Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-28 (2015) 478–483

A BAYESIAN NON-PARAMETRIC APPROACH TO FREQUENCY ESTIMATION

Martina Favaro * **Giorgio Picci** **

* *Department of Economics and Statistics, Guangzhou university, Guangzhou, China (e-mail: martinafavaro@yahoo.it)*
** *Department of Information Engineering, University of Padova, Padova, Italy (e-mail: picci@dei.unipd.it)*

Abstract: Although frequency estimation is a nonlinear parametric problem, it can be cast in a non-parametric framework. By assigning a natural a priori probability to the unknown frequency, the covariance of the prior signal model is found to admit an eigenfunction expansion alike the famous *prolate spheroidal wave functions*, introduced by D. Slepian in the 1960's. This leads to a technique for estimating the hyperparameters of the prior distribution which is essentially linear. This is in contrast to standard parametric estimation methods which are based on iterative optimization algorithms of local nature. The approach seems to be new and quite promising.

1. INTRODUCTION

Frequency estimation is an old nonlinear problem of paramount importance in various branches of engineering which has generated an enormous literature, see e.g. Quinn and Hannan [2001]. Besides spectral analysis, which tends to produce nonconsistent estimates, in the literature the problem has been mostly approached by optimization techniques which by their very nature are generally local and non robust. In the last two decades, mostly due to the availability of ultrafast and very powerful computers, there has been a dramatic growth in research and applications of nonparametric Bayesian identification and estimation methods. These tools have made non-parametric statistics based on the Bayesian paradigm almost routine. In this paper we shall attempt to formulate and indicate a possible solution of the frequency estimation problem by using a nonparametric Bayesian approach. There seems to be very little prior research in this area and we may adventure to say that our approach seems to be original.

The advantage of the Bayesian point of view is that it provides a rich probabilistic setting in which parameters may appear at a higher level as *hyperparameters* in the a priori distribution. As we shall see this is also the case for our problem setting. In this way frequency estimation, which is naturally an exquisitely parametric problem, can be recast in a Bayesian framework as a problem of hyperparameter estimation for a natural class of prior distributions.
We use a very natural prior density on the unknown frequency which describes the samples of the observed signal as those of a particular class of stationary processes, called *Bandlimited white noise processes*. Stationary a posteriori descriptions of harmonic signals are also considered in Lázaro-Gredilla et al. [2010] but our work goes far beyond this reference. We exploit the fact that the covariance of these processes turns out to be of the *modulated* Sinc *type*. It came as a pleasant surprise for us to discover that the eigenfunction expansion of Sinc-type kernels has been well studied in the 60's and 70's by David Slepian and coworkers in a famous series of papers the first of which Slepian and Pollak [1961] has about 1700 citations. These eigenfunction expansions have the remarkable property of involving only a finite and known number of terms, as most eigenvalues decay very fast to zero for index greater than a known a priori computable number (the so-called *Slepian frequency*).

The paper is organized as follows:
In Sect. 2 and 3 we formulate the frequency estimation problem in a Bayesian framework. We derive the covariance Kernel of the oscillatory signal $\mathbf{x}$ and of the observation process.
In Section 3 we discuss the problem of estimating the hyperparameter of the prior distribution. The estimate can be computed by a prediction error method. The general structure of the linear Bayesian predictor is discussed in Sect. 4.
In Sect. 5 we illustrate by experiments the spectral expansion of the covariance kernel (Sinc and Modulated Sinc kernels) and pinpoint the special properties of the eigenvalues eigenfunctions of these kernels, discovered by Slepian. The Bayes predictor can be expressed in terms of these spectral data.
Section 6 uses the special properties of the eigenstructure to reveal the dependency of the predictor on the center frequency and the prediction error estimate.
The appendix contains a restatement of some facts which relate to the eigenstructure of the covariance Sinc and modulated Sinc Kernel.
We deal with signals with just one hidden sinusoidal component; signals with multiple harmonic components of unknown frequency can be treated in a similar way by assigning non-overlapping rectangular (uniform) prior distribution to the unknown frequencies $\boldsymbol{\omega}_k$; $k = 1, 2, \ldots$ so that the a posteriori process model results in a sum of

2405-8963
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2015.12.174

uncorrelated components each having spectrum supported on non-overlapping frequency intervals. In this way the overall covariance kernel becomes the sum of the individual covariances of the harmonic components and Multiple Kernel methods as Hoffmann et al. [2008], Bach et al. [2004] can be applied; although some of the details still need to worked out. This issue is a subject of current research and will be discussed in a future publication.

2. PROBLEM STATEMENT

Consider the following simple model of a (wide-sense) stationary random oscillation in additive noise

$$\mathbf{y}(t)=\mathbf{a}\cos(2\pi\mathbf{f}t)+\mathbf{b}\sin(2\pi\mathbf{f}t)+\mathbf{w}(t):=\mathbf{x}(t)+\mathbf{w}(t),\quad t\in\mathbb{Z}, \tag{1}$$

where, by stationarity, $\mathbf{a}$, $\mathbf{b}$ must be uncorrelated zero-mean random variables of the same variance σ^2. We shall model the normalized angular frequency $\mathbf{f}$ as a random variable taking values in the interval $[-1/2,\ 1/2]$, independent of $\mathbf{a}$, $\mathbf{b}$. The process $\mathbf{w}(t)$ is a stationary white noise of variance $\sigma^2_{\mathbf{w}}$ independent of everything else.

