Automatica 142 (2022) 110362

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Empirical Bayes identification of stationary processes and approximation of Toeplitz spectra ☆

Giorgio Picci [a,*], Bin Zhu [b]

[a] Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131 Padova, Italy
[b] School of Intelligent Systems Engineering, Sun Yat-sen University, Waihuan East Road 132, 510006 Guangzhou, China

ARTICLE INFO

Article history:
Received 16 November 2020
Received in revised form 5 December 2021
Accepted 4 April 2022
Available online 9 May 2022

Keywords:
Toeplitz spectra approximation
Empirical Bayes estimation
Prolate spheroidal wave functions
Subspace methods
Frequency estimation

ABSTRACT

We study Statistical Consistency of an approximate subspace identification procedure for the infinite dimensional *a posteriori* model of a frequency estimation problem in an Empirical Bayesian framework. By first imposing a natural uniform prior probability density on the unknown frequencies, the estimation of the hyperparameters of the *a priori* distribution can be accomplished by a sequence of subspace identification techniques. These techniques exploit the special structure of the covariance matrix of the *a posteriori* process which is discovered by making a connection with classical results on energy concentration in deterministic signal processing. The convergence of the spectrum of the subspace estimates to the (nonrational) spectral density of the *a posteriori* process has an analytic counterpart in the approximation of symmetric positive definite Toeplitz matrices by submatrices of finite rank. This is proven by a weak-sense convergence theorem for Toeplitz spectra.

1. Introduction and motivation

The Empirical Bayes approach to parameter estimation (Lehmann & Casella, 1998) has in certain cases shown superior performance, providing a smaller Mean Square Error (MSE) than Maximum Likelihood (Reinsel, 1985; Yuan et al., 2016). There has consequently been some interest in approaching linear system identification by using this methodology. The procedure is however complicated since it requires as a preliminary step choosing a parametric class of prior distributions and estimate their relative parameters (which are called *hyperparameters* in the literature) from the available data. Successively one implements a standard Bayesian procedure based on the estimated prior, see Lehmann and Casella (1998, p. 262) and Aravkin et al. (2012), Chiuso (2016), Efron (2010, 2014).

A particular class of problems where this approach has shown to be feasible is the identification of a purely oscillatory linear system. For these systems the output signal (which is assumed scalar for simplicity) is the sum of sinusoidal oscillations with unknown deterministic frequencies and random amplitudes (called a *quasi-periodic process*), and a zero mean uncorrelated white noise. It is well-known that quasi-periodic processes are (wide-sense stationary and) purely deterministic. Casting the identification problem in a Bayesian setting, the unknown frequencies are modeled as *random variables*. The *a priori* description of these random variables is naturally chosen as a combination of simple local distributions concentrated about some nominal frequencies which are *a priori* unknown (see e.g., Zacharias et al., 2013) and play the role of hyperparameters. This model seems to be appropriate to describe a variety of applications where the frequencies can fluctuate slightly about nominal values.

A remarkable fact is that under such a natural class of *a priori* distributions on the unknown frequencies, the *a posteriori* description of the quasi-periodic message signal is still a purely deterministic process but this "a posteriori" process turns out to have a nonrational spectral density. In other words, it is infinite dimensional. The *a posteriori spectrum* of the message signal is no longer composed of a finite combination of spectral lines (Dirac functions) as for standard quasi periodic signals, but can generally be of the continuous type, although the spectral density should have a smaller support than the whole frequency range $[-\pi, \pi]$. A specific example with some remarkable structural properties will be described in the next section. Therefore, a faithful second-order description of such *a posteriori* signals requires infinite dimensional models. Their identification can therefore only be approached by inventing a suitable finite-dimensional approximation.

In an abstract Empirical Bayesian setting the posterior distribution of the signal (which is still parametric) becomes a function

☆ B. Zhu was partially supported by the "Hundred-Talent Program" of Sun Yat-sen University. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Tianshi Chen under the direction of Editor Torsten Söderström.

