Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006

A NEW SPECTRUM ESTIMATION METHOD IN UNEVENLY SAMPLING SPACE

JUN XIAN[1,2], SHUAN-HU WU[1,3], ALAN LIEW[4], DAVID SMITH[5], HONG YAN[1,6]

[1]Department of Electronic Engineering, City University of Hong Kong, Hong Kong
[2]Department of Mathematics, Zhejiang University, Hangzhou, 310027, China
[3]School of Computer Science and Technology, Yantai University, Yantai 264005, China
[4]Department of Computer Science and Engineering Chinese University of Hong Kong, Shatin, Hong Kong
[5]Department of Biochemistry, University of Hong Kong, Pok Fu Lam, Hong Kong
[6]School of Electronic and Information Engineering, University of Sydney, NSW2006, Australia
E-MAIL: junxian@cityu.edu.hk, itwush@cityu.edu.hk, wcliew@cse.cuhk.edu.hk, dsmith@hku.hk, h.yan@cityu.edu.hk

Abstract:

Spectrum estimation is a popular method for identifying periodically expressed genes in microarray time series analysis. For unevenly sampled data, a common technique is applying the Lomb-Scargle algorithm. The performance of this method suffers from the effect of noise in the data. In this paper, we propose a new spectrum estimation algorithm for unevenly sampled data. The new method is based on signal reconstructing technic in aliased shift-invariant signal spaces and a direct spectrum estimation formula was derived based on B-spline basis. The new algorithm is very flexible and can reduce the effect of noise by adjusting the order of B-spline basis. The test on simulated noisy signal and typical periodically expressed gene data shows our algorithm is accurate compared with Lomb-Scargle algorithm.

Keywords:

Spectrum estimation; eriodically expressed gene; unevenly sampled data; Lomb-Scargle algorithm; signal reconstruction; B-spline

1. Introduction

Spectrum 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.

Ruf 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 [6]. Recently, Bohn, Hinderlich, Hütt, Kaiser and Lüttge have used the Lomb-Scargle periodogram to detect rhythmic components in the circadian cycle of the Crassulacean acid metabolism plants [7]. The Lomb-Scargle periodogram was originally developed for analysis of noisy unevenly sampled data from astronomical observations. Since it assumes there is a single stationary sinusoid wave that has infinite support, it may introduce some illusive periodic components for finite data. Also due to the effect of the noise in the data, it may produce inaccurate estimation results.

In this paper, we propose a new spectrum 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 [8-24]. In our method, a direct spectrum estimation formula is derived based on the B-spline basis that has a finite support and its Fourier transform do not introduce illusive components. The proposed algorithm is also flexible and can reduce the effect of noise by adjusting the order of the B-spline basis. Tests on simulated noisy signals and periodically expressed gene data showed our algorithm is accurate compared with the Lomb-Scargle algorithm.

1-4244-0060-0/06/$20.00 ©2006 IEEE

2. Mathematical theory and algorithm

2.1. Mathematical model

In the following, we first review existing work on signal analysis in the shift-invariant signal space, then derive the new spectrum estimation algorithm.

Shannon's signal sampling and reconstruction theorem states:

Theorem A:

If $f \in B_\Omega = \{f \ : \operatorname{supp} \hat{f} \subset [-\Omega, \Omega]\}$ and $0 < T\Omega \le 2\pi$, then

$$f(x) = \sum_{n \in Z} f(nT) \sin c(\Omega(x - nT)) \tag{1}$$

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

Equation (1) shows that the space of bandlimited signal is identical to the space:

$$V(\sin c) = \{f(x) = \sum_{k \in Z} c_k \sin c(x-k) : (c_k) \in \ell^2\} \tag{2}$$

Dowski et al. have introduced a reconstruction formula for unevenly sampled signal that is the special case of (2) [10]:

$$\{f \ : \ f(x) = \sum_{n=0}^{N-1} c_n \sin c(x-n)\}. \tag{3}$$

Since the *sinc* function has an infinite support and slow decay, it is seldom adopted in real applications. In [23], Xian and Lin find 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 introduce a signal space that is called the shift-invariant (also called time-invariant) signal space [8-24]:

$$V(\phi) = \{f : f(x) = \sum_{k \in Z} c_k \phi(x-k) : (c_k) \in \ell^2\} \tag{4}$$

Its aliased version is defined by

$$V_L(\phi) = \{f : f(x) = \sum_{k \in Z} c_k \phi(x - Lk) : (c_k) \in \ell^2\} \tag{5}$$

where $L > 0$ is a constant. Reconstruction coefficients $\{c_i\}$ are related to the choice of basis function, ϕ. We leave the detailed computational procedure to Section 2.2.