Our goal is to estimate the unknown random parameters of the model and their variances. To this end we shall propose a nonparametric approach which seems to be original. Note that the model is linear in $\mathbf{a}$, $\mathbf{b}$ and hence estimation of the amplitudes and their variance when the frequency is given is just a standard linear estimation problem. For this reason, in this paper we shall just concentrate on the estimation of frequency.

The covariance function of the process $\mathbf{y}$ has the form

$$\Sigma(t,s)=\mathbb{E}\,\mathbf{y}(t)\mathbf{y}(s)=K(t,s)+\sigma^2_{\mathbf{w}}\delta(t,s)$$

where δ is the Kronecker symbol and K is the a priori covariance of the signal $\mathbf{x}(t)$, that is

$$\begin{aligned}K(t,s)=\mathbb{E}\,\{&\mathbf{a}^2\cos(2\pi\mathbf{f}t)\cos(2\pi\mathbf{f}s)+\mathbf{ab}\cos(2\pi\mathbf{f}t)\sin(2\pi\mathbf{f}s)\\&+\mathbf{ab}\sin(2\pi\mathbf{f}t)\cos(2\pi\mathbf{f}s)+\mathbf{b}^2\sin(2\pi\mathbf{f}t)\sin(2\pi\mathbf{f}s)\}\end{aligned}$$

We shall put a prior distribution on the normalized frequency $\mathbf{f}$ assuming a uniform distribution on the frequency band $[f_0-W,\ f_0+W]$ where $0\le W\le 1/2$.

Computing first the conditional expectation with respect to $\mathbf{f}=f$ and then integrating with respect to the prior on $\mathbf{f}$ one gets

$$\begin{aligned}K(t,s)&=\mathbb{E}\,\left[\sigma^2\cos 2\pi\mathbf{f}(t-s)\right]=\\&2\sigma^2\int_{f_0-W}^{f_0+W}\cos 2\pi f(t-s)\frac{df}{4W}\\&=\sigma^2\cos 2\pi f_0\tau\frac{\sin 2\pi W\tau}{2\pi W\tau}\end{aligned} \tag{2}$$

where $\tau:=(t-s)$. The signal $\mathbf{x}(t)$ is therefore stationary. For $f_0=0$, the function K is the well-known *Sinc Kernel*. For $f_0\neq 0$ it will be called a *Modulated-sinc Kernel* Khare [2006]. Since the Sinc kernel is the inverse Fourier transform of a rectangular function of frequency

$$\frac{\sin 2\pi W\tau}{2\pi W\tau}=\int_{-W}^{W}\frac{1}{2W}e^{j2\pi f\tau}df$$

it follows that for $f_0=0$ the process $\mathbf{x}(t)$ has a uniform spectral density supported on the normalized frequency interval $[-W,\ W]$

$$\boldsymbol{\phi}_{\mathbf{x}}(f)=\sigma^2\frac{1}{2W}\chi_{[-W,W]}(f)$$

where χ_Δ denotes the characteristic function of the set Δ. In other words, the process $\mathbf{x}(t)$ is just a *bandlimited white noise* signal within the frequency band $[-W,\ W]$.

For $W<1/2$ the process is purely deterministic with an absolutely continuous spectral distribution (the logarithm of the density being obviously not integrable) while for $W=1/2$ the a priori model of the process is just a stationary white noise of variance σ^2.

When $f_0\neq 0$, assuming $|f_0-W|<1/2$, the spectral density is supported on the two intervals $[f_0-W,\ f_0+W]$ and $[-f_0-W,\ -f_0+W]$ and has the expression

$$\boldsymbol{\phi}_{\mathbf{x}}(f)=\frac{1}{2}\sigma^2\frac{1}{2W}\chi_{[f_0-W,\ f_0+W]}+\frac{1}{2}\,\sigma^2\frac{1}{2W}\chi_{[-f_0-W,\ -f_0+W]}$$

The signal $\mathbf{x}$ can therefore be described as a deterministic carrier of frequency $2\pi f_0$, amplitude-modulated by the bandlimited white noise process described before.

To simplify notations we shall denote the center frequency of the prior by the symbol $\theta:=2\pi f_0$ with $f_0\le 0.5$. In the literature both θ and W are called *hyperparameters* of the prior distribution. In the Bayesian nonparametric setting estimation is posed as estimation of the hyperparameters; see Chiuso [2015] for a critical survey and an extensive bibliography.