* Corresponding author.
E-mail addresses: picci@dei.unipd.it (G. Picci), zhub26@mail.sysu.edu.cn (B. Zhu).

https://doi.org/10.1016/j.automatica.2022.110362
0005-1098/

of the hyperparameters of the prior and the statistical problem can in principle be brought back to classical inference, now formulated in terms of hyperparameters. The question of *consistency*, once suitably interpreted, remains therefore a central issue. Because of the infinite dimensionality of the posterior model, identification, by whatever method, say subspace algorithms, needs to be implemented on an *approximate model*. Although subspace algorithms could *per se* provide consistent estimates for finite dimensional models, the problem of consistency now involves the approximation of the underlying infinite-dimensional true model, and its proof remains a challenging task.

1.1. Contributions

The main contributions of this paper are:

(1) We study a simple prototype of Empirical Bayes identification problem for a finite-dimensional quasi-periodic process. The computation of the *a posteriori* model can be solved explicitly and analyzed quite in detail. Examples of complete analytic solutions of problems of this kind are very seldom found in the literature and seem to only regard static linear regression problems.
(2) We address the hyperparameter estimation of the posterior model by a suitable sequence of approximate subspace methods. This can be interpreted as a specialization to our setting of the standard *marginal likelihood* approach suggested in the literature and considered for example in Aravkin et al. (2012) and Lázaro-Gredilla et al. (2010) which however needs to be solved by numerical optimization algorithms. Our work uses instead the special structure of the data process and need not involve optimization.
(3) Since the *a posteriori* model describes a process whose spectral density is not rational and hence cannot be modeled exactly by a linear stochastic system, the subspace method must by nature be approximate. Hence there is a natural question of *statistical consistency* which needs to be addressed. We approach the consistency of the subspace procedure by studying covariance approximation. The covariance of the *a posteriori* signal process is an infinite Toeplitz matrix which, contrary to finite-dimensional quasi periodic processes, does not have a finite rank. The proof of consistency hinges on a result of approximation of infinite Toeplitz matrices by Toeplitz matrices of finite rank which is believed to be new and of interest to researchers in stochastic systems theory.
(4) We propose a possible stochastic metric for this approximation and prove convergence of the spectra, although only in a weak sense. This is however sufficient for assessing statistical consistency (in mean square) of the subspace algorithm.

The final step of the identification procedure should be to compute a truly Bayesian *Maximum A Posteriori* (MAP) estimate of the system parameters given the estimated prior. This has been done in Picci and Zhu (2020, 2021), but including this part in the present manuscript would have increased its length beyond reasonable limits. The completion will appear in a future publication.

2. Preliminaries

Notations: All random variables considered in this paper will be real, zero-mean with finite variance. Most random processes are scalar wide sense (w.s.) stationary defined on the integer lattice $\mathbb{Z}$. Bold symbols like $\mathbf{y}$ or $\boldsymbol{\Sigma}$ are reserved for random variables or for *infinite* arrays, say stochastic processes or infinite covariance matrices. Hatted symbols will denote sample estimates. We shall use the standard stochastic real Hilbert space setting with inner product of (scalar) random variables $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle$ defined by the covariance $\mathbb{E}\boldsymbol{\xi}\boldsymbol{\eta}$ and denote by $\mathbf{H}(\mathbf{v})$ the subspace linearly generated by the components of a zero-mean process $\mathbf{v}$. The abbreviations p.d. and p.n.d. are shorthands for purely deterministic and purely nondeterministic processes.

Consider a scalar observation process $\mathbf{y}$ which is the sum of a purely deterministic message process $\mathbf{x}$ plus an uncorrelated white noise process $\mathbf{w}$ of variance ρ^2. Assume that the covariance function

$$\sigma(\tau) = \mathbb{E}\mathbf{x}(t+\tau)\mathbf{x}(t), \qquad \tau \in \mathbb{Z} \tag{1}$$

admits a Fourier transform

$$\varphi(e^{i\theta}) = \sum_{\tau=-\infty}^{+\infty} e^{-i\theta\tau}\sigma(\tau)$$