Signal reconstructing in the shift-invariant space is an active research area and there are many mathematical theories and computer algorithms on the topic [8-24]. When the signal $f \in V(\phi)$ or $V_L(\phi)$, we hope to reconstruct signal f from sampled value $\{f(x_i)\}$, where $\{x_i\}$ is the sampling point set. If $\{x_i\}$ is an evenly sampling point set, this problem can be regarded as signal reconstruction in an evenly sampling space. Otherwise, this is a signal reconstruction problem in an unevenly sampling space.

In fact, the well-known autoregressive (AR) model can be regarded as a special case of signal reconstruction in the above signal space. For a given discrete data sequence $x[n]$ for $0 \le n \le N-1$, the sample at time index n is approximated by a linear combination of previous K samples in the sequence based on the AR prediction model that can be written as,

$$x[n] = \tilde{x}[n] + e[n] = -\sum_{k=1}^{K} a_k x[n-k] + e[n], \ \ (n \ge K) \tag{6}$$

where $\tilde{x}[n]$ and $e[n]$ represent the estimation of $x[n]$ and the corresponding estimation error, respectively. Comparing (5) with (6), it is obvious that (6) is only a special case of (5).

For the signal reconstructing in an aliased shift-invariant space, we can obtain the following theorem characterizing its energy density spectrum $S_{xx}(\omega)$ according to (5).

Theorem 1. If $f(x) \in V_L(\phi)$, then the energy density spectrum

$$S_{xx}(\omega) = |\sum_k c_k e^{-i2\pi\omega Lk} \hat{\phi}(\omega)|^2 \tag{7}$$

where $\hat{\phi}$ is the Fourier transform of ϕ defined by $\hat{\phi}(\omega) = \int_{-\infty}^{+\infty} \phi(t) e^{-i2\pi\omega t} dt$

Proof: From the definitions of energy density spectrum and the signal space $V_L(\phi)$, we can easily deduce $S_{xx}(\omega) = |\sum_k c_k e^{-i2\pi\omega Lk} \hat{\phi}(\omega)|^2$ according to (5)

In order to avoid instructing illusive periodic components caused by basis function ϕ with infinite support, for example, *sinc*, 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 supp ϕ to indicate the support of basis function ϕ. Assume that $\operatorname{supp} \phi \subset [-\Omega, \Omega]$ and $f(x) \in V_L(\phi)$ that is defined in finite interval, then we have the following theorem:

Theorem 2. If supp $\phi \subset [-\Omega, \Omega]$ and $f(x) \in V_L(\phi)$ defined in the 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

$$f(x) = \sum_{k=\frac{A_1-\Omega}{L}+1}^{\frac{A_2+\Omega}{L}-1} c_k \phi(x - Lk) \tag{8}$$

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

Proof: Since supp $(\phi) \subset [-\Omega, \Omega]$ and $f(x) \in V_L(\phi)$ is defined in the interval $[A_1, A_2]$, we have

$$A_1 - \Omega \le x - \Omega < Lk < x + \Omega \le A_2 + \Omega,$$

that is, $\frac{A_1-\Omega}{L} < k < \frac{A_2+\Omega}{L}$. It follows that

$$f(x) = \sum_{k \in \mathbb{Z}} c_k \phi(x - Lk) = \sum_{k=\frac{A_1-\Omega}{L}+1}^{\frac{A_2+\Omega}{L}-1} c_k \phi(x - Lk).$$

