Detecting Periodically Expression in Unevenly Spaced Microarray Time Series

Jun Xian¹, Jinping Wang², and Dao-Qing Dai^{1,3}

¹ Department of Mathematics, Sun Yat-Sen (Zhongshan) University, Guangzhou, 510275, China

² Department of Mathematics, Ningbo University, Ningbo, Zhejiang, 315211, China

³ Center for Computer Vision, Sun Yat-Sen (Zhongshan) University, Guangzhou, 510275, China

xianjun@mail.sysu.edu.cn, stsddq@mail.sysu.edu.cn

Abstract. Spectral estimation has important applications to microarray time series analysis. For unevenly sampled data, a common spectral estimation technique is to use the Lomb-Scargle algorithm. In this paper, we introduce a new reconstruction algorithm and singular spectrum analysis (SSA) method to deal with unevenly sampled microarray time series. The new reconstruction method is based on signal reconstruction technique in aliased shift-invariant signal spaces and a direct implemental algorithm is developed based on the B-spline basis. We experiments on simulated noisy signals and gene expression profiles show different effects for our designed three methods. The three methods are based on our presented reconstruction algorithm, SSA, classical FFT periodogram method and Lomb-Scargle periodogram method.

Keywords: Spectral estimation; periodically expressed genes; unevenly sampled data; the Lomb-Scargle method; signal reconstruction; B-splines.

1 Introduction

Spectral estimation has been a classical research topic in signal processing communities. Many approaches have been proposed in the past decades, including the modified periodogram, autoregressive (AR) model, the MUSIC algorithm and the multitaper method [1,2]. Although all these algorithms have their own advantages, they are all developed based on a basic assumption: the input signal is evenly sampled. However, in many real-world applications, the data can be unevenly sampled. For example, in DAN microarray gene expression experiments, a time series may be obtained with different time sampling intervals [3-5]. Furthermore, an evenly sampled time series may contain missing values due to corruption or absence of some expression measurements. A time series with missing values can be considered as one with unevenly data samples in general.

Lu et al. [6] proposed a periodic-normal mixture (PNM) model to fit transcription profiles of periodically expressed genes in cell cycle microarray experiments and obtained some nice results. Ruf [7] is one of the first to treat evenly sampled gene expression time series with missing values as unevenly sampled data for spectral

analysis using the Lomb-Scargle periodogram. Recently, Bohn et al. [8] have used the Lomb-Scargle periodogram in their attempt to detect rhythmic components in the circadian cycle of the Crassulacean acid metabolism plants. The Lomb-Scargle periodogram was originally developed for analysis of noisy unevenly sampled data from astronomical observations.

Although spectral analysis can be applied directly to the original data, noise and outliers would degrade the final results. In DNA microarray analysis, a major challenge is to effectively dissociate actual gene expression values from noise. As a robust model-free technique in time series analysis, SSA can be used to extract as much reliable information as possible from short and noisy time series without prior knowledge of the dynamics underlying. It has been proven to be a powerful tool for processing many types of time series in geophysics, economics, medicine and other sciences [9]. SSA method was originally developed for analysis of noisy unevenly sampled data from geophysical observations. We will introduce SSA method into DAN microarray gene expression analysis. To the best of our knowledge, we firstly introduce SSA into DAN microarray gene expression data analysis.

In this paper, we propose a new spectral estimation technique for unevenly sampled data. Our method models the signal in the aliased shift-invariant signal space that is a generalization of shift-invariant signal space, for which many theories and algorithms are available [10-22]. The proposed algorithm is also flexible in adjusting the order of the B-spline basis based on the number of sampling points. We experiments on simulated noisy signals and gene expression profiles show different effects for our designed three methods. The three methods are based on our presented reconstruction algorithm, SSA, classical FFT periodogram method and Lomb-Scargle periodogram method.

2 Mathematical Theory and Algorithm

2.1 Reconstruction Algorithm of Signal in Aliased Shift-Invariant Signal Spaces

In the following, we first review existing work on signal analysis in the shift-invariant signal space.