3. A BAYESIAN PARADIGM FOR FREQUENCY ESTIMATION

Estimation of the hyperparameter is usually done by maximum likelihood based on a (hopefully) long data set. Let $y_N(t):=[y(t)\ y(t-1)\ \dots y(t-N+1)]^\top$ be a string of N successive observed data and let

$$Y_N=y_N(t)y_N(t)^\top$$

The Gaussian log-likelihood function based on N data can be written Hannan and Deistler [1988]

$$l(\theta)=-\frac{N}{2}\log 2\pi-\frac{N}{2}\log\det\boldsymbol{\Sigma}(\theta,W)-\frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}(\theta,W)^{-1}Y_N\} \tag{3}$$

where $\boldsymbol{\Sigma}(\theta,W)$ is the covariance matrix of $\mathbf{y}_N$. One should minimize this function with respect to θ,W. For N moderately large, say of the order of hundred data points, the numerical maximization of this function turns however out to be an impossible task. This is so also if we approximate the problem by just attempting to minimize the trace $\mathrm{tr}\{\boldsymbol{\Sigma}(\theta,W)^{-1}Y_N\}$. The reason being that the inversion of large Toeplitz matrices is an extremely time-consuming and ill-conditioned problem. One may try several variants of so-called "fast" Toeplitz inversion algorithms e.g. Golub and van Loan [1983] but going beyond $N=100$ turns out to be very hard anyway. See however Stoica et al. [2011] for a brute-force solution of a similar problem.

To circumvent the Toeplitz inversion problem we shall resort to a Prediction-Error (PE) minimization approach. It is well-known that under very mild assumptions on the true model describing the data, the two procedures are asymptotically equivalent; see e.g. Ljung [1999]. It is shown for example in Ljung's book that a PE estimator will generally be consistent and asymptotically Gaussian. The estimation by a prediction-error method, although being not exactly equivalent to maximum likelihood for small data lenght, has the advantage of not requiring the inversion of large Toeplitz matrices.

Assume that we have a (long enough) data series of $2N$ samples $\{y(-N+1),\dots,y(0),\dots,y(T)\}$ and for each $t=0,\dots,T-1$ construct a one step ahead linear predictor,

say $\hat{\mathbf{y}}(t+1 \mid t)$ of $\mathbf{y}(t+1)$ based on the past history $\mathbf{y}_N(t) := \{\mathbf{y}(s)\,;\, t \geq s \geq t-N+1\}$ of the most recent N observations at time t. Such a predictor will be based on a linear Bayesian estimation approach. The details will be given in Section 4. Note that since in the model (1) the additive noise is uncorrelated with the signal $\mathbf{x}$ one actually has the identity

$$\hat{\mathbf{y}}(t+1 \mid t) = \hat{\mathbf{x}}(t+1 \mid t)\,, \tag{4}$$

so that the predictor $\hat{\mathbf{y}}(t+1 \mid t)$ will coincides with the predictor of the process $\mathbf{x}$ having covariance kernel $K(t-s)$, based on observations corrupted by the additive white noise $\mathbf{w}$. As such, this predictor will be a linear function of the past data $\mathbf{y}_N(t)$, denoted

$$\hat{\mathbf{y}}(t+1 \mid t) = a(\theta)^\top \mathbf{y}_N(t) \tag{5}$$

where the vector a depends and on the hyperparameter θ and W (not shown) but is independent of t by stationarity. Now it is quite obvious and can be formally justified that for $W \to 0$ the process $\mathbf{x}$ becomes a sinusoidal signal of deterministic frequency equal to the a priori centerfrequency f_0. As W increases, there trivially will be more uncertainty associated to $\mathbf{f}$. Intuitively W could be interpreted as a *confidence interval* about the true signal center frequency. In the approach proposed in this paper the minimization of the prediction error will be done only with respect to θ keeping W fixed. We shall just minimize the sample prediction error variance with respect to θ; i.e.

$$\min_{\theta}\ \frac{1}{T}\sum_{t=1}^{T}(y(t+1) - \hat{\mathbf{y}}(t+1 \mid t))^2 \tag{6}$$

which will yield a PE estimate $\hat{\theta}_T$ of course depending on W. The parameter W will then be adjusted iteratively in function of the statistical dispersion of the center frequency estimate.

In general, even if the minimization (6) does not involve large Toeplitz matrix inversion, it needs to be done numerically. Later we shall argue that our problem has a special structure which will make this task quite straightforward. Note that, by consistency of PE estimation, for $N \to \infty$ the residual average squared prediction error minimized in (6), approximates the one step prediction error variance. Therefore a consistent estimate of the noise variance can be obtained as the average residual squared prediction error

$$\hat{\sigma}^2_{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}(y(t+1) - a(\hat{\theta}_T)^\top y_N(t))^2\,. \tag{7}$$

which will converge when $T \to \infty$ to $\mathbb{E}\,(\mathbf{y}(t+1) - \hat{\mathbf{y}}(t+1 \mid t))^2$, the innovation variance of $\mathbf{y}(t)$. Since (for $W < 1/2$) $\mathbf{x}(t)$ is a purely deterministic process, the innovation of the process $\mathbf{y}(t)$ is just the additive white noise $\mathbf{w}(t)$ in the model (1) and hence, for T large enough we have $\hat{\sigma}^2_{\mathbf{w}} \simeq \sigma^2_{\mathbf{w}}$. Now stack each scalar linear predictor next to each other for $t = 0, 1, \ldots, N-1$ to form a vector

$$\hat{\mathbf{y}}_T := [\hat{\mathbf{y}}(1 \mid 0)\ \hat{\mathbf{y}}(2 \mid 1)\ \ldots\ \hat{\mathbf{y}}(T \mid T-1)]^\top \tag{8}$$

and let

$$Y_N := \begin{bmatrix} y_N(0)^\top \\ y_N(1)^\top \\ \cdots \\ y_N(T-1)^\top \end{bmatrix}. \quad \mathbf{y}_T := \begin{bmatrix} y(1) \\ y(2) \\ \cdots \\ y(T) \end{bmatrix}$$