2.2. PSD estimation algorithm

In terms of the definition of the power spectrum density (PSD), we can obtain the following estimation function according to (8)

$$P_{xx}(\omega) = \frac{1}{A_2 - A_1} \left| \sum_{k=\frac{A_1-\Omega}{L}+1}^{\frac{A_2+\Omega}{L}-1} c_k e^{-i2\pi\omega Lk} \hat{\phi}(\omega) \right|^2 \tag{9}$$

Coefficients, $\{c_k\}$, can be calculated according to following steps:

(1) Given sampling points $x_1, \cdots, x_J \in [A_1, A_2]$ and corresponding discrete function value $y = (y_1, \cdots, y_J)$, computing matrix: $U = (U_{jk}),\ T = (T_{kl})$, where

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

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

(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.

Compared with traditional PSD estimation algorithms, our method can directly compute PSD for unevenly sampling signal from (9). Also we can control the effect of noise by selecting different basis function ϕ and parameter L.

3. Experimental results

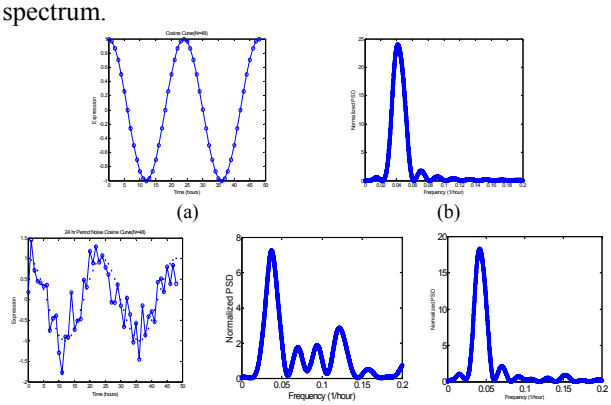
In our numerical test, we choose B-spline of order N as basis function ϕ for spectrum estimation. B-spline of order N can be defined as the convolutions of (N+1) B-splines of order 0, i.e. $\phi = \overbrace{\chi_{[-\frac{1}{2},\frac{1}{2}]} * \cdots * \chi_{[-\frac{1}{2},\frac{1}{2}]}}^{N+1}$. It is obvious that supp $\phi \subset [-\frac{N+1}{2}, \frac{N+1}{2}]$ for B-spline of order N and its Fourier transform can be easily computed as follows: $\hat{\phi} = (\frac{\sin \pi\omega}{\pi\omega})^{N+1}$

Through an easy substitution, we can obtain an explicit repression of PSD estimation as follows:

$$P_{xx}(\omega) = \frac{1}{A_2 - A_1} \left| \sum_{k=\frac{2A_1-N-1}{2L}+1}^{\frac{2A_2+N+1}{2L}-1} c_k e^{-i2\pi\omega Lk} (\frac{\sin \pi\omega}{\pi\omega})^{N+1} \right|^2 \tag{10}$$

We first tested our spectral estimation algorithm on simulated signal for comparing the estimation accuracy with the Lomb-Scargle method. A cosine curve has been used to represent the ideal expression of a gene that goes from "on" state, to an "off" state, and then back to "on" [25]. In Figure 1(a), a cosine curve was generated to simulate the expression of a gene that has a 24-hours period with data samples taken every half-hours and its corresponding periodogram (see Figure 1(b)) showing a peak at the frequency of 1/24 Hz. Figure 1(c) shows a the same cosine signal, but it is now corrupoted with Gassian noise and unevenly sampled. Its periodograms obtained by using Lomb-Scargle method and our algorithm are shown in Figure 1(d) and Figure 1(e), respectively. The frequency corresponding to the peak in the periodograms obtained by using the Lomb-Scargle method and our method is 1/27Hz and 1/24 Hz, respectively (see Figure 1(d) and (e)). Clearly our method is more accurate than the Lomb-Scargle

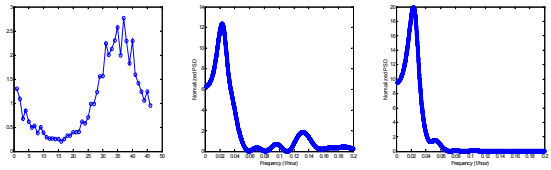
algorithm. Our method also produces less false peak in the spectrum.