which is a piecewise smooth function of θ, called the *spectral density* of the process. For example φ is continuous and bounded when σ belongs to ℓ^1. In the example studied in this paper, this function will *vanish on a set of positive Lebesgue measure*. By a well-known criterion, e.g., Lindquist and Picci (2015, Theorem 4.7.5), this implies that $\mathbf{x}$ is purely deterministic. The process $\mathbf{y}$ is described by the infinite covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{y}} = [\sigma(t-s)]_{t,s\in\mathbb{Z}_+} + \rho^2\mathbf{I} := \boldsymbol{\Sigma}_{\mathbf{x}} + \rho^2\mathbf{I}$$

and the spectral density $\varphi_{\mathbf{y}}(e^{i\theta}) = \varphi(e^{i\theta}) + \rho^2$. Note that φ can be easily recovered from $\varphi_{\mathbf{y}}$ by subtracting the baseline constant and similarly for the infinite covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$. In this paper we shall essentially study this last object and for simplicity denote it by $\boldsymbol{\Sigma}$. We shall assume that $\varphi(e^{i\theta})$ is piecewise continuous and bounded, as for example is a rectangular function. Then $\boldsymbol{\Sigma}$ induces a bounded linear operator in ℓ^2 (Akhiezer & Glazman, 1961; Hartman & Wintner, 1954). The function φ is also called the *symbol* of $\boldsymbol{\Sigma}$.

3. Bayesian frequency estimation

In recent investigations (Favaro & Picci, 2015; Picci & Zhu, 2019, 2020) we have studied the Bayesian identification of the ubiquitous signal plus noise model

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{w}(t), \quad t \in \mathbb{Z} \tag{2}$$

where $\mathbf{x}$ is the sum of ν *random* oscillatory components (a quasi periodic process), that is,

$$\mathbf{x}(t) := \sum_{\ell=1}^{\nu} \mathbf{a}_\ell \cos(\boldsymbol{\omega}_\ell t) + \mathbf{b}_\ell \sin(\boldsymbol{\omega}_\ell t), \tag{3}$$

and $\mathbf{w}$ is additive Gaussian white noise. The angular frequencies $\boldsymbol{\omega}_\ell$ are unknown random variables but their number ν is fixed in advance. The model is specified as follows:

- the amplitude pairs $\mathbf{a}_k, \mathbf{b}_k$ are zero-mean pairwise and mutually uncorrelated for all k and the two components $\mathbf{a}_k, \mathbf{b}_k$ have equal variance: $\sigma_{0,k}^2 := \mathrm{var}[\mathbf{a}_k] = \mathrm{var}[\mathbf{b}_k]$, $k = 1, \dots, \nu$;
- each angular frequency $\boldsymbol{\omega}_\ell$ is a *random variable* taking values in the interval $[0, \pi]$, independent of the amplitudes;
- The noise $\mathbf{w}(t)$ is zero-mean stationary white Gaussian, of variance $\sigma_{\mathbf{w}}^2$, independent of everything else.

Let $\boldsymbol{\omega} := [\boldsymbol{\omega}_1 \ \dots \ \boldsymbol{\omega}_\nu]^\top$ and denote by $\mathbf{a}$, $\mathbf{b}$ two similarly arranged amplitude vectors. Note that the model is linear in $\mathbf{a}$, $\mathbf{b}$, and hence estimation of the amplitudes and their variance is just a standard linear estimation problem when the frequencies are known.

A simplified *a priori* model for the components $\boldsymbol{\omega}_\ell$ of the random vector $\boldsymbol{\omega}$ is a uniform distribution on the frequency band $[\theta_\ell - W_\ell, \theta_\ell + W_\ell]$ such that the symmetrized sets with respect to the origin

$$S_\ell := [\theta_\ell - W_\ell, \theta_\ell + W_\ell] \cup [-\theta_\ell - W_\ell, -\theta_\ell + W_\ell] \tag{4}$$

for $\ell = 1, \dots, \nu$ do not overlap. For simplicity we shall assume that the assigned bandwidth is the same for different frequencies, i.e., $W_1 = \cdots = W_\nu = W$. Here $0 \le \theta_\ell \le \pi$ is called a *center frequency* and $0 \le W \le \pi$ the *bandwidth*. Both θ and W are the *hyperparameters* of the *a priori* distribution for the frequency $\boldsymbol{\omega}$.

