Bayesian Spectrum Estimation of Unevenly Sampled Nonstationary Data

EDICS: 2-TIFR, 2-SPEC

Yuan Qi* , Thomas P. Minka, and Rosalind W. Picard

Y. Qi is with Media Laboratory of MIT, Cambridge, Massachusetts, 02139, U.S.A., (E-mail: yuanqi@media.mit.edu)

T. P. Minka is with Department of Statistics of Carnegie Mellon University, Pittsburgh, Pennsylvania, 15123, U.S.A., (E-mail: minka@stat.cmu.edu)

R. W. Picard is with Media Laboratory of MIT, Cambridge, Massachusetts, 02139, U.S.A.,(E-mail: picard@media.mit.edu),

Abstract

Spectral estimation methods typically assume stationarity and uniform spacing between samples of data. The non-stationarity of real data is usually accommodated by windowing methods, while the lack of uniformly-spaced samples is typically addressed by methods that "fill in" the data in some way. This paper presents a new approach to both of these problems: we use a Bayesian framework, which includes a non-stationary Kalman filter, to jointly estimate all spectral coefficients instantaneously. The new method works regardless whether the samples are evenly or unevenly spaced; moreover, it provides a new approach to enabling processing when it is desirable to virtually eliminate aliasing by unevenly sampling. An amplitude-preservation property of the new method can be used to detect if aliasing occurred. Finally, we propose an efficient algorithm for sparsifying the spectrum estimates when we know a priori that the signal is narrow-band in the frequency domain. We illustrate the new method on several data sets, showing that it can perform well on unevenly sampled nonstationary signals without the use of any sliding window, that it can estimate frequency components beyond half of the average sampling frequency when the signal is unevenly sampled, and that it can provide more accurate estimation than several other important recent and classical methods.

I. Introduction

Spectrum estimation has been a classical research topic in signal processing communities for decades. Many approaches have been proposed, including the modified periodogram, estimation based on auto-regressive modeling, the MUSIC algorithm, and the Multitaper method [1], [2], [3]. Although all these algorithms have their own advantages, they all have two basic assumptions: first, the signal samples are evenly spaced; second, the signal is stationary, at least over the duration of a window. But in many real world applications– for example in several electrocardiogram analysis problems where frequency analysis is performed on the beat arrival times –signals are unevenly sampled and nonstationary.

Unevenly sampled signal may be obtained for many reasons including the random nature of the sampling time, missing data, and deterministically designed sampling schemes. Compared to the rich research activities on spectral estimation of evenly sampled signals, there has been less research done for unevenly sampled signals. In the following paragraphs, we briefly review some of the previous work on spectral analysis of unevenly sampled signals.

To estimate power spectra of a laser Doppler velocimetry signal that is unevenly sampled, Ouahabi et al. [4] first interpolate the signal to be evenly spaced and then apply classical FFT-based methods, while Banning [5] models the sampling time as a Poisson distribution and uses

Kalman filtering to estimate evenly spaced "samples". Also, Dowski [6] utilizes interpolation to obtain spectrum estimates from unevenly sampled signals.

Besides using interpolation, another direction is to perform spectrum estimation directly from the unevenly spaced samples. Lomb and Scargle proposed the Lomb-Scargle periodogram to deal with unevenly sampled data The Lomb-Scargle periodogram [7], [8] models the data as a single stationary sinusoid wave; this method was originally proposed using the Maximum likelihood principle. Later Bretthorst gave it a Bayesian interpretation based on Laplace's approximation [9]. The Lomb-Scargle periodogram has been applied to economic time series, Wolf's relative sunspot numbers, and heart rate data [9], [10]. Also, Bronez applies generalized prolate spheroidal sequences to spectrum estimation of unevenly sampled data [11]. Due to the high computation burden of this method, he suggests to use approximation techniques in practice.

For nonstationary signals, most methods explicitly or implicitly use sliding windows, such as short-time FFT and time-varying multitaper methods [12], [13].

In this paper, we propose a new Bayesian spectrum estimation method for unevenly sampled nonstationary data. This method models the signal by a linear dynamic system and formulates spectrum estimation as a probabilistic inference problem. The new method has the following main features:

a. Instantaneously estimating all the spectral coefficients at the time a new sample arrives, while neither doing interpolation when samples are unevenly spaced, nor using a sliding window when the signal has a time-varying spectrum.

b. Jointly estimating all the frequencies, while many classical methods, for example, the Lomb-Scargle periodogram, estimate each of the frequencies separately.

c. Preserving the virtually-aliasing-free property of unevenly sampled data, and providing a means of detection of aliasing via an amplitude conservation property.

d. Modeling the observation noise, in contrast with most spectrum estimation or time-frequency analysis methods that are based on deterministic transformations without noise modeling or stochastic signal modeling.

e. Applying Bayesian inference to spectrum estimation. In this sense, this method can be viewed as an extension of the Lomb-Scargle periodogram. Instead of just assigning values to spectral coefficients, it provides a joint probability distribution over spectral coefficients, and easily allows incorporation of prior information about this distribution.

Furthermore, we propose an efficient maximum likelihood method to sparsify spectral coefficients, which is useful for narrow-band signals.

The rest of the paper is organized as follows. Section II presents the new spectrum estimation method, and discusses its properties, followed by section III, describing a Kalman-based approach for updating probabilities. Section IV introduces the sparsification algorithm for the narrow-band case. Section V compares the new method with classical spectrum estimation methods, demonstrates its accuracy, manifests that it can correctly estimate frequencies beyond the range of half of the average sampling frequency, and discusses how it can be applied to resolving ambiguity in time-varying spectrum estimation. Finally, section VI summarizes the new method and describes future research directions.

II. A Bayesian Framework for Nonstationary Spectrum Estimation

In this section, we present a Bayesian framework for estimating the nonstationary spectrum of a given signal. This framework does not assume any short time stationarity to the signals, in contrast with classical spectrum estimation approaches. The method works both for evenly and unevenly sampled data.

For the spectrum estimation problem, we observe the data $\mathbf{x}$: $\mathbf{x} = [x_1, x_2, \ldots, x_n, \ldots, x_N]^T$, where x_n is sampled at time t_n. When the data is unevenly sampled, $\mathbf{t} = [t_1, \ldots, t_N]^T$ contains useful information for spectrum estimation. We model the data as

$$x_n = a_{n0} + \sum_{k=1}^{M} a_{nk} \sin(2\pi f_k t_n) + b_{nk} \cos(2\pi f_k t_n) + v_n \qquad (1)$$
$$\text{for } n = 1, \ldots, N.$$

where v_n is a noise variable. The number and value of frequency bases, M and f_k, can be chosen based on prior knowledge. These frequency bases can be made to be evenly or unevenly spaced in the frequency domain; however, later in this paper we simply choose all the frequency bases to be equally spaced so that equation (1) is a truncated Fourier expansion. The signal energy will project on these sinusoid and cosine bases.

Both a_{nk} and b_{nk} have real values. Note that for a nonstationary signal, a_{nk}, b_{nk}, and v_n depend on the sampling time t_n. The use of a_{nk} and b_{nk} allows the signal to have a time-varying amplitude $\sqrt{a_{nk}^2 + b_{nk}^2}$ and a changing phase $\arctan(\frac{b_{nk}}{a_{nk}})$ for the k^{th} frequency band at time t_n .