(a) (b) (c) (d) (e)

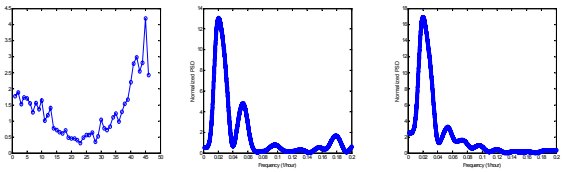
Figure 1. Comparison of spectral estimation for simulated data: (a) simulated cosine signal with even sampling; (b) periodogram of the simulated signal in (a) obtained by using the Fourier transform; (c) simulated noisy cosine signal with uneven sampling; (d) periodogram obtained by using the Lomb-Scargle method for the signal in (c); (e) periodogram obtained by our method for the signal in (c).

We have also tested our algorithm on real gene expression data of Plasmodium falciparum, which is one of the species that cause human malaria [26]. The gene expression time series from the asexual intraerythrocytic developmental cycle (IDC) of Plasmodium falciparum are strongly periodic. Identifying periodically expressed genes is useful for understanding the genome of Plasmodium falciparum and designing effective vaccines for prevention of human malaria. In the gene expression database from [26], data values at 23-rd and 29-th hours are completely missing. With missing values, the time series can be in general treated as unevenly sampled data. An example of gene expression profile from the database is shown in Figure 2(a), and the periodograms obtained by using the Lomb-Scargle algorithm and our algorithm are shown in Figure 2(b) and Figure 2(c), respectively. The frequency corresponding to the peak in the periodograms obtained by usingthe Lomb-Scargle method and our method is 1/44.15 Hz and 1/43.23 Hz, respectively (see Figure 1(b) and (c)). Another example is shown in Figure 3. The frequency corresponding to the peak in the periodograms obtained by using the Lomb-Scargle method and our method is 1/48.78 Hz and 1/49.75 Hz, respectively (see Figure 1(b) and (c)). We can see from these diagrams that our algorithm can effectively reduce the spurious oscillation components in the spectra.



(a) (b) (c)

Figure 2. Spectral estimation for the expression time series of gene a12797_1 in the Plasmodium falciparum microarray data: (a) gene expression pattern; (b) periodogram obtained by using the Lomb-Scargle method; (c) periodogram obtained by using our algorithm.



(a) (b) (c)

Figure 3. Spectral estimation the gene expression of gene 13725_4 in the Plasmodium falciparum microarray data: (a) gene expression pattern; (b) periodogram obtained by using the Lomb-Scargle method; (c) periodogram obtained by using our algorithm.

4. Conclusions

In this paper, we have proposed a new spectrum estimation algorithm based on signal reconstructing technique in an unevenly sampled space. The advantages of our algorithm over conventional the Lomb-Scargle spectral estimation method is that new algorithm can effectively reduce the effects of noise and spurious oscillation components and therefore improve the estimation accuracy. Also new algorithm is flexible since the order of B-spline basis function can be adjusted. Experiments on simulated signal and real gene expression data show that our method is effective and can be applied to identifying periodically expressed genes.

Acknowledgements

This work is supported by a CityU interdisciplinary research grant (project 9010003) and a grant from the Hong Kong Research Grant Council (project CityU122005) and the Mathematical Tianyuan Foundation of China NSF (10526036).