The assumptions imply that for each fixed frequency value ω the ν components, say $\mathbf{x}_\ell$, $\ell = 1, \dots, \nu$ of the signal (3) are stationary uncorrelated processes. Hence the covariance function of the process $\mathbf{y}$ for a fixed (deterministic) sample value ω has the form

$$\sigma_y(t, s \mid \omega) := \mathbb{E}\{\mathbf{y}(t)\mathbf{y}(s) \mid \omega\} = \sigma(t, s \mid \omega) + \sigma_{\mathbf{w}}^2\, \delta(t, s) \tag{5}$$

where $\delta(t, s)$ is the Kronecker symbol, and

$$\sigma(t, s \mid \omega) := \sum_{\ell=1}^{\nu} \mathbb{E}\{\mathbf{x}_\ell(t)\mathbf{x}_\ell(s) \mid \omega\} = \sum_{\ell=1}^{\nu} \sigma_\ell(t, s \mid \omega)$$

is the *a priori* conditional covariance of the signal $\mathbf{x}$ given $\boldsymbol{\omega} = \omega$. To lighten the notation, we suppress the subscripts. The formulas below should be interpreted as holding for a generic index ℓ. By the stated assumptions, the following computation is straightforward:

$$\begin{aligned}\sigma(t, s \mid \omega) &= \mathbb{E}\left\{\mathbf{a}^2 \cos(\omega t)\cos(\omega s) + \mathbf{ab}\cos(\omega t)\sin(\omega s)\right.\\ &\quad \left. + \mathbf{ab}\sin(\omega t)\cos(\omega s) + \mathbf{b}^2 \sin(\omega t)\sin(\omega s)\right\}\\ &= \sigma_0^2 \cos\omega\tau\end{aligned} \tag{6}$$

where $\tau := t - s$, and then computing the *a posteriori* covariance by integrating the function with respect to the uniform prior density, one gets

$$\begin{aligned}\sigma(t, s) = \sigma_0^2\, \mathbb{E}(\cos\boldsymbol{\omega}\tau) &= \sigma_0^2 \int_{\theta-W}^{\theta+W} \cos(\omega\tau)\frac{1}{2W} d\omega\\ &= \sigma_0^2 \cos(\theta\tau)\frac{\sin W\tau}{W\tau}.\end{aligned} \tag{7}$$

Since the covariance function depends only on τ, the signal $\mathbf{x}$ is stationary, and so is $\mathbf{y}$. In the following, we will write $\sigma(\tau)$ in place of $\sigma(t, s)$. Note that the *a posteriori* covariance of the signal is *no longer a finite dimensional kernel* and, as we shall see in a minute its spectrum is far from being rational.

For $\theta = 0$, the covariance function $\sigma(\tau)$ is the well-known *Sinc function*, which is the inverse Fourier transform of a rectangular function with a support $[-W, W]$, namely

$$\sigma_0^2 \frac{\sin W\tau}{W\tau} = \frac{\sigma_0^2}{2W}\int_{-W}^{W} e^{i\omega\tau} d\omega. \tag{8}$$

It follows that a zero-frequency component of the process $\mathbf{x}$ must have a uniform spectral density $\frac{\pi\sigma_0^2}{W}\chi_{[-W,W]}(\omega)$. When $W = \pi$, the process is just a usual stationary white noise of variance σ_0^2. For $W < \pi$, the process $\mathbf{x}$ is nontrivial, called a *bandlimited white noise* within the frequency band $[-W, W]$. In this case, it is a purely deterministic process with an absolutely continuous spectral distribution, since the logarithm of the density is obviously not integrable, see e.g., Lindquist and Picci (2015, p. 144).