The space of a bandlimited signal is introduced, which is inspired by Shannon's signal sampling and reconstruction theorem:

$$V(\text{sinc}) = \{f(x) = \sum_{k \in \mathbb{Z}} c_k \text{sinc}(x - k) : (c_k) \in \ell^2\} \quad (1)$$

where $\text{sinc}(x) = \frac{\sin(x)}{x}$.

Since the sinc function has an infinite support and slow decay, it is not convenient in real applications. Xian and Lin [21] found a good decay function that can replace the sinc basis function, but the new function still has an infinite support. To replace the sinc function with a general function ϕ , we can introduce a signal space that is called the shift-invariant (also called time-invariant) signal space [10, 13-14, 17-18]. Its aliased version is defined by

$$V_L(\phi) = \{f : f(x) = \sum_{k \in \mathbb{Z}} c_k \phi(x - Lk) : (c_k) \in \ell^2\} \quad (2)$$

where $L > 0$ is a constant.

Signal reconstructing in the shift-invariant space is an active research area and there are many mathematical theories and computer algorithms on the topic [10-22]. When the signal $f \in V_L(\phi)$, we hope to reconstruct signal f from sampled value $\{f(x_i)\}$, where $\{x_i\}$ is the sampling point set.

In fact, the well-known autoregressive (AR) model can be regarded as a special case of signal reconstruction in the above signal space. Yeung et al. [23] have presented an application of the AR model to microarray gene expression time series analysis.

In order to avoid introducing spurious periodic components caused by basis function ϕ with infinite support, for example, the *sinc* function, here we only consider basis functions with compact support. The B-spline function is a good choice and is widely adopted in the wavelet reconstruction theory. We use $\text{supp } \phi$ to indicate the support of basis function ϕ . Assume that $\text{supp } \phi \subset [-\Omega, \Omega]$ and $f(x) \in V_L(\phi)$ that is defined in finite interval $[A_1, A_2]$, then f can be determined completely by the coefficients $\{c_k\}$ for $k \in (\frac{A_1 - \Omega}{L}, \frac{A_2 + \Omega}{L}) \cap \mathbb{Z}$ with

$$f(x) = \sum_{k=\frac{A_1 - \Omega}{L} + 1}^{\frac{A_2 + \Omega}{L} - 1} c_k \phi(x - Lk) \quad (3)$$

where $\text{supp } \phi = \overline{\{x : \phi(x) \neq 0\}}$.

The coefficients, $\{c_i\}$, can be calculated according to following steps:

(1) Given sampling points $x_1, \dots, x_J \in [A_1, A_2]$ and corresponding discrete

function value $y = (y_1, \dots, y_J)$. Assume that $J \geq \frac{A_2 - A_1 + 2\Omega}{L} - 1$ and

that the truncated matrix T defined below is invertible. Computing matrix: $U = (U_{jk})$, $T = (T_{kl})$, where

$$U_{jk} = \phi(x_j - Lk), T_{kl} = \sum_{j=1}^J \overline{\phi(x_j - Lk)} \phi(x_j - Ll),$$

$$j = 1, \dots, J, k, l = \frac{A_1 - \Omega}{L} + 1, \dots, \frac{A_2 + \Omega}{L} - 1.$$

(2) Compute $c = T^{-1}b$ according to $b = \overline{U}y$, where $\overline{U}$ denotes complex conjugate of U and T^{-1} is the inverse of T .

Remark

- (1) The construction of matrix T has the advantage that it is a positive operator on $\ell^2(Z)$ [24].
- (2) In the case of B-spline, that matrix T is invertible under the condition that

$$\sup_j (x_{j+1} - x_j) \leq \frac{L(A_2 - A_1)}{A_2 - A_1 + 2\Omega - L} \quad (4)$$

That is, the Schoenberg-Whitney Theorem implies that T is invertible [25].

- (3) Since the numerical solution of (c_k) could be sensitive to particular sets of intersample spacings, we can approximate the inverse of the singular or ill-conditioned T by its pseudoinverse using singular valued decomposition (SVD).