References

[1] M. H. Hayes, Statistical digital signal processing and modeling. John Wiley and Sons, Inc, 1996.
[2] S. M. Kay, S. L. Marple Jr, "Spectrum analysis-A modern perspective", Proceeding of IEEE, Vol 69, No. 11, pp. 1380-1418, 1981.
[3] S. Chu, J. DeRisi, et al., "The transcriptional program of sporulation in budding yeast", Science. Vol 282, pp. 699-705, 1998.
[4] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", PNAS. Vol 95, pp. 14863-14868, 1998.
[5] T. S. Spellman, G. Sherlock, et al., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisia by microarray hybridization", Mol. Biol. Cell. Vol 9, pp. 3273-3297 , 1998.
[6] T. Ruf, "The Lomb-Scargle Periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series", Biological Rhythm Research, Vol 30, pp. 178-201, 1999.
[7] A. Bohn, S. Hinderlich, M-T. Hütt, F. Kaiser and U. Lüttge, "Identification of rhythmic subsystems in the circadian cycle of crassulacean acid metabolism under thermoperiodic perturbations", Biol. Chem.,Vol 384, pp. 721-728, 2003.
[8] C. K. Chui, An introduction to Wavelet, Academic Press, New York, 1992
[9] I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.
[10] E. R. Dowski, C. A. Whitmore and S. K. Avery, "Estimation of randomly sampled sinusoids in additive noise", IEEE Trans. Acoustics. Speech. Signal Processing. Vol 36, No 12, pp. 1906-1908, 1988.
[11] W. Chen, S. Itoh and J. Shiki, "Irregular sampling theorem for wavelet subspaces", IEEE Trans. Information Theory, Vol 44, No. 3, pp. 1131-1141, 1998.
[12] S. Ericsson and N. Grip, "An analysis method for sampling in shift-invariant spaces", Int. J. Wavelet. Multi. Information. Processing. Vol 3, No. 3, pp. 301-319, SEP. 2005
[13] S. S. Goh, I. G. H. Ong, "Reconstruction of bandlimited signals from irregular samples", Signal. Processing. Vol 46, No. 3, pp. 315-329, 1995.
[14] J. J. Benedetto, "Irregular sampling and frams", in: C. K. Chui (Ed.), wavelets: A Tutorial in theory and Applications. pp. 445-507, 1992.
[15] W. Chen, S. Itoh and J. Shiki, "On sampling in shift invariant spaces", IEEE Trans. Information Theory, Vol 48, No. 10, pp. 2802 – 2810, 2002.
[16] R. Q. Jia, "Shift-invariant spaces and linear operator equations", Israel Math. J., Vol 103, pp. 259-288, 1998.
[17] J. J. Lei, R. Q. Jia and E. W. Cheney, "Approximation from shift-invariant spaces by integral operators", SIAM J. Math. Anal., Vol 28, No. 2, pp. 481-498, 1997.
[18] W. Chen, B. Han, R. Q. Jia, "Estimate of aliasing error for non-smooth signals prefiltered by quasi-projections into shift-invariant spaces", IEEE Trans. Signal. Processing. Vol 53, No. 5, pp.1927-1933, 2005
[19] R. M. Lewitt, "Alternatives to voxels for image representation in iterative reconstruction algorithm", Phys. Med. Biol, Vol 37, pp. 705-716, 1992.
[20] Y. Liu, "Irregular sampling for spline wavelet subspaces", IEEE Trans. Information Theory, Vol 42, No. 2, pp. 623 – 627, 1996.
[21] Y. Liu and G. G. Walter, "Irregular sampling in wavelet subspaces", J. Fourier. Anal. Appl., Vol 2, No. 2, pp. 181-189, 1995.
[22] W. Chen, B. Han, R. Q. Jia, "On simple oversampled A/D conversion in shift-invariant spaces", IEEE Trans. Inform. Theory. Vol 51, No. 2, pp. 648-657, 2005.
[23] J. Xian, W. Lin, "Sampling and reconstruction in time-warped spaces and their applications", Appl. Math. Comput, Vol 157, pp. 153-173, 2004.
[24] J. Xian, S. P. Luo and W. Lin, "Weighted sampling and signal reconstruction in spline subspaces", Signal. Processing, Vol 86, No. 2, pp. 331-340, 2006.
[25] S. Wichert, "Identifying periodically expressed transcripts in microarray time series data", Bioinformatics, Vol 20, pp. 5-20, 2004.
[26] Z. Bozdech, M. Llinas, B. L.Pulliam, E. D. Wong,et al., "The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum", PLoS Biology, Vol 1, pp. 1-16, 2003.
[27] L. K. Yeung, L. K. Szeto, A. W. C. Liew, H. Yan, "Dominant spectral component analysis for transcriptional regulations using microarray time-series data", Bioinformatics, Vol 20, No. 5, pp. 742-749, 2004.