For equation (1), we define

$$\begin{aligned} \mathbf{s}_n &= [a_{n0}, a_{n1}, a_{n2}, \ldots, a_{nM}, b_{n1}, b_{n2}, \ldots, b_{nM}]^T && (2) \\ \mathbf{c}_n &= [1, \sin(2\pi f_1 t_n), \ldots, \sin(2\pi f_M t_n), \\ &\quad \cos(2\pi f_1 t_n), \ldots, \cos(2\pi f_M t_n)] && (3) \end{aligned}$$

For nonstationary spectrum estimation, our goal is to estimate the state vector $\mathbf{s}_n$ instantaneously at the sampling time t_n. To this end, we assume that the hidden states $\mathbf{s}_1 \ldots \mathbf{s}_N$ form a Markov chain that emits a time series of observations $x_1 \ldots x_N$:

$$\begin{aligned} \mathbf{s}_n &= \mathbf{s}_{n-1} + \mathbf{w}_n && (4) \\ x_n &= \mathbf{c}_n \mathbf{s}_n + v_n && (5) \end{aligned}$$

where $\mathbf{w}_n$ is the process noise at the sampling time t_n, and v_n is the observation noise at t_n. We can model the process observation noises by Gaussian distributions or by heavy-tailed non-Gaussian distributions. However, using non-Gaussian distributions invokes the use of numerical approximation techniques in the inference procedure.

According to this model, the joint distribution of hidden states and observations can be computed as

$$p(\mathbf{s}_{1:N}, x_{1:N}) = p(x_1|\mathbf{s}_1)p(\mathbf{s}_1) \prod_{n=2}^{N} p(x_n|\mathbf{s}_n)p(\mathbf{s}_n|\mathbf{s}_{n-1}) \quad (6)$$

where $\mathbf{s}_{1:N} = [\mathbf{s}_1, \ldots, \mathbf{s}_N]^T$ and $x_{1:N} = [x_1, \ldots, x_N]^T$ denotes collections of states and observations from time t_1 to t_N.

The filtering distribution $p(\mathbf{s}_n|x_{1:n})$ can be sequentially estimated as follows

$$\begin{aligned} p(\mathbf{s}_n|x_{1:n-1}) &= \int_{\mathbf{s}_{n-1}} p(\mathbf{s}_n|\mathbf{s}_{n-1})p(\mathbf{s}_{n-1}|x_{1:n-1}) && (7) \\ p(\mathbf{s}_n|x_{1:n}) &= \frac{p(\mathbf{x}_n|\mathbf{s}_n)p(\mathbf{s}_n|x_{1:n-1})}{p(x_n|x_{1:n-1})} && (8) \end{aligned}$$

Then the spectrum at time t_n can be summarized by the mean of $p(\mathbf{s}_n|x_{1:n})$.

III. Spectrum Estimation by Kalman Filtering

A. Algorithm

If we use linear Gaussian models in equations (4) and (5):

$$\mathbf{w}_n \sim \mathcal{N}(\mathbf{0}, \Gamma_n) \tag{9}$$

$$v_n \sim \mathcal{N}(\mathbf{0}, \sigma_n^2), \tag{10}$$

then $p(\mathbf{s}_n|x_{1:n-1})$ is also Gaussian, and we can use Kalman filtering to efficiently update these probabilities. To deal with uneven sampling, we set the variance of the process noise proportional to the time interval between two consecutive samples,i.e.,

$$\Gamma_n = \mathbf{Z}(t_n - t_{n-1}); \tag{11}$$

where $\mathbf{Z}$ is a pre-defined constant matrix, which we say more about below. The intuition behind this equation is that the longer the sampling interval between time t_n and t_{n+1}, the larger the uncertainty about the spectral coefficients at time t_{n+1} conditional on those at time t_n.

Denote $\mathbf{m}_n$ and $\mathbf{V}_n$ as the mean and covariance matrix of $p(\mathbf{s}_n|x_{1:n})$. We have the following Kalman filtering update [14], [15] equations:

$$\mathbf{m}_n = \mathbf{m}_{n-1} + \mathbf{K}_n(x_n - \mathbf{c}_n\mathbf{m}_{n-1}) \tag{12}$$

$$\mathbf{V}_n = (\mathbf{I} - \mathbf{K}_n\mathbf{c}_n)\mathbf{P}_{n-1} \tag{13}$$

where

$$\mathbf{P}_{n-1} = \mathbf{V}_{n-1} + \Gamma_{n-1} \tag{14}$$

$$\mathbf{K}_n = \mathbf{P}_{n-1}\mathbf{c}_n^T(\mathbf{c}_n\mathbf{P}_{n-1}\mathbf{c}_n^T + \sigma_n^2)^{-1} \tag{15}$$

and $\mathbf{I}$ is an identity matrix. Note that we have a nonstationary Kalman filtering algorithm; both $\mathbf{c}_n$ and Γ_{n-1} vary with time.

The recursions start off with

$$\mathbf{m}_1 = \mathbf{m}_0 + \mathbf{K}_1(x_1 - \mathbf{c}_1\mathbf{m}_0) \tag{16}$$

$$\mathbf{V}_1 = (\mathbf{I} - \mathbf{K}_1\mathbf{c}_1)\mathbf{V}_0 \tag{17}$$

$$\mathbf{K}_1 = \mathbf{V}_0\mathbf{c}_1^T(\mathbf{c}_1\mathbf{V}_0\mathbf{c}_1^T + \sigma_1^2)^{-1} \tag{18}$$

where $\mathbf{m}_0$ and $\mathbf{V}_0$ are pre-defined hyper-parameters for the prior distribution $p(\mathbf{s}_0)$, which we say more about below.

If we want to utilize not only the past information, but also the future information in the data set to estimate the spectrum, we may want to compute $p(\mathbf{s}_i|x_{1:N})$ where $x_{1:N}$ is the whole data set. As a well-known technique, Kalman smoothing can be employed to compute this posterior distribution [14].

B. Model Parameters and Hyperparameters

Given the algorithm in the last section, it is natural to ask how to assign the values of the Gaussian variances, Γ_n and σ^2, and the hyperparameters $\mathbf{m}_0$ and $\mathbf{V}_0$. As a Bayesian method, the algorithm allows you to easily incorporate prior knowledge into the estimation, which on the other hand is relatively hard for a Maximum likelihood approach.

First, if we have no information about the frequency characteristics of the data, we may set Γ_n or more exactly $\mathbf{Z}$ in equation (11) to be a scaled identity matrix

$$\mathbf{Z} = z\mathbf{I} \tag{19}$$

where z controls the variablity of the amplitude of the estimated frequencies. Also, we may use a so-called noninformative prior distribution $p(\mathbf{s}_0)$ by assigning $\mathbf{V}_0$ to be a scaled identity matrix, and $\mathbf{m}_0$ to be a zero vector.

Second, if we think the data might contain only some known frequencies, a simply way of representing this belief is to only use those frequency bases in the data model (1). But if we are not certain if there are other frequency components in the data, this approach will be too aggressive. A better approach will be assigning $\mathbf{V}_0$ to be a diagonal matrix with small variances for the 0 elements in $\mathbf{m}_0$. In this way, we are more likely to obtain the nonzero estimate of the preferred frequencies. But at the same time, we can still obtain nonzero estimates of the other frequencies if the data suggest that such exist.

Third, since for natural signals the amplitudes of low frequency components may tend to change slower than those of the high frequency components, we might want to assign smaller process noise for low frequency components, and larger process noise for high frequency components. By doing so, more data points are needed to change the estimates of low frequency coefficients and fewer data points to change the estimates at high frequencies. As a popular time-frequency analysis tool, wavelets share the similar property.

In summary, there is room in this new algorithm to represent prior knowledge and use it to guide the estimation while employing information from the data at the same time.