2.2 Lomb-Scargle(L-S) Periodogram Method

In the following, we briefly give an expression of Lomb-Scargle periodogram method. For a times series $X(t_i)$, where $i = 1, 2, \dots, N_0$, the periodogram as a function of the frequency ω is defined as

$$P_{xx}(\omega) = \frac{1}{2\sigma^2} \left\{ \frac{\left[\sum_{j=1}^{N_0} X(t_j) \cos \omega(t_j - \tau) \right]^2}{\sum_{j=1}^{N_0} \cos^2 \omega(t_j - \tau)} + \frac{\left[\sum_{j=1}^{N_0} X(t_j) \sin \omega(t_j - \tau) \right]^2}{\sum_{j=1}^{N_0} \sin^2 \omega(t_j - \tau)} \right\} \quad (5)$$

where τ is defined by the equation

$$\tan(2\omega\tau) = \left(\sum_{j=1}^{N_0} \sin 2\omega t_j \right) / \left(\sum_{j=1}^{N_0} \cos 2\omega t_j \right)$$

$$\text{and } \sigma^2 = \frac{1}{N_0 - 1} \sum_{i=1}^{N_0} (X(t_i))^2.$$

The above Lomb-Scargle periodogram method gives superior results on unevenly sampled data with classical FFT method: it weights the data on a “per point” basis instead of on a “per time interval” basis. Lomb proved that his periodogram is the same as the classical periodogram (see [26]) in the case of equally spaced data. But since it assumes there is a single stationary sinusoid wave that has infinite support in Lomb-Scargle method, it may introduce some spurious periodic components for finite data. Also due to the effect of noise in the given data, it may produce inaccurate estimation results.

2.3 Singular Spectrum Analysis (SSA) Method

Although spectral analysis can be applied directly to the original data, noise and outliers would degrade the final results. In DNA microarray analysis, a major challenge is to effectively dissociate actual gene expression values from noise.

As a robust model-free technique in time series analysis, the SSA can be used to extract as much reliable information as possible from short and noisy time series without prior knowledge of the dynamics underlying [9]. The main idea of the SSA is to perform singular value decomposition (SVD) of the so-called trajectory matrix obtained from the original time series with subsequent reconstruction of the series. It is aimed at an expansion of the original time series into a sum of a small number of independent and additive components that typically can be interpreted as trend components, oscillatory components and noise components. These properties make SSA a useful technique for gene expression data processing. The SSA decomposition into three parts is indeed related to the separation of the spectral range into three regions: low frequencies corresponding to the trend, the higher frequencies describing the oscillatory component, and the residual as noise. The Fourier analysis demonstrates that the power spectrum of the expression profiles are dominated by low frequency components. Therefore, SSA could be used to extract the trend curve that represents the dominative component in order to reduce the effect of noise, which in fact is a process of data fitting. The detail of SSA can be found in [9].

2.4 Fisher's Test

To search for periodic gene expression, a formal test should be carried out to determine a peak in periodogram whether or not is significant or not. Fisher [7] proposed a test statistic and derived the null distribution of the Fisher's G-statistic for finite samples, and this is what will be utilized next. In the context of microarray gene expression data, the observed significance value or p-value for the hypothesis testing of the periodicity of a fixed gene g , using G-statistic as the test statistic, can be obtained by explicit expression. The G-statistic value calculated from the Fisher's G-statistic. So Fisher's Test gives a p-value that allows to test whether a gene behaves like a purely random process or whether the maximum peak in the periodogram is significant. For details on the G test statistic, its null distribution and the percentage points of the G test statistics, please refer to [27-29].

3 Experimental Results

3.1 Design of Experiment

In our numerical test, we choose B-spline of order N as the basis function ϕ in our reconstruction algorithm. It is obvious that $\text{supp } \phi \subset [-\frac{N+1}{2}, \frac{N+1}{2}]$.

L-S method can directly based on unevenly sampling data, but it may be deduced by missing data and noise. Though SSA can be used to extract as much reliable information as possible from short and noisy time series without prior knowledge of the dynamics underlying, SSA are based on evenly sampling data. The classical periodogram method also has the disadvantage. So we need a reconstruction data from original raw unevenly signal for SSA and the classical periodogram method. We adopt our presented reconstruction algorithm in Section 2.1.