The minimization (6) will be done in two steps. First step will be a least squares minimization of the Euclidean norm

$$\min_{a \in \mathbb{R}^N}\ \|\mathbf{y}_T - \hat{\mathbf{y}}_T\|^2 \equiv \min_{a \in \mathbb{R}^N}\ \|\mathbf{y}_T - Y_N a\|^2\,. \tag{9}$$

Next (see Sect. 6) we shall compute the estimate $\hat{\theta}_T$ using the analytic expression of $a(\theta)$ which will be derived in the following two sections of the paper.

4. STRUCTURE OF THE LINEAR PREDICTOR

The normalized $N \times N$ covariance matrix [1] of an observed string $\mathbf{y}_N(t)$ of N data, has the form

$$\boldsymbol{\Sigma}_{\mathbf{y}}(i,k) := \frac{1}{\sigma^2}\,[\,\mathbb{E}\,\mathbf{y}(i)\mathbf{y}(k)\,]_{i,k=1,2,\ldots,N} \tag{10}$$

$$= \cos\theta(i-k)\frac{\sin[2\pi W(i-k)]}{2\pi W(i-k)} + \rho^2\delta(i-k) \tag{11}$$

where $\rho^2 := \sigma^2_{\mathbf{w}}/\sigma^2$ is the inverse of the signal to noise ratio, which we shall assume is known a priori. This "noise perturbed" Modulated-Sinc Kernel matrix depends on θ and on W. Likewise let $\mathbf{K}$ be the $N \times N$ normalized covariance of the vector $\mathbf{x}_N(t)$, the N-vector of previous signal samples $\mathbf{x}(s)$, $t \leq s \leq t-N+1$. We have

$$\mathbf{K}_{i,k} = \cos\theta(i-k)\frac{\sin[2\pi W(i-k)]}{2\pi W(i-k)}, \qquad i,k = 0,1,\ldots,N-1 \tag{12}$$

The linear predictor (4) can be computed by the classical formula

$$\hat{\mathbf{y}}(t+1 \mid t) = \frac{1}{\sigma^2}\mathrm{Cov}\,\{\mathbf{y}(t+1), \mathbf{y}_N(t)\}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y}_N(t)$$

$$= \frac{1}{\sigma^2}\mathrm{Cov}\,\{\mathbf{x}(t+1), \mathbf{x}_N(t)\}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y}_N(t) \tag{13}$$

Here $\mathbf{x}_N(t)$ is substituted in place of $\mathbf{y}_N(t)$ due to the uncorrelated additve noise structure of the model. Note that all covariances do not depend on t by stationarity. The cross-covariance string $\mathrm{Cov}\,\{\mathbf{x}(t+1), \mathbf{x}_N(t)\}$ is just the first row of the $N \times N$ matrix $\mathrm{Cov}\,\{\mathbf{x}_N(t+1), \mathbf{x}_N(t)\}$, which, after normalization has entries

$$[\mathbf{K}_1]_{i,k} = \cos\theta(i-k+1)\frac{\sin[2\pi W(i-k+1)]}{2\pi W(i-k+1)},$$

and is also made of the same modulated Sinc Kernel expressions as in formula (12) but shifted by one time unit. Letting $\mathbf{k}_1$ be the first row of $\mathbf{K}_1$, the predictor at time t can be represented by the explicit Bayes formula

$$\hat{\mathbf{y}}(t+1 \mid t) = \mathbf{k}_1\,\boldsymbol{\Sigma}_N^{-1}\,\mathbf{y}_N(t) = \mathbf{k}_1\,\left[\,\mathbf{K} + \rho^2 I_N\,\right]^{-1}\,\mathbf{y}_N(t) \tag{14}$$

(here we have simplified σ^2 in both members).

Formula (14) requires the inversion of a large covariance matrix. To this end it is customary to introduce its spectral expansion in term of eigenvalues/eigenfunction and express the estimate directly in terms of the eigenfunctions Wahba [1990].

Consider then the eigendecomposition of the $N \times N$ symmetric positive semidefinite matrix $\mathbf{K}$,

$$\mathbf{K} = \sum_{k=0}^{N-1}\mu_k\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^\top, \quad \mathbf{K}\boldsymbol{\varphi}_k = \mu_k\boldsymbol{\varphi}_k \quad k = 0,1,2,\ldots,N-1 \tag{15}$$

[1] Covariance matrices of random variables will also be denoted by bold symbols.

where the eigenvalues are all positive and ordered in decreasing magnitude. In vector notation the expansion (15) can be written

$$\mathbf{K} = \Phi\,\Lambda\,\Phi^\top \qquad \text{where} \quad \Lambda := \text{diag}\{\mu_0, \dots, \mu_{N-1}\} \quad (16)$$

where the matrix $\Phi := [\boldsymbol{\varphi}_0 \ \dots \ \boldsymbol{\varphi}_{N-1}]$ is unitary, that is $\Phi\Phi^\top = \Phi^\top\Phi = I_N$. Likewise we shall have

$$\mathbf{K}_1 = \Phi_1\,\Lambda\,\Phi^\top \quad (17)$$

where Φ_1 is made of the eigenvectors of the matrix (15) bordered to dimension $(N+1)\times(N+1)$ with the first row chopped off so as to make its columns of length N. Assuming N large enough, the first N eigenvalues of the bordered matrix are with very good accuracy the same of the original $N \times N$ kernel $\mathbf{K}$. See later for a discussion of this point.