C. No Fixed Window

It appears a little surprising that this new method can estimate the frequencies of a periodic signal without imposing any window on the signal. A closer look at the formula will help illuminate what is going on.

Let us rewrite equation (12) in another form:

$$\mathbf{m}_n = (\mathbf{I} - \mathbf{K}_n\mathbf{c}_n)\mathbf{m}_{n-1} + \mathbf{K}_n x_n \tag{20}$$

$$= (\mathbf{I} - \mathbf{K}_n\mathbf{c}_n)(\mathbf{I} - \mathbf{K}_{n-1}\mathbf{c}_{n-1})\mathbf{m}_{n-2} + (\mathbf{I} - \mathbf{K}_n\mathbf{c}_n)\mathbf{K}_{n-1}x_{n-1} + \mathbf{K}_n x_n \tag{21}$$

$$= \cdots$$

$$= \prod_{d=1}^{n}(\mathbf{I} - \mathbf{K}_d\mathbf{c}_d)\mathbf{m}_0 + \sum_{k=1}^{n-1}\prod_{d=k+1}^{n}(\mathbf{I} - \mathbf{K}_d\mathbf{c}_d)\mathbf{K}_k x_k + \mathbf{K}_n x_n \tag{22}$$

Define $\mathbf{K}_0 = 1$ and $g_k = \mathbf{K}_k \prod_{d=k+1}^{n}(\mathbf{I}-\mathbf{K}_d\mathbf{c}_d)$ for $k = 0, 1, \ldots, (n-1)$. From equation (22), we can see that $\mathbf{m}_n$ is a weighted average of x_k and $\mathbf{m}_0$, where g_k serves as the weighting coefficient for the k^{th} term.

Thus, the new method can be considered to construct an adaptive weighted window, which decays fast over the past history. Both observation and process noise variances play important roles in the weighting coefficient g_k. The sampling time t_k also affects g_k through its influence on Γ_k.

For example, consider $\mathbf{K}_n x_n$ in equation (22). If σ_n^2 is large, then due to equation (15) $\mathbf{K}_n$ is small. In other words, the influence of the current noisy obervation x_n on the spectral estimate $\mathbf{m}_n$ is damped. Similarly when the data points before the k^{th} data point are clean or the k^{th} data point is noisy, the influence of x_k on $\mathbf{m}_n$ is reduced by a small g_k.

Finally, we want to emphasize that this adaptive windowing mechanism works for the new estimation method automatically given the model and the data. In contrast with other sliding window methods, there is no need to manually tune explicit parameters for a window shape and size, or for a set of such windows.

D. Conquer Aliasing by Unevenly Sampling

Recently [16], Bretthorst showed that a generalization of the discrete Fourier transformation (DFT) can handle the case when data is unevenly sampled, resulting in a much larger effective bandwidth than when the DFT is used on evenly sampled data. For the new Bayesian spectrum estimation method, the similar effect of uneven sampling holds: the critical frequency beyond which aliasing occurs may be almost infinite for unevenly sampled data.

Let us first review the reason why aliasing exists for evenly sampled data. When the data are evenly sampled, we have

$$t = n\Delta t, \quad \text{for } n = 1, \ldots, N. \tag{23}$$

$$f_s = \frac{1}{\Delta t} \tag{24}$$

$$f_{Ny} = \frac{f_s}{2} = \frac{1}{2\Delta t} \tag{25}$$

where Δt is the time interval between any two consecutive samples, f_s is the sampling frequency, and f_{Ny} is the Nyquist frequency. In this case, it is well known that aliases occur at multiples of the sampling frequency: evenly spaced samples of both $x(t) = A\cos(2\pi f_0 t + \phi)$ and $y(t) = A\cos(2\pi(f_0 + kf_s)t + \phi)$, replacing $t = n\Delta t$, will be identical for all integer k . When components of (3) include such aliases, then these repeated components will receive similar probabilities.

In contrast, if the data are unevenly sampled, so that the time intervals between two samples may differ, then we can denote the largest common factor of all t_n's as $\Delta t'$. Then it follows

$$t_n = k_n \Delta t', \quad \text{for } n = 1, \ldots, N. \tag{26}$$

where k_n is an integer. For evenly sampled data, $k_n = 0, 1, \ldots, N-1$. For unevenly sampled data, k_n may start from a large number.

Then we define the new cut-off frequency f'_{Ny} for irregular sampled data as

$$f'_{Ny} = \frac{1}{2\Delta t'} \tag{27}$$

Note that $\Delta t'$ is less than or equal to the smallest time interval between data points. When the sampling is random, $\Delta t'$ may be as small as the numerical resolution of the system. For example, if t_n is stored by a 32 bit number, $\Delta t'$ will be around 2^{-32} and f'_{Ny} will be around 2^{31} Hz.

In other words, when the data are randomly sampled, or unevenly sampled in a well-designed way, the new spectrum estimate will have an almost infinite cut-off frequency f'_{Ny}, thus providing an effectively infinite bandwidth. This effect is illustrated in Fig. 1.

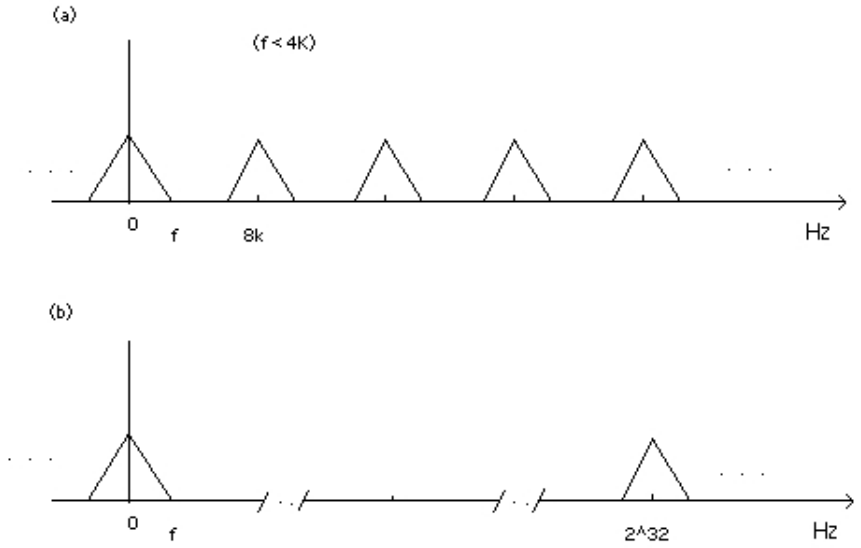


Fig. 1. The effect of uneven sampling on spectra estimated by the new method. (a) Spectrum of signal evenly sampled at 8KHz. The bandwidth of the signal is less than 8K/2 = 4K Hz. With the Nyquist frequency at 4K Hz, the first alias will appear around 8K Hz. (b) Spectrum of signal randomly sampled at 2^{32} Hz. With the new cut-off frequency at 2^{31} Hz, the first alias will appear around 2^{32} Hz.

Finally, notice that this "essentially no aliasing" property does not imply preservation of estimation accuracy when the average sampling rate declines dramatically, even if the samples are unevenly spaced. Dramatically reducing aliasing is not the same as preserving estimation accuracy. As the number of samples decreases, the estimation accuracy will decrease smoothly.

E. Advantages of Joint Estimation

As shown by Bretthorst [16], the power spectrum of the (generalized) DFT actually utilizes a single frequency model. In order to obtain the estimation for different frequencies, the (generalized) DFT basically applies the same model to every frequency basis and repeats the estimation procedure again and again. Similarly, the Lomb-Scargle periodogram has a single frequency model.

