method for the idenfication of cell-cyclic components in gene expression profiles which is robust to the uncertainties inherent in microarray data. The method has been shown to perform well with both real and synthetic datasets. Given that no good biologically-inspired models are currently available, parametric estimators are inappropriate as of yet. However, the use of non-parametric spectral estimators allowing for the incorporation of limited domain knowledge through uncertainty shaping is a promising area of research.

5. REFERENCES

[1] P. Brown and D. Botstein, "Exploring the new world of the genome with dna microarrays," *Nature Genetics*, vol. 21, pp. 33–37, 1999.

[2] P. T. Spellman, G. Sherlock, M. Zhang, V. Lyer, 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 hydridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273–3297, 1998.

[3] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L.Wodicka, T. Wolfsberg, A. Gabrielan, D. Landsman, D. Lockhart, and R. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Biology of the Cell*, vol. 2, pp. 65–73, 1998.

[4] T. Bowles, J. Chambers, and A. Jakobsson, "Advanced spectral estimation for the identification of cell-cycle regulated genes," in *IEEE EMBS UK and RI postgraduate conference in biomedical engineering and medical physics*, pp. 19–20, 2003.

[5] G. Sherlock, "Personal communication." 2003.

[6] J. Li, P. Stoica, and Z. Wang, "On Robust Capon Beamforming and Diagonal Loading," *IEEE Trans. Signal Processing*, vol. 51, pp. 1702–1715, July 2003.

[7] J. Li, P. Stoica, and Z. Wang, "Robust capon beamforming," *IEEE Signal Processing Letters*, vol. 10, no. 6, pp. 172–175, 2003.

[8] E. G. Larsson, J. Li, and P. Stoica, *"High-Resolution Nonparametric Spectral Analysis: Theory and Applications"*. In *High-resolution and robust signal processing*, Y. Hua, A. B. Gershman and Q. Cheng, Eds., New York, N.Y., USA: Marcel-Dekker, 2003.

[9] M. Jansson and P. Stoica, "Forward-only and forward-backward sample covariances-a comparative study," *Signal Processing*, vol. 77, pp. 235–245, 1999.