Denote by $\mathbf{c}$ the first row of Φ_1. Then, the first row of (17) is $\mathbf{k}_1 = \mathbf{c}\,\Lambda\,\Phi^\top$ and

$$\begin{aligned}\hat{\mathbf{y}}(t+1 \mid t) &= \mathbf{c}\,\Lambda\,\Phi^\top \left[\mathbf{K} + \rho^2 I_N\right]^{-1}\mathbf{y}_N(t)\\ &= \mathbf{c}\,\left[\Phi^\top\mathbf{K}\Phi + \rho^2 I_N\right]^{-1}\Phi^\top\mathbf{y}_N(t) \quad (18)\\ &= \mathbf{c}\,\text{diag}\,\{\frac{\mu_0}{\mu_0+\rho^2}, \dots, \frac{\mu_{N-1}}{\mu_{N-1}+\rho^2}\}\Phi^\top\mathbf{y}_N(t)\end{aligned}$$

so that each predictor is a linear combination of the generalized Fourier coefficients

$$\bar{\mathbf{y}}_N(t) := \left[\boldsymbol{\varphi}_0^\top\mathbf{y}_N(t) \ \dots \ \boldsymbol{\varphi}_{N-1}^\top\mathbf{y}_N(t)\right]^\top \in \mathbb{R}^N$$

The last expression in (18) will be central in the following. We need to understand how the prediction error depends on the hyperparameters. Although in general it may look like the eigenvalues should depend on θ, W and on N, the situation turns actually out to be much simpler due to the very special structure of the kernels. We shall first analyze the case of center frequency equal to zero.

It has been shown by Slepian [1978] that for $\theta = 0$, in the eigendecomposition of the *Sinc kernel*,

$$\mathbf{S}_{i,k} = \frac{\sin[2\pi W(i-k)]}{2\pi W(i-k)} = \frac{\sin 2\pi W(i-k)}{2\pi W(i-k)},\ i,k = 0,1,\dots,N \quad (19)$$

there is a magic integer number n very closely approximated by $2NW$, sometimes called the *Slepian frequency*, such that the eigenvalues λ_k of $\mathbf{S}$ are, with very good approximation, all equal to one when $k < n$, while the others are very small and can be neglected. This fact will be also discussed in some detail in the appendix A.The fact that there are only a relatively small number of non-negligible eigenvalues is of course of great practical importance. That this should also be the case for the eigenvalues μ_k of the *modulated Sinc kernel* $\mathbf{K}$ has been conjectured in Khare [2006] but as far as we know, there is no rigorous proof in the literature. This is of course of great importance for our application. In the paper we shall keep this fact for granted. Extensive simulations indicate that this is indeed the case. See Section 5 below.

Assuming that $\mu_0 = \mu_1 = \dots = \mu_n \simeq 1$ and $\mu_{n+k} \simeq 0$ for $k > 0$, the last expression in (18) can be simplified to involve only the first n eigenfunctions. Letting

$$\mathbf{K} = \Phi_n\Lambda_n\Phi_n^\top, \quad \Lambda_n = \text{diag}\{\mu_0, \dots, \mu_{n-1}\} \quad (20)$$

$$\mathbf{z}(t) := \Phi_n^\top\mathbf{y}_N(t) \quad \in \mathbb{R}^n \quad (21)$$

where $\Phi_n \in \mathbb{R}^{N\times n}$ is the matrix of the first n eigenvectors and $\mathbf{z}(t)$ is an n-dimensional random vector of covariance Λ. The expression of the predictor simplifies to

$$\hat{\mathbf{y}}(t+1 \mid t) = \mathbf{c}\,\text{diag}\,\{\frac{1}{1+\rho^2}, \dots, \frac{1}{1+\rho^2}\}\mathbf{z}(t) = \frac{1}{1+\rho^2}\mathbf{c}\,\mathbf{z}(t) \quad (22)$$

which depends on $\mathbf{y}_N(t)$ only through its components with respect to the first n eigenfunctions.

In the next section we shall provide some experimental evidence that this behavior of the eigenvalues is actually occurring also for the modulated Sinc kernel $\mathbf{K}$, as conjectured in Khare [2006].

5. EXPERIMENTAL STUDY OF THE EIGENVALUES OF THE SINC AND MODULATED SINC KERNEL

As for the Sinc kernel, we have generated a matrix $\mathbf{S}$ of dimension NXN with $N = 1000$ for values of the bandwidth $W := 0.02$. Here $n = 2NW$ is then equal 40. Figures 1c shows the behavior of the eigenvalues λ_k of the Sinc kernel for these two values of N, W. We can clearly see that for $n < 40$ the eigenvalues are all equal to the same constant while for $n > 40$ the λ_k very quickly decrease to zero.