However, for the new method, all the frequencies in the spectrum (1) are used together to explain the data and jointly estimated, which gives the new method several estimation advantages over the methods based upon a single frequency model. These advantages are discussed in the remainder of this section.

E.1 Amplitude Conservation and Aliasing Detection

For the DFT and the Lomb-Scargle periodogram, when aliases appear, the amplitude estimates of the aliases will be the same regardless of how many aliases are present over the frequency range being considered; that is, the amplitude estimation of one frequency is independent of the estimation of the others. This behavior is a consequence of both the DFT and the Lomb-Scargle periodogram using a single frequency model that treats each frequency separately.

But for the new method,when aliasing occurs, the signal energy will be equally distributed on all the repeated elements in the basis $\mathbf{c}_n$ due to the symmetry of the projection on these identical elements. Then it is easy to see from equation (1) that the sum of the absolute values of the projection coefficients on two identical frequency elements in $\mathbf{c}_n$ will be the same as the absolute value of the projection coefficient on one of them if we remove the other one from the basis $\mathbf{c}_n$. In other words, the sum of the spectral coefficients is the same no matter if we use a larger or smaller frequency basis $\mathbf{c}_n$. We call this behavior the amplitude conservation property, and emphasize that it arises because of the joint estimation of all the frequencies. The DFT and the Lomb-Scargle periodogram do not have this property.

The amplitude conservation property of the new method can be used for aliasing detection. Specifically, we can reduce the frequency range of the sinusoid and cosine basis to see if amplitude estimates change or not, in order to see if there are aliases in the estimation. In contrast, for the DFT and the Lomb-Scargle periodogram, this property does not hold.

E.2 Super-resolution

Now consider the case when a signal contains multiple frequency components. If the frequency components are well-separated in the frequency domain, using a method based on a single frequency model may achieve a good approximation since there is not much estimation interference between those frequency components. But if the frequency components are very close to each other, the single-frequency methods suffer from interference between the closely-spaced frequency components, which in turn reduces the estimation accuracy.

On the other hand, by jointly estimating all the frequency components and conditioning on the estimation results of past data points, the new method can separate very close frequency components. This will be demonstrated in Sec. V.

IV. Sparse Estimation

If a signal contains only a few frequency components, and we know this a priori, then we can sparsify the Kalman filtering result in order to reduce the number of estimated frequency components and remove the estimation ambiguity as discussed later in Sec. V-D.

In this section, we propose an efficient sparsification algorithm. This algorithm can be viewed as a generalization of the Optimal Brain Surgeon algorithm [17] that has been proposed in the machine learning community for pruning irrelevant features.

A. Sparsification by Lagrangian Optimization

To sparsify the mean estimation $\mathbf{m}$ of the spectrum, we maximize the probability of the estimation at that time t_n while setting some values in the mean vector $\mathbf{m}$ to be zeros. For the simplicity of the notation, we drop the subscript n in $\mathbf{m}_n$, $\mathbf{V}_n$, and $\mathbf{s}_n$ in this section. Let $\mathbf{m}_\star$ be the new mean vector after pruning some elements to be zeros and modifying the rest of the nonzero elements, and let the vector $\mathbf{q}$ indicate which elements of $\mathbf{m}$ are pruned to obtain $\mathbf{m}_\star$.

Formally speaking, we have the following problem:

$$\max_{\mathbf{q}} \min_{\mathbf{m}_\star} \log p(\mathbf{s}|\mathbf{m}_\star, \mathbf{V}) \tag{28}$$

$$\text{i.e. } \max_{\mathbf{q}} \min_{\mathbf{m}_\star} -\frac{1}{2}(\mathbf{s}-\mathbf{m}_\star)^T\mathbf{V}^{-1}(\mathbf{s}-\mathbf{m}_\star) \tag{29}$$

Now, let h be the length of $\mathbf{m}$, and let $\mathbf{e}_i$ be a vector with all elements being zero except its i^{th} element being one. Then, $\mathbf{E_q}$ is the matrix obtained by extracting columns specified in $\mathbf{q}$ from the h by h identity matrix. For example, if $\mathbf{q} = [1, 65]$, then $\mathbf{E_q}$ consists of the first and sixty-fifth column of the h by h identity matrix.

The above optimization problem can then be shown to be equivalent to the following one:

$$\min_{\mathbf{q}} \min_{\delta\mathbf{m}} \quad \delta\mathbf{m}^T\mathbf{V}^{-1}\delta\mathbf{m} \tag{30}$$

Subject to

$$\mathbf{E}_\mathbf{q}^T\delta\mathbf{m} - \mathbf{m_q} = \mathbf{0} \tag{31}$$

where $\delta\mathbf{m} = \mathbf{m} - \mathbf{m}_\star$.

We form a Lagrangian from equations (30) and (31):

$$l = \delta\mathbf{m}^T\mathbf{V}^{-1}\delta\mathbf{m} + \boldsymbol{\lambda}^T(\mathbf{E}_\mathbf{q}^T\delta\mathbf{m} - \mathbf{m_q}) \tag{32}$$

where $\boldsymbol{\lambda}$ is a Lagrange multiplier.

By taking derivatives of l and using the constraint (31), we obtain

$$\delta\mathbf{m} = \mathbf{V}_{1:M;\mathbf{q}}(\mathbf{V}_{\mathbf{q};\mathbf{q}})^{-1}\mathbf{m}_{\mathbf{q}} \tag{33}$$

$$l_{\mathbf{q}} = \mathbf{m}_{\mathbf{q}}^T(\mathbf{V}_{\mathbf{q};\mathbf{q}})^{-1}\mathbf{m}_{\mathbf{q}} \tag{34}$$

where $\mathbf{V}_{1:M;\mathbf{q}}$ consists of the $\mathbf{q}$ columns of the matrix $\mathbf{V}$, $\mathbf{V}_{\mathbf{q};\mathbf{q}}$ the intersection of $\mathbf{q}$ columns and $\mathbf{q}$ rows of $\mathbf{V}$, and $\mathbf{m}_{\mathbf{q}}$ the $\mathbf{q}$ elements of $\mathbf{m}$.

If we want to prune l elements of $\mathbf{m}$ to be zeros, we compute $l_{\mathbf{q}}$ for all possible vectors $\mathbf{q}$ that indicate the positions of those l elements. Then we set the optimal $\mathbf{q}^*$ as

$$\mathbf{q}^* = \arg\min_{\mathbf{q}} l_{\mathbf{q}},$$

and compute $\delta\mathbf{m}$ based on $\mathbf{q}^*$.

B. Using Schur Complements

The above sparsification algorithm involves a matrix inverse $(\mathbf{V}_{\mathbf{q};\mathbf{q}})^{-1}$, which becomes expensive when we want to prune many elements in $\mathbf{m}$. Using Schur Complements, we can efficiently obtain $\delta\mathbf{m}$ and l_q without computing $(\mathbf{V}_{\mathbf{q};\mathbf{q}})^{-1}$ explicitly.