One is primarily interested in the case $\theta_\ell \neq 0$, for which we make the assumption that $|\theta_\ell| > W$ for all ℓ, so that each single support set S_ℓ in (4) is composed of two *disjoint* intervals. Then the last expression in (7) can be rewritten as

$$\begin{aligned}\sigma_0^2 \cos(\theta\tau)\frac{\sin W\tau}{W\tau} &= \frac{\sigma_0^2}{4W}\int_{-\pi}^{\pi} \cos(\omega\tau)\chi_S(\omega) d\omega\\ &= \frac{\pi\sigma_0^2}{2W}\int_{-\pi}^{\pi} e^{i\omega\tau}\chi_S(\omega)\frac{d\omega}{2\pi}\end{aligned} \tag{9}$$

where χ_S is the indicator function of S, and the second equality holds due to the symmetry of the integrand. From the above relation, one sees that the spectral density of the process $\mathbf{x}$ is the sum of ν disjoint spectral terms, each of a rectangular shape:

$$\varphi_{\mathbf{x}_\ell}(\omega) = \frac{\pi\sigma_{0,\ell}^2}{2W}\left(\chi_{[\theta_\ell - W,\, \theta_\ell + W]} + \chi_{[-\theta_\ell - W,\, -\theta_\ell + W]}\right), \tag{10}$$

where $\ell = 1, \dots, \nu$. The signal $\mathbf{x}$ can then be described as a sum of ν independent deterministic carriers, each of angular frequency θ_ℓ, amplitude-modulated by a bandlimited white noise process as described before. Note that the generated subspace $\mathbf{H}(\mathbf{x})$ is always infinite dimensional. The single covariance function (7) has been called a *modulated Sinc kernel* in Khare (2006), where it arises in a different deterministic context.

A remarkable fact is that the *a posteriori* covariance operator with kernel $\sigma(t-s)$ has very similar properties to those that have been uncovered in the 60s and 70s by D. Slepian and coworkers in a famous series of papers concerning the energy concentration problems of time and band limited signals (Landau & Pollak, 1961; Landau & Widom, 1980; Slepian, 1978; Slepian & Pollak, 1961). The key property of the covariance operator in question is that its eigenvalues decay extremely fast to very small values (nearly zero) for indices greater than an *a priori* computable number n, called the *Slepian frequency*.

In practice we can only observe sample paths of finite length N from the process $\mathbf{y}$. Collect the observed random variables into a column vector, and in particular, let $\mathbf{x}_N := [\mathbf{x}(t), \mathbf{x}(t+1), \dots, \mathbf{x}(t+N-1)]^\top$. Then consider the $N \times N$ covariance matrix

$$\begin{aligned}\Sigma_N &:= \mathbb{E}\{\mathbf{x}_N\mathbf{x}_N^\top\}\\ &= \begin{bmatrix} \sigma(0) & \sigma(1) & \cdots & \sigma(N-1)\\ \sigma(1) & \sigma(0) & \cdots & \sigma(N-2)\\ \vdots & \vdots & \ddots & \vdots\\ \sigma(N-1) & \sigma(N-2) & \cdots & \sigma(0)\end{bmatrix}.\end{aligned} \tag{11}$$

This symmetric Toeplitz structure of the covariance matrix comes from the fact that the process is stationary and real-valued. Similarly, we can define the $N \times N$ covariance matrix of the process $\mathbf{y}$, say $\Sigma_{\mathbf{y},N}$, and we have the relation

$$\Sigma_{\mathbf{y},N} = \Sigma_N + \sigma_{\mathbf{w}}^2 I_N. \tag{12}$$

Analysis of the eigen-structure of Σ_N will be of great importance to our frequency estimation problem, and that will be the content of the next section.