As for the modulated Sinc kernel, we shall only report here the eigenvalues of a matrix $\mathbf{K}$ of dimension $N = 1000$ and $f_0 = 0.3$, for $W = 0.02$. In Fig. 2 we see that the eigenvalues have exactly the same behavior as those of the Sinc kernel. Only the value of n such that for $k > n$, $\mu_k \simeq 0$ is now equal to $2(2NW)$; i.e. twice the value of n for the Sinc kernel. Moreover the amplitudes of the eigenvalues μ_k for $k < n$ are half of those of the Sinc kernel, for equal values of W. This follows the from the symmetry of the spectrum and matches exactly the findings of Khare [2006]. We should note that in Slepian [1978] the eigenvalues of a slightly different Sinc kernel equal to $2W\mathbf{S}$ are analyzed. Their behavior is the same of ours except that the normalization makes the λ_k all practically equal to one for $k < n$. In order to get the same normalization we just need to substitute λ_k with $2W\lambda_k$. This fact is evident in the simulations and will be explained in the appendix A.

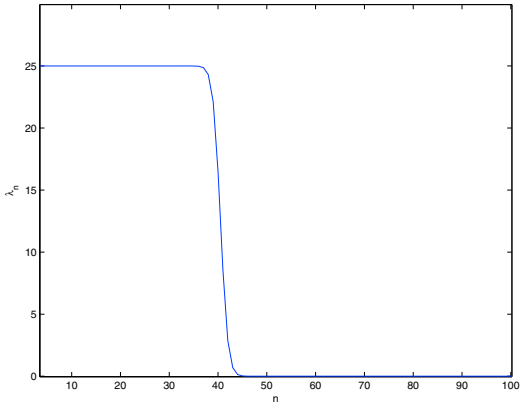


Fig. 1. *Sinc Kernel Eigenvalues,* $n = 40$

As regards the modulated Sinc kernel, in order to get the largest eigenvalues equal to one, a normalization should be made by substituting λ_k with $4W\lambda_k$. This also agrees with the findings of Khare [2006].

6. COMPUTING THE ESTIMATE

For a fixed N and W we shall here assume that the $N \times N$ covariance kernel $\mathbf{K}$ has exactly rank $n \simeq 4NW$.

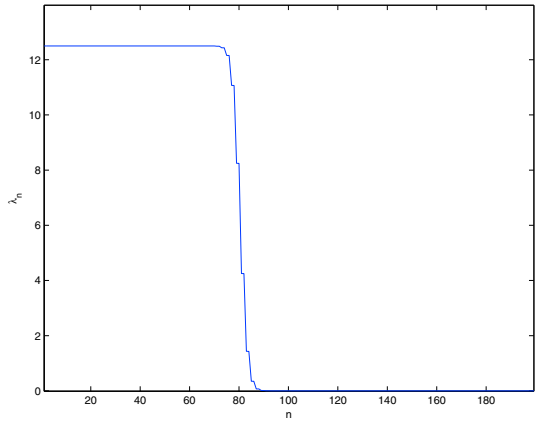


Fig. 2. *Modulated Sinc Kernel Eigenvalues,* $n = 80$

Proposition 1. There are a $n \times n$ matrix A and an n-dimensional row vector c such that

$$\mathbf{z}(t+1) = A\mathbf{z}(t) \tag{23}$$
$$\mathbf{y}(t) = c\mathbf{z}(t) + \mathbf{w}(t) \tag{24}$$

where $\mathbf{z}(t)$ is the n-dimensional vector defined in (21).

Proof : In fact by the analysis of the previous section, all covariances obtained by bordering $\mathbf{K}$ will still have rank n. Now a rank-deficient covariance matrix (of rank n) must necessarily be the covariance of a purely deterministic process [2] which can be represented by a deterministic linear recursion of order n [Lindquist and Picci, 2015, p. 138, 276] or equivalently, by a n-dimensional state space model. In geometric terminology this means that the Hilbert space $\mathbf{H}$ of random variables spanned by the random components of the vectors $\mathbf{x}_N(t)$ will be n-dimensional and the same as that spanned by $\mathbf{x}_M(t)$ for all $M \geq N$. In fact this will be also equal to the Hilbert spaces spanned by $\mathbf{x}_M(t+k)$ for any $k \geq 0$. Any n-dimensional basis vector $\boldsymbol{\xi} = [\,\boldsymbol{\xi}_1\, \boldsymbol{\xi}_2\, \dots \boldsymbol{\xi}_n\,]^\top$ in $\mathbf{H}$ yields such a state space representation for the process $\mathbf{x}(t)$, say

$$\boldsymbol{\xi}(t+1) = A\boldsymbol{\xi}(t)$$
$$\mathbf{x}(t) = c\boldsymbol{\xi}(t)$$

so that $\mathbf{y}(t)$ can be represented as $\mathbf{y}(t) = c\boldsymbol{\xi}(t) + \mathbf{w}(t)$. It is easy to check that for such a model the one step ahead predictor of $\mathbf{y}(t)$ is a linear function of the state at time t, given by

$$\hat{\mathbf{y}}(t+1 \mid t) = cA\boldsymbol{\xi}(t)$$

but the predictor (22) has a similar form, involving the n-dimensional random vector $\mathbf{z}(t)$ defined in (21) in place of $\boldsymbol{\xi}$. Since the predictor must be a linear function of the state, it follows that $\mathbf{z}(t)$ can also serve as a state process and hence $\mathbf{y}(t)$ can be described by the model (23), (24). □