Denote by $\mathbf{H}$ the inverse of the covariance matrix $\mathbf{V}$. and partition $\mathbf{H}$ into four sub-matrices:

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{\mathbf{p};\mathbf{p}} & \mathbf{H}_{\mathbf{p};\mathbf{q}} \\ \mathbf{H}_{\mathbf{q};\mathbf{p}} & \mathbf{H}_{\mathbf{q};\mathbf{q}} \end{pmatrix} \tag{35}$$

where $\mathbf{p}$ is the complementary vector of $\mathbf{q}$. For example, if $\mathbf{H}$ is a 3 by 3 matrix and $\mathbf{q} = [1,\ 3]$, then $\mathbf{p} = [2]$ and

$$\mathbf{H}_{\mathbf{p};\mathbf{p}} = \mathbf{H}_{2;2} \qquad \mathbf{H}_{\mathbf{p};\mathbf{q}} = [\mathbf{H}_{2;1}\mathbf{H}_{2;3}]. \tag{36}$$

As shown in the appendix, we have

$$\delta\mathbf{m} = \mathbf{V}_{1:M;\mathbf{q}}(\mathbf{H}_{\mathbf{q};\mathbf{q}} - \mathbf{H}_{\mathbf{q};\mathbf{p}}\mathbf{H}_{\mathbf{p};\mathbf{p}}^{-1}\mathbf{H}_{\mathbf{p};\mathbf{q}})\mathbf{m}_{\mathbf{q}} \tag{37}$$

$$\begin{aligned} l_{\mathbf{q}} &= \mathbf{m}^T\mathbf{H}\mathbf{m} - \mathbf{m}_{\mathbf{p}}^T\mathbf{H}_{\mathbf{p};\mathbf{p}}\mathbf{m}_{\mathbf{p}} - 2\mathbf{m}_{\mathbf{p}}^T\mathbf{H}_{\mathbf{p};\mathbf{q}}\mathbf{m}_{\mathbf{q}} - \\ &\quad \mathbf{m}_{\mathbf{q}}^T\mathbf{H}_{\mathbf{p};\mathbf{q}}^T\mathbf{H}_{\mathbf{p};\mathbf{p}}^{-1}\mathbf{H}_{\mathbf{p};\mathbf{q}}\mathbf{m}_{\mathbf{q}} \\ &= \mathbf{m}^T\mathbf{H}\mathbf{m} - \mathbf{m}_{\mathbf{p}}^T\mathbf{H}_{\mathbf{p};\mathbf{p}}\mathbf{m}_{\mathbf{p}} - 2\mathbf{m}_{\mathbf{p}}^T\mathbf{G} - \mathbf{G}^T\mathbf{H}_{\mathbf{p};\mathbf{p}}^{-1}\mathbf{G} \end{aligned} \tag{38}$$

where $\mathbf{G} = \mathbf{H}_{\mathbf{p};\mathbf{q}}\mathbf{m}_{\mathbf{q}}$.

When $\mathbf{p}$ reduces to a scalar, only one non-zero element in the new mean $\mathbf{m}_\star$ is kept, and no matrix inversion is needed for computing $l_\mathbf{q}$ since $\mathbf{H}_{\mathbf{p};\mathbf{p}}^{-1}$ becomes a scalar division. By pre-computing $\mathbf{m}^T\mathbf{H}\mathbf{m}$, we can update $l_\mathbf{q}$ efficiently through equation (38) for every $\mathbf{q}$.

We embed this sparsification algorithm in the Kalman filtering procedure. After obtaining $\mathbf{V}$ and $\mathbf{m}$ for the current data point, we sparsify the mean $\mathbf{m}$ and continue the Kalman filtering. In practice, we apply the sparsification algorithm every few steps of the Kalman filtering instead of every single step, in order to enhance the smoothness of $\mathbf{m}$ in time and reduce the computation.

C. Greedy Approximation

When the length of $\mathbf{p}$ is comparable to that of $\mathbf{q}$, then using the Schur Complement in equation (47) is no longer efficient. In this case, we may use a greedy approach to prune two elements of $\mathbf{m}$, corresponding to the same frequency, at each step.

If a signal has known frequency phases, we can reduce the model size by using only those sinusoid basis functions with an additional known phase parameter in the vector $\mathbf{c}_n$.

For such a reduced model, the greedy approximation prunes one element each time and becomes the so-called Optimal Brain Surgeon algorithm [17].

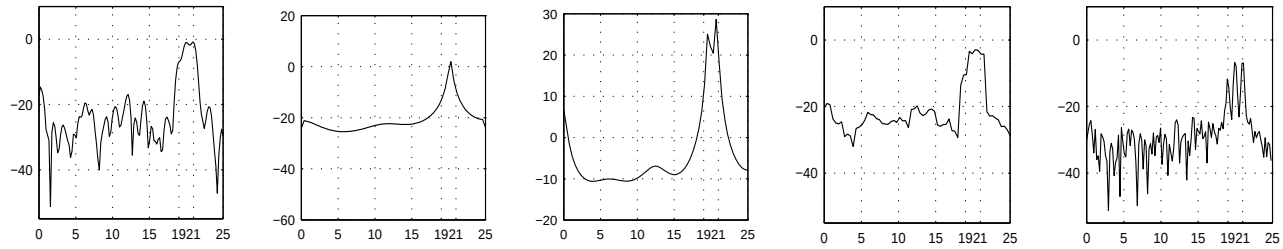
V. Experiments and Discussions

A. Comparison with Classical Spectrum Estimation Algorithms

First, we compare several classical methods with the new method on evenly sampled data to illustrate the new method's ability to resolve closely spaced signal frequency components. In Fig. 2 the signal is the sum of three sinusoids and noise. For Welch's algorithm, we use a window size of 100 data points, with 50 points of overlap. For Burg's algorithm, we chose a 6^{th} order AR model. For the MUSIC algorithm, we set the the signal subspace dimension to be 6. For the multitaper method, we used the standard Matlab implementation, and set the time-bandwidth product for the discrete prolate spheroidal sequences to be 2. These parameters were chosen by trial and error, trying to get the best performance out of each method. (For example, Matlab for multitaper recommends a time-bandwidth parameter of 4, but we found 2 to work better for this data set.) For the new method, we set the process noise variance to be a scaled identity matrix ($\mathbf{Z} = \mathbf{I}$) in equation (11), and used a stationary observation noise ($\sigma_n^2 = 0.1$ for all n). Also, we set $M = 127$ so that the length of the frequency basis vector $\mathbf{c}$ is 255, and utilized a noninformative prior on $p(\mathbf{s}_0)$ ($\mathbf{V}_0 = 1 \times 10^3$).

Note that the Y axis in Fig. 2 is on a log scale. The new method, because of its joint frequency estimation capability, successfully resolves the closely-spaced peaks without problems arising from their interference.

In addition, we carried out a sensitivity analysis. We scaled the noninformative prior and the noninformative model parameters 100 times larger and smaller to see if the changes affect the estimation results significantly. In these experiments, the results based on the new parameters were basically the same as those using the original parameters. This indicates that the probability model used is a good fit of the data and robust to the variation of the parameters.



(a) Welch's algorithm (b) Burg's algorithm (c) MUSIC algorithm (d) Multitaper (e) New

Fig. 2. Comparison with classical spectral estimation algorithms on evenly sampled stationary data. The signal is the sum of 19, 20, and 21 Hz real sinusoid waves with amplitudes 0.5, 1, and 1 respectively. The variance of the additive white noise is 0.1. The signal is evenly sampled 128 times at 50 Hz.

B. Estimation accuracy: fast decaying amplitude sinusoid

We synthesize an unevenly sampled signal that contains one 125 Hz sinusoid wave modulated with an exponentially fast decaying amplitude. In the case of unevenly sampled data, the Lomb-Scargle method is widely used in many applications.