So experiments are designed with the following three cases:

- I. Raw data with missing points → L-S Method
- II. Raw data with missing points → Our Reconstruction Algorithm → Classical Periodogram Method
- III. Raw data with missing points → Our Reconstruction Algorithm → SSA → Classical Periodogram Method

3.2 Simulated Data

First, we use simulated data to test the proposed method with the three cases. For a gene g and expression level observed at time t_i , we denote the time series by $Y_g(t_i)$ for $i = 1, \dots, N$ and $g = 1, \dots, G$.

$$Y_g(t_i) = \beta \cos\left(\frac{t_i \pi}{12}\right) + \varepsilon_g(t_i) \quad (6)$$

In the case of the simulated data, the length of a whole ideal time series (genes) is evenly 48 points. That is, $t_i = i (i = 1, \dots, 48)$.

Fig. 1. Comparison of spectral estimation for simulated data: (a) simulated cosine signal with even sampling ($N=48$), (b) simulated noisy cosine signal with uneven sampling, (c) the periodogram of the simulated signal in (a) obtained using the Fourier transform, and the periodograms of the signal in (c) obtained using the three methods mentioned in Section 3.1 are shown in (d), (e) and (f), respectively

First, we test a gene that presented in (6). We give a classical periodogram of 24 hours period cosine with evenly 48 sampling points as a standard contrastive figure.

Let FFT periodogram based on evenly sampling points be a standard periodogram (See Figure 1(c)). Method III produces fewer and smaller false peaks than Method II in Figure 1(e), (f). Method II produces fewer and smaller false peaks than Method I in Figure 1(d), (e). The peaks are highest among the three methods and closest the standard (ture) peak value with Method III (see Figure 1(f)).

3.3 Genes Data

The gene expression for *S. Pombe* in [30] were gathered from three experiments. Hence the data sets are called *cdc25*, *elutA* and *elutB*. In our test, we only apply our methods to deal with the gene expression data from the *cdc25* experiment.

From the Table 1, we observe that Method III can find more periodic genes than Method I and Method II under the same FDR level threshold. The reasons are the reconstruction algorithm repairing the missing data and SSA reducing noise. Because the missing rate of *cdc25* data is lower, the effect of Method I is almost as same as Method II. We believe that the effect of Method II is better than Method I in a high missing rate data.

Table 1. Data sets analyzed with different methods

Cell type	Experiment	Number of periodic expression		
		Method I	Method II	Method III
<i>S. Pombe</i> (Olivien et al., [30])	<i>cdc25</i>	420	411	493

Notation: The number of periodic genes that are statistically significant for a FDR level of $q=0.03$.

4 Conclusions

In this paper, we have proposed a new spectral estimation algorithm based on a signal reconstruction technique in an unevenly sampled space. The new algorithm is flexible since the order of the B-spline basis function can be adjusted. Experiments on simulated signals and real gene expression data show that Method III (Reconstruction→SSA→Classical Periodogram Method) is more effective in identifying periodically expressed genes. Due to the limitation of eight pages, we omit some simulated tests and real data tests.

Acknowledgements

This project is supported in part by NSF of China (60175031, 10231040, 60575004), Grants from the Ministry of Education of China (NCET-04-0791), NSF of Guang Dong and Grants from Sun Yat-Sen University 10471123, the Mathematical Tianyuan Foundation of China NSF (10526036) and China Postdoctoral Science Foundation

(20060391063), Natural Science Foundation of Zhejiang Province (Y606093) and Scientific Research Fund of Zhejiang Provincial Education Department(20051760) .