The matrices c and A can be computed from the matrix Φ_n made of the first n columns of Φ, which are clearly also eigenvectors of the output covariance $\boldsymbol{\Sigma}$, the eigenvalues being now $\mu_k + \rho^2$, by a standard "shift-invariance" procedure of subspace identification. The eigenvector matrix Φ_n and its one block shifted counterpart must have the structure

$$\Phi_n = \begin{bmatrix} c \\ cA \\ \dots \\ cA^{\nu-1} \end{bmatrix}, \quad \downarrow \Phi_n := \begin{bmatrix} cA \\ cA^2 \\ \dots \\ cA^{\nu} \end{bmatrix}$$

from which one can extract c and the dynamic matrix A by solving (in a least squares sense) the equation $\downarrow \Phi_n = \Phi_n A$. Since, as we have seen, the eigenvalues do not depend on θ, the whole dependence on the centerfrequency must be in cA. Hence equating $a = cA(\theta)$, where $a = a(\theta)$ is the predictor vector introduced in (5), estimated from the real data by least squares, provides a rule to compute an estimate of θ.

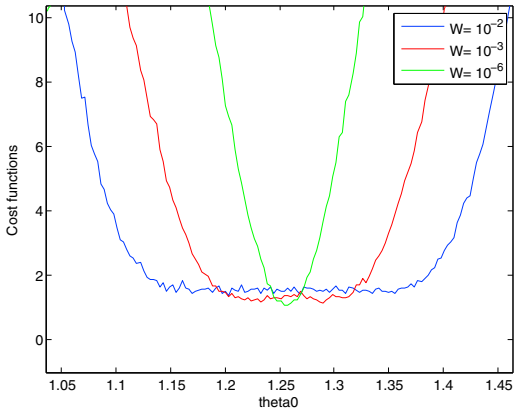
In conclusion, the estimation of θ can be done by first computing the least squares estimate $\hat{a}$ of a and then selecting the fundamental frequency among the zeros of the polynomial $\lambda^n + \sum_{k=1}^{n} \hat{a}_k \lambda^{n-k}$ which describes the difference equation for $\mathbf{y}$ equivalent to the state space model (23), (24). Subspace methods such as those developed in Favaro and Picci [2012] which are properly adapted to oscillatory signals could also be used. In the algorithm one must constrain the eigenvalues of the estimate of $A(\theta)$ to lie on the unit circle to extract easily the fundamental frequency.

6.1 Simulation results

Below we show plots of the PEM cost function to be minimized w.r.to the center frequency hyperparameter θ

$$V(\theta_0) := \frac{1}{T}\sum_{t=1}^{T}(y(t+1) - a(\theta)^\top z(t))^2$$

It is convenient to use the variable bandwidth W as a design parameter. In the figure below the cost function $V(\hat{\theta}_T)$ is shown for three values of the bandwidth W: 10^{-2},10^{-3},10^{-6}.



Notice that as the bandwidth W decreases, the cost function $V(\hat{\theta}_T)$ becomes sharper and the minimum becomes more pronounced. This minimum can be chosen as a new center frequency in an iterative algorithm where the prior distribution is iterated by refining the center frequency and concurrently restricting W in a suitable way. We skip the details for reasons of space.

7. CONCLUSIONS

The frequency estimation problem is an intrinsecally non linear parametric problem which has been approached

[2] For $W < 1/2$ $\mathbf{x}(t)$ is indeed a purely deterministic process.

in the literature by a variety of techniques which most of the times lead to iterative optimization algorithms of local nature. By formulating the problem as nonparametric Bayesian estimation of the hyperparameters of a prior distribution, the solution can be based on linear techniques, say subspace identification. This permits to rephrase the problem in a linear way. Simulations seem to indicate that the performance of this approach could be good but much further analysis is need to better understand and evaluate the idea. The approach could be extended to the estimation of signals with multiple harmonic components. This will be discussed in future publications.

REFERENCES

F. R. Bach, G.R.G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality and the smo algorithm. In *Proc o fthe 21st Int. Conf. on Machine Learning*, Banff Canada, 2004.

A. Chiuso. Regularization and bayesian learning in dynamical systems: Past, present and future (invited paper). In *Proc of the SYSID15*, Beijing China, 2015.

M. Favaro and G. Picci. A subspace algorithm for extracting periodic components from multivariable signals in colored noise. In *Proc. of 16th IFAC Symposium on System Identification (SYSID)*, pages 1150 –1155, Bruxelles, 2012.

G. Golub and C van Loan. *Matrix Computations*. Johns Hopkins U.P., N.Y., 1983.

E. J. Hannan and M. Deistler. *The statistical theory of linear systems*. John Wiley, N.Y., 1988.

T. Hoffmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36 (3):1171–1220, 2008.

Kedar Khare. Bandpass sampling and bandpass analogues of prolate spheroidal functions. *Signal Processing*, pages 1550 – 1558, 2006.