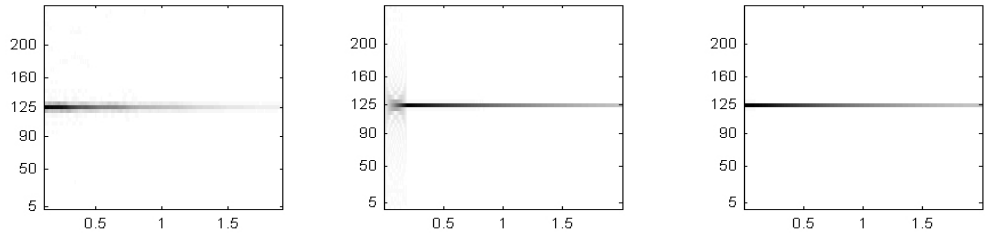
We compare the Lomb-Scargle periodogram, the new method, and the new method with smoothing. For the Lomb-Scargle periodogram, we use a short window size of 60 data points, with 59 points of overlap; less overlap yields visible "blocking" effects, and this value appeared to optimize its performance.

For the new method and its smoothing version, we set $z = 1000$ and $\sigma_n = 1$, and assigned a noninformative prior on $p(\mathbf{s}_0)$ ($\mathbf{V}_0 = 1 \times 10^{10}$).

The estimated spectra by these three methods are shown in Fig. 3. The true amplitude and the estimated amplitudes of 125 Hz components are plotted in Fig. 4. For the Lomb-Scargle periodogram, the mean square error is 0.0384. Except for the initialization (0.2 seconds) for the

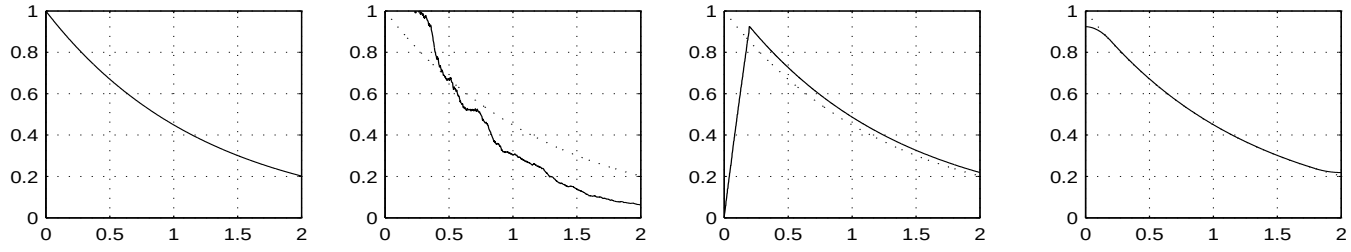
new method, the mean square error of the estimated amplitudes along the time axis is 0.0016; for the new method with smoothing, the mean square error drops to 0.0000080. Note that the Lomb-Scargle periodogram also has an initialization period due to its sliding window; its error, given above, omits this initialization period as well.

The new method, in both the original and smoothing versions, yields accurate estimates of the frequency and fast decaying amplitude, while the Lomb-Scargle periodogram fails to track the changing amplitude. Also, the Lomb-Scargle periodogram contains more sidelobe energy than do the spectrograms obtained from the new method. This is partly because the new method jointly estimates all the frequency bands and thus has the so-called "explaining-away' effect: if the signal is well explained by one or some of the frequency bands, the influence of other frequency bands will be reduced.



(a) Lomb-Scargle (b) New (c) New with Smoothing

Fig. 3. Estimated spectrograms for an unevenly sampled signal that contains one 125 Hz sinusoid modulated with an exponentially fast decaying amplitude. Note that the Y axes in this figure are log scale.



(a) True Amplitudes (b) Lomb-Scargle (c) New (d) New with Smoothing

Fig. 4. True and estimated amplitudes for the signal shown in Fig. 3.

C. Sparsification for Resolving Changing Frequencies

In this section, we demonstrate the sparsification technique described in section IV-B on a signal where the frequency jumps abruptly over time. The signal is unevenly sampled from a uniform distribution with a 30 dB SNR. The frequency of the signal jumps from 20 Hz to 40 Hz at time -0.833 second, and then jumps from 40 Hz to 60 Hz at time 0.833 second. The results, comparing the new method to Lomb-Scargle, are shown in figure 5.

As shown in the figure, by reducing the sliding window size from 200 data points to 100 data points, the Lomb-Scargle periodogram increases its frequency resolution, but at the same time, the stronger blocking effect results in a much larger detection delay of the signal's frequency switching. Here two consective windows overlap half of the window size. Actually, when dealing with fast changing frequencies, this problem is almost unavoidable for all sliding-window based methods, such as the short-time FFT, regardless whether the signals are evenly or unevenly sampled. For those methods, there is always a tradeoff between the capability of tracking fast changing frequencies and the frequency resolution.

For such abrupt changes in frequency, although the new method can still do a decent estimation, it cannot detect the frequency change very quickly as shown in Fig. 5 (c). Coupled with sparsification, the new method achieves the result shown in Fig. 5 (d), which not only has high frequency accuracy, but also quickly detects the frequency change. In this example, the sparsification algorithm is applied every 40 filtering steps, and the length of $\mathbf{q}$ is chosen a priori to be $h-2$.

D. Adjusting Model Parameters for Resolving Estimation Ambiguity

There is often more than one possible way to interpret a given signal. For example, given a signal $\mathbf{x}$

$$\mathbf{x}_n = \sin(2\pi f_1 t_n)\cos(2\pi f_2 t_n) \tag{39}$$

$$= \frac{1}{2}\big(\sin(2\pi t_n(f_1+f_2)) + \sin(2\pi t_n(f_1-f_2))\big)\ , \tag{40}$$

the traditional way to interpret this is via the second equation, so that its spectrogram has two sinsoids, one at f_1+f_2 and one at f_1-f_2. On the other hand, it might be desirable, in accord with the alternate interpretation of equation (39), to produce a "non-stationary spectrogram", that contains one single sinusoid at f_1, with a time-varying amplitude of value $\cos(2\pi f_2 t_n)$.