References

1. Hayes, M. H., Statistical digital signal processing and modeling. John Wiley and Sons, Inc. (1996)
2. Kay, S.M., Marple Jr, S.L.: Spectrum analysis-A modern perspective. Proceeding of IEEE. 69(11) (1981) 1380-1418
3. Chu, S., DeRisi, J., et al. : The transcriptional program of sporulation in budding yeast. Science. 282(1998) 699-705
4. Eisen, M.B., Spellman, P.T., Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. PNAS. 95(1998) 14863-14868
5. Spellman, T.S., Sherlock, G., et al: Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisia* by microarray hybridization. Mol. Biol. Cell. 9(1998)3273-3297.
6. Lu, X., Zhang, W., Qin, Z.H., Kwast, K.E., Liu, J.S.: Statistical resynchronization and Bayesian detection of periodically expressed genes. Nucleic Acids Research. 32(2)(2004) 447-455
7. Ruf, T.: The Lomb-Scargle Periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series. Biological Rhythm Research. 30(1999) 178-201
8. Bohn, A., Hinderlich, S., Hutt, M.T., Kaiser, F., Luttge, U.: Identification of rhythmic subsystems in the circadian cycle of crassulacean acid metabolism under thermoperiodic perturbations. Biol. Chem. 384(2003)721-728
9. Golyandina N., Nekrutkin V., Zhigljavsky A., Analysis of time series structure: SSA and related techniques, Chapman & Hall/CRC, 2001.
10. Benedetto, J.J.: Irregular sampling and frames, in: C. K. Chui (Ed.), wavelets: A Tutorial in Theory and Applications (1992)445-507
11. Chen, W., Han, B., Jia, R.Q.: Estimate of aliasing error for non-smooth signals prefiltered by quasi-projections into shift-invariant spaces. IEEE Trans. Signal. Processing. 53(5) (2005) 1927-1933
12. Chen, W., Han, B., Jia, R.Q.: On simple oversampled A/D conversion in shift-invariant spaces. IEEE Trans. Inform. Theory. 51(2) (2005) 648-657
13. Chui, C.K.: An introduction to Wavelet, Academic Press, New York 1992
14. Daubechies, I.: Ten Lectures on Wavelets, SIAM 1992
15. Ericsson, S., Grip, N.: An analysis method for sampling in shift-invariant spaces. Int. J. Wavelet. Multi. Information. Processing. 3 (3) (2005) 301-319
16. Goh, S. S., Ong, I. G. H: Reconstruction of bandlimited signals from irregular samples. Signal. Processing. 46(3) (1995)315-329
17. Jia, R.Q.: Shift-invariant spaces and linear operator equations. Israel Math. J. 103(1998)259-288
18. Lei, J.J, Jia, R.Q., Cheney, E.W.: Approximation from shift-invariant spaces by integral operators. SIAM J. Math. Anal. 28(2) (1997) 481-498
19. Liu, Y.: Irregular sampling for spline wavelet subspaces. IEEE Trans. Information Theory. 42(2) (1996) 623 – 627
20. Liu, Y., Walter, G.G.: Irregular sampling in wavelet subspaces. J. Fourier. Anal. Appl. 2(2) (1995)181-189

21. Xian, J., Lin, W.: Sampling and reconstruction in time-warped spaces and their applications. *Appl. Math. Comput.* 157(2004)153-173
22. Xian, J., Luo, S.P., Lin, W.: Weighted sampling and signal reconstruction in spline subspaces. *Signal. Processing.* 86(2)(2006)331-340
23. Yeung, L.K., Szeto, L.K., Liew, A.W.C., Yan, H.: Dominant spectral component analysis for transcriptional regulations using microarray time-series data. *Bioinformatics.* 20(5)(2004)742-749
24. Golub, G.H., Van Loan, C.F.: *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, (1996)
25. Schumaker, L.L.: *Spline functions: Basic Theory*, Wiley-Interscience, New York (1981)
26. Lomb, N.R.: Least-squares frequency analysis of unequally spaced data. *Astrophys Space Science.* 39(1976)447-462
27. Fisher, R.A.: Tests of significance in Harmonic Analysis. *Proceedings of the Royal Society of London, Series A.* 125(1929)54-59.
28. Davis, H.T.: *The Analysis of Economic Time Series*. Principia Press, Bloomington, Indiana; (1941)
29. Priestley, M.B.: *Spectral Analysis and Time Series*. Academic Press: San Diego. (1981)
30. Oliva, A., Rosebrock, A., Ferrezuelo, F., et al.: The cell cycle-regulated genes of *Schizosaccharomyces pombe*. *Plos Biology.* 3 (7)(2005)1239-1260