M. Lázaro-Gredilla, J. Q. Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

A. Lindquist and G. Picci. *Linear Stochastic Systems: a Geometric Approach to Modeling Estimation and Identification*. Springer Verlag, 2015.

L. Ljung. *System Identification; theory for the user*. Prentice Hall, Upper Saddle River N.J., 1999.

B. G. Quinn and E. J. Hannan. *The Estimation and Tracking of Frequency*. Cambridge U.P., 2001.

David Slepian. Prolate spheroidal wave functions, fourier analysis and uncertainty v: The discrete case. *Bell Syst. Tech. Jour.*, 57(5):1371–1430, 1978.

David Slepian and H.O. Pollak. Prolate spheroidal wave functions, fourier analysis and uncertainty i. *Bell Syst. Tech. Jour.*, 40:43–63, 1961.

Petre Stoica, Prabhu Babu, and Jian Li. Spice: a sparse covariance-based estimation method for array processing. *IEEE Trans. on Signal Process.*, 59(2):629–638, 2011.

Grace Wahba. *Spline models for observational data*. SIAM, Philadelphia, USA, 1990.

Appendix A. THE DISCRETE PROLATE SPHEROIDAL (DPS) SEQUENCES

One can relate the eigensequences (written as column vectors) $\boldsymbol{\varphi}_k$ to the Slepian's *discrete prolate spheroidal sequences* Slepian [1978] introduced in Slepian and Pollak [1961].

Let N be a fixed natural number. Without loss of generality we shall assume hereafter that N is odd equal to $2M+1$ for some integer M. The function $\dfrac{\sin \pi N f}{\sin \pi f}$ is called a *Dirichlet Kernel* denoted $D_N(f)$. For $f \in \mathbb{R}$ this is a periodic function. It is well-known that for $N \to \infty$ the Dirichlet kernel acts like a delta function, which is the commonly made approximation when using the FFT. The discrete prolate spheroidal sequences are solutions to the eigenvalue problem

$$\int_{-W}^{+W} \frac{\sin N\pi(f-n)}{\sin \pi(f-n)} \hat{\boldsymbol{\psi}}_{N,k}(n)\, dn = \lambda_k\, \hat{\boldsymbol{\psi}}_{N,k}(f)\,, \quad \text{(A.1)}$$

Since the kernel of the integral operator above has finite rank N,there are just N eigenvalues $\lambda_0 \geq \dots \lambda_k \geq \dots \lambda_{N-1} > 0$. Since the kernel is defined for all $f \in \mathbb{R}$, the eigenfunctions are also defined (and periodic) on the whole real line.

Since for $N \to \infty$ the operator tends to the identity all eigenvalues must tend to 1. On the other hand, for $W = 1/2$ and N finite the frequency convolution theorem (for finite Fourier transforms) implies that (A.1) has the time domain counterpart

$$\mathrm{rect}_{[-M,\,M]}(t)\boldsymbol{\psi}_{N,k}(t) = \lambda_k\, \boldsymbol{\psi}_{N,k}(t)\,,\ \boldsymbol{\psi}_{N,k}(t) = \mathcal{F}^{-1}(\hat{\boldsymbol{\psi}}_{N,k})$$

which also implies that all eigenvalues must be equal to one. In this case there are exactly N linearly independent eigenfunctions, that is N trigonometric polynomials which can be chosen orthogonal, say, each supported on a single time point in the interval $[-M,\, M]$. In other words, when $W = 1/2$ the $\boldsymbol{\psi}_{N,k}$ are *time limited functions*.When $W < 1/2$ it is not a priori clear if the support of $\boldsymbol{\psi}_{N,k}(t)$ is finite. In fact it is; see Theorem 2 below.

Consider now the Sinc function $\mathrm{Sinc}_W(k) := \dfrac{\sin(2\pi W k)}{2\pi W k}$, $k = 0, \pm 1, \pm 2 \dots$ where, as before $0 \leq W \leq 1/2$ and form the $N \times N$ matrix (N odd)

$$\mathbf{S}_N := \left[\frac{\sin 2\pi W(t-s)}{2\pi W(t-s)}\right], \qquad |t| \leq \frac{N-1}{2} \quad |s| \leq \frac{N-1}{2}$$

which is a positive definite Toeplitz matrix with eigendecomposition

$$\mathbf{S}_N \bar{\boldsymbol{\varphi}}_k = \bar{\lambda}_k \bar{\boldsymbol{\varphi}}_k \quad k = 0, 1, 2, \dots, N-1 \qquad \text{(A.2)}$$

The following is the key result which permits a direct computation of the eigenexpansion of (A.1).

Theorem 2. The (nonzero) eigenvalues of the two operators (A.1) and (A.2) coincide modulo the factor $2W$; i.e. $\lambda_k = 2W\bar{\lambda}_k$, $k = 0, 1, \dots, N-1$. The eigenvectors coincide modulo a factor of modulus one, namely $\boldsymbol{\psi}_{N,k}(t) = 2\pi W_k\, \bar{\boldsymbol{\varphi}}_k(t) \qquad t \in \mathbb{Z}\,.$