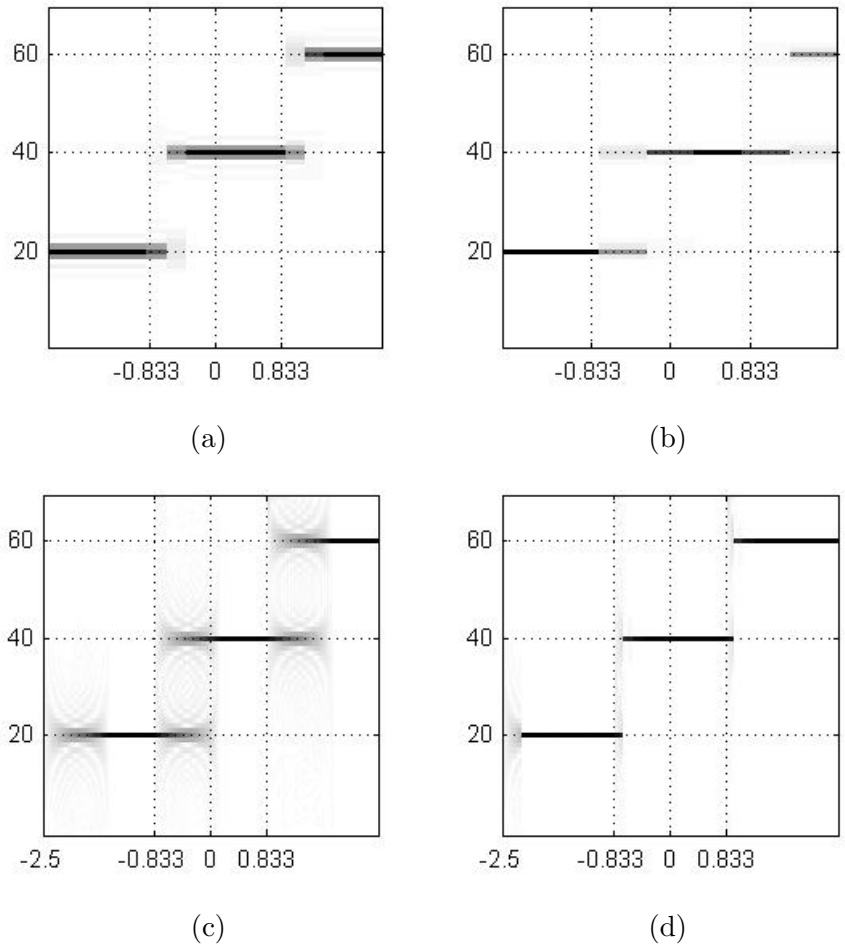


Fig. 5. Spectral analysis for an unevenly sampled signal where the frequency jumps from 20 Hz to 40 Hz at the sampling time -0.833 second, and then jumps from 40 Hz to 60 Hz at 0.833 second. (a) Lomb-Scargle periodogram with a window size of 100 points (b) Lomb-Scargle periodogram with a window size of 200 points (b) Spectrogram by the new method (d) Spectrogram by the new method coupled with sparsification

As discussed in the previous section III-E.2, the new method prefers the smoothness of the estimation along the time axis. Thus if we use a noninformative prior and model parameters that represent ignorance about the frequency property of the data (here $\mathbf{Z} = \mathbf{I}$ and $\sigma^2 = 10$), the estimator will tend to interpret the signal as what equation (40) says – there are two sinusoids in the signal $\mathbf{x}$. This is verified in the simulation shown in Fig. 6 (d), where the new method recovers the two sinusoids in the spectrogram that are very close to each other.

However, if we use so-called informative model parameters, which incorporate a priori knowledge about the data, we can bias the estimation to have only one frequency component (i.e., we set a very large process variance, 10^9, corresponding to both the sine and cosine components at

40Hz in $\mathbf{Z}$ and a small process noise variance, 10, for the rest frequency components). In this case, we obtain an estimation result showing that this signal is a single sinusoid at 40 Hz with a cosine modulated amplitude, as seen in Fig. 6 (e).

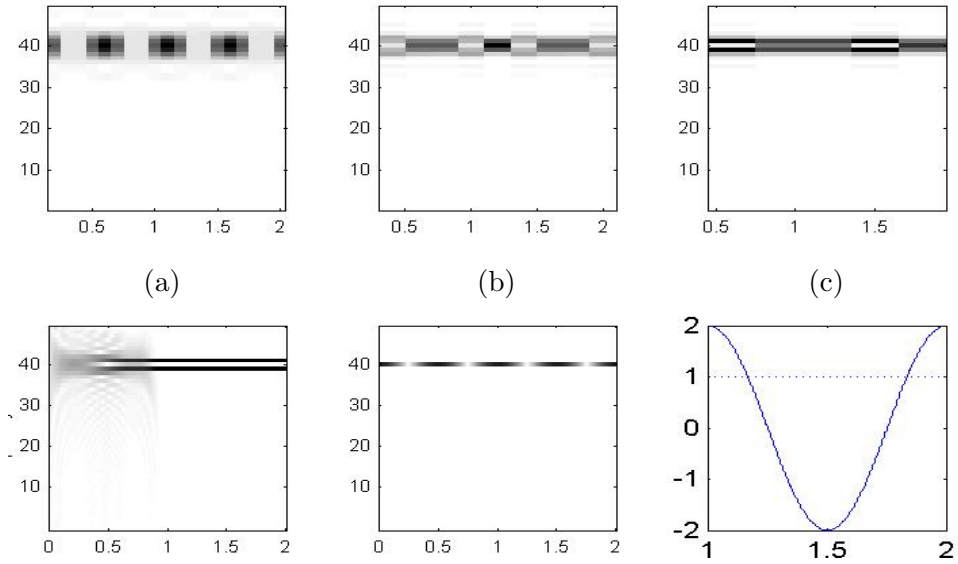


Fig. 6. Spectral analysis for an unevenly sampeled signal that contains 39 and 41 Hz sinusoids (a-c) Lomb-Scargle periodograms with sliding windows of 100, 200, and 300 data points respectively, which illustrate the interference of neighboring frequencies in Lomb-Scargle periodograms. (d) Spectrogram by the new method with noninformative model parameters: There is no interference between neighbor frequencies, again demonstrating the super-resolution property of this new method. (e) Spectrogram by the new method with informative model parameters. (f) Dotted curve: estimated amplitudes of 39 and 41 Hz by the new method with a noninformative model; solid curve: the estimated amplitude of the 40 Hz component by the new method with an informative model.

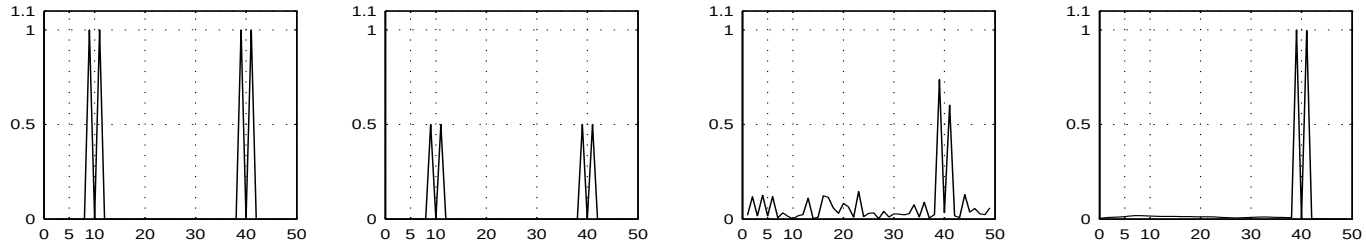
E. Sparsification or Informative Model Parameters?

We can either apply the sparsification algorithm or use informative model parameters to obtain a sparse spectrogram. We have performed a number of experiments to compare the two approaches on the data set used in the previous section, and the results are visibly indistinct from those we show above. In general, the sparsification method costs more in terms of computation, but requires less prior information than using informative model parameters.

F. Sampling Rate and Aliasing

Finally, we illustrate in Fig. 7 the use of the new method in preserving the property that uneven sampling diminishes aliasing. When this signal containing 39 and 41 Hz sinusoids is evenly sampled, it requires $f_s > 82$Hz in order to avoid aliasing. Here we show both even and uneven sampling at a rate of 100 times over 2 seconds. In the evenly sampled case, the aliased components show up as expected at 9 and 11 Hz. However, in the unevenly sampled case, there is no such aliasing for this signal. Thus, when aliasing is a concern, a method such as this new technique, that works well on unevenly sampled data, may provide a bandwidth advantage.

Comparing Fig. 7 (c) and (d), we see that the new method has an amplitude conservation property, i.e., the estimated amplitudes are equally distributed in the true and aliasing frequencies in (c). This illustrates the use of the amplitude conservation property to detect if aliasing occurs or not.



(a) Lomb-Scargle, even (c) New, even (b) Lomb-Scargle, uneven (d) New, uneven

Fig. 7. Lomb-Scargle periodogram and the spectra estimated by the new method for a signal $\mathbf{x} = \sin(2\pi 39\mathbf{t}) + \sin(2\pi 41\mathbf{t})$ sampled 100 times over 2 seconds, with samples either evenly or randomly (unevenly) spaced.

VI. Conclusion and Future Work

This paper has proposed a Bayesian method for spectrum estimation for unevenly-sampled noisy non-stationary data. By utilizing a non-stationary Kalman filter, the new method jointly estimates all the amplitudes and phases of frequency bands of interest instantaneously without the use of a fixed window or a fixed set of windows. Additionally, we showed how the Bayesian method is able to accommodate prior information about noise and signal structure. The new method appears to provide outstanding frequency resolution, even on small data sets. When

coupled with the sparsification algorithm, it can accurately estimate switching frequencies. When data are unevenly sampled, it can estimate frequency components beyond half of the average sampling rate.

One direction of future work is using non-Gaussian process and observation noises to estimate quick frequency changes. To this end, we are applying a deterministic Bayesian approximation technique to the probabilistic inference of spectrum. Some initial work has been done.

In addition, we are interested in employing different signal bases besides sinusoid and cosine functions. An example is wavelets, which, combined with Bayesian inference, will result in a fast stochastic time-frequency analysis algorithm. It is also possible to utilize multiple discrete or generalized prolate spheroidal sequences [11] as our signal basis, in order to reduce the estimation variance.

VII. Appendix

The pruning procedure in section (IV-A) is pretty expensive. At each iteration over the possible $\mathbf{q}$, it involves the computation of the inverse of the matrix $\mathbf{V}_{\mathbf{q};\mathbf{q}}$. In case we want to prune most elements of $\mathbf{m}$, we can use Schur complements to efficiently compute the inverse of partitioned matrices [18]. If we partition a matrix $\mathbf{P}$ into four sub-matrices $\begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix}$ then the Schur complement of $\mathbf{P}_{11}$ in $\mathbf{P}$ is

$$(\mathbf{P}|\mathbf{P}_{11}) = \mathbf{P}_{22} - \mathbf{P}_{21}\mathbf{P}_{11}^{-1}\mathbf{P}_{12} \tag{41}$$

Then if we similarly partition the inverse matrix $\mathbf{P}^{-1}$ into four sub-matrices $\begin{pmatrix} (\mathbf{P}^{-1})_{11} & (\mathbf{P}^{-1})_{12} \\ (\mathbf{P}^{-1})_{21} & (\mathbf{P}^{-1})_{22} \end{pmatrix}$ then

$$(\mathbf{P}^{-1})_{22} = (\mathbf{P}|\mathbf{P}_{11})^{-1}. \tag{42}$$

Then, from equations (35), (42) and (41) we have

$$(\mathbf{V}_{\mathbf{q};\mathbf{q}})^{-1} = (\mathbf{H}|\mathbf{H}_{\mathbf{p};\mathbf{p}}) \tag{43}$$

$$= \mathbf{H}_{\mathbf{q};\mathbf{q}} - \mathbf{H}_{\mathbf{q};\mathbf{p}}\mathbf{H}_{\mathbf{p};\mathbf{p}}^{-1}\mathbf{H}_{\mathbf{p};\mathbf{q}} \tag{44}$$

So we have

$$\delta\mathbf{m} = \mathbf{V}_{1:M;\mathbf{q}}(\mathbf{H}_{\mathbf{q};\mathbf{q}} - \mathbf{H}_{\mathbf{q};\mathbf{p}}\mathbf{H}_{\mathbf{p};\mathbf{p}}^{-1}\mathbf{H}_{\mathbf{p};\mathbf{q}})\mathbf{m}_{\mathbf{q}} \tag{45}$$

$$l_{\mathbf{q}} = \mathbf{m}_{\mathbf{q}}^T(\mathbf{H}_{\mathbf{q};\mathbf{q}} - \mathbf{H}_{\mathbf{q};\mathbf{p}}\mathbf{H}_{\mathbf{p};\mathbf{p}}^{-1}\mathbf{H}_{\mathbf{p};\mathbf{q}})\mathbf{m}_{\mathbf{q}} \tag{46}$$

$$= \mathbf{m}^T\mathbf{H}\mathbf{m} - \mathbf{m}_{\mathbf{p}}^T\mathbf{H}_{\mathbf{p};\mathbf{p}}\mathbf{m}_{\mathbf{p}} - 2\mathbf{m}_{\mathbf{p}}^T(\mathbf{H}_{\mathbf{p};\mathbf{q}}\mathbf{m}_{\mathbf{q}}) - (\mathbf{H}_{\mathbf{p};\mathbf{q}}\mathbf{m}_{\mathbf{q}})^T\mathbf{H}_{\mathbf{p};\mathbf{p}}^{-1}(\mathbf{H}_{\mathbf{p};\mathbf{q}}\mathbf{m}_{\mathbf{q}}) \tag{47}$$

References

[1] M. H. Hayes, *Statistical digital signal processing and modeling*, John Wiley and Sons, Inc, 1996.

[2] S.M. Kay and S.L. Marple Jr., "Spectrum analysis-A modern perspective," *Proceedings of IEEE*, vol. 69, no. 11, pp. 1380–1418, November 1981.

[3] D.J. Thomson, "An overview of multiple-window and quadratic-inverse spectrum estimation methods," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Australia, April 1994, vol. 6.

[4] A. Ouahabi, C. Depollier, L. Simon, and D. Kouamé, "Spectrum estimation from randomly sampled velocity data," *IEEE Transactions on instrumentation and measurement*, vol. 47, no. 4, pp. 1005–1012, August 1998.

[5] R. Banning, "Spectral analysis methods for Poisson sampled measurements," *IEEE Transactions on instrumentation and measurement*, vol. 46, no. 4, pp. 882–887, August 1997.

[6] E.R. Dowski, C.A. Whitmore, and S.K. Avery, "Estimation of randomly sampled sinusoids in additive noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 12, pp. 1906–1908, December 1988.

[7] N. R. Lomb, "Least-squares frequency analysis of unevenly spaced data," *Astrophysical and Space Science*, pp. 447–462, 1976.

[8] J. D. Scargle, "Studies in astronomical time series analysis ii. statistical aspects of spectral analysis of unevenly sampled data," *Astrophysical Journal*, pp. 835–853, 1982.

[9] G. L. Bretthorst, "Bayesian spectrum analysis and parameter estimation," in *Lecture Notes in Statistics*, 48. Springer-Verlag, 1988.

[10] G.B. Moody, "Spectral analysis of heart rate without resampling," *Computers in Cardiology*, vol. 20, pp. 715–718, 1993.

[11] T.P. Bronez, "Spectral estimation of irregularly sampled multidimensional processes by generalized prolate spheroidal sequences.," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 12, pp. 1862–1873, December 1988.

[12] M. Bayram and R.G. Baraniuk, "Multiple window time-varying spectrum estimation," in *Nonlinear and Nonstationary signal processing*, pp. 292–316. Cambridge University press, 2000.

[13] D.J. Thomson, "Multitaper analysis of nonstationary and nonlinear time series data," in *Nonlinear and Nonstationary signal processing*, pp. 317–394. Cambridge University press, 2000.

[14] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, N.J., 1979.

[15] T. P. Minka, "From hidden markov models to linear dynamical systems," Tech. Rep. 531, Vision and Modeling Group of Media Lab, MIT, 1998.

[16] G. L. Bretthorst, "Nonuniform sampling: Bandwidth and aliasing," in *Maximum Entropy and Bayesian Methods in Science and Engineering*, 2000, pp. 1–28.

[17] B. Hassibi and D.G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," in *NIPS*, 1993, vol. 5, pp. 164–171.

[18] T. P. Minka, "Old and new matrix algebra useful for statistics," http://www.stat.cmu.edu/~minka/papers/matrix.html, 1997.