4. Eigenvalues of the covariance matrix

Let S be the set that is a union of a finite number of pairwise disjoint closed subintervals of $[-\pi, \pi]$, like the one in (4). Obviously the spectrum $\phi_{\mathbf{x}}(\omega)$ has the same support of the characteristic function $\chi_S(\omega)$. Consider the inverse Fourier transform

$$\rho(t) := \frac{1}{2\pi}\int_S e^{it\omega} \mathrm{d}\omega, \quad t \in \mathbb{Z}. \tag{13}$$

This is just the impulse response of the ideal bandpass filter $\chi_S(\omega)$ which in our context can be interpreted as a spectral density, a sort of "global" renormalization of (10) so that for $\nu = 1$ the covariance function $\sigma(t)$ is just a scalar multiple of $\rho(t)$ via $\sigma(t) =$

$\frac{\pi\sigma_0^2}{2W}\rho(t)$. Observe that the function ρ has the symmetry $\rho(-t) = \overline{\rho(t)}$ where z^* means the complex conjugate (transpose) of $z \in \mathbb{C}$.

Define the Hermitian Toeplitz matrix

$$R = \begin{bmatrix} \rho(0) & \rho(-1) & \cdots & \rho(-N+1) \\ \rho(1) & \rho(0) & \cdots & \rho(-N+2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(N-1) & \rho(N-2) & \cdots & \rho(0) \end{bmatrix}. \tag{14}$$

By well-known properties of covariance functions, for an arbitrary complex vector $y = [y_0, \dots, y_{N-1}]^\top$ of length N, the quadratic form y^*Ry is positive definite. Notice that when the set S is symmetric with respect to the origin as in (4), the integral in (13) reduces to $\int_S \cos(t\omega)\mathrm{d}\omega$. In that case, ρ is an even function of time, and the matrix R is real symmetric.

Now, the largest eigenvalue of R is well-known to be equal to the maximum of the Rayleigh quotient associated to R. By the min–max theorem, this maximum is attained when y is the corresponding eigenvector. It is also obvious that the eigenvalues of R cannot exceed 1, simply because the maximum of the corresponding spectrum $\chi_S(\omega)$ is just equal to 1.

4.1. Asymptotic behavior of the eigenvalues of R

We shall now allow the dimension of R to increase. In other words, the integer N introduced earlier is considered as a variable tending to infinity. For notational consistency we shall then add the subscript $_N$ to R. Let $\lambda_j(N)$ be the j-th eigenvalue (arranged in nonincreasing order) of R_N. We know from the previous subsection that $0 < \lambda_j(N) \leq 1$ for all $j = 1, \dots, N$. It also follows easily from (13) that

$$\sum_{j=1}^{N} \lambda_j(N) = \operatorname{tr} R_N = N\rho(0) = \frac{\mathrm{m}(S)}{2\pi}N, \tag{15}$$

where the notation $\mathrm{m}(\cdot)$ denotes the Lebesgue measure of a set. Now for a real number $0 < \gamma < 1$, define $n(\gamma, N)$ to be the number of eigenvalues of R_N that are no less than γ. Note that this number is just the *numerical rank* of the matrix R_N with threshold γ. The next result is a first-order description of the asymptotic eigenvalue distribution of the matrix R_N.

Theorem 1. *It holds that*

$$\lim_{N\to\infty} \frac{n(\gamma, N)}{N} = \frac{\mathrm{m}(S)}{2\pi} \tag{16}$$

independent of γ. Equivalently, for $N \to \infty$, the covariance matrix R_N has numerical rank equal to

$$n = N\mathrm{m}(S)/2\pi, \tag{17}$$

independent of γ and all the nonzero eigenvalues tend to 1.

Proof. Formula (16) follows from a famous theorem of Szegö for the asymptotic eigenvalue distribution of Toeplitz matrices (Grenander & Szegő, 1958, p. 65) as the support of the "spectral density" $\chi_S(\omega)$ associated to the normalized covariance function ρ is precisely the set S. In plain words, the fraction of *dominant* positive eigenvalues of R_N in the integer interval $[0, N]$ is asymptotically equal to the fraction of the measure of the spectral support of its symbol ρ to that of the whole frequency domain. □

Remark 2. The above proof has also been reported in Picci and Zhu (2021). A more precise formula for the asymptotic eigenvalue distribution of R_N is given in Landau and Widom (1980) for the continuous-time case. The number of "transient" eigenvalues (those that lie between the dominant and the ones which are essentially zero) is shown to be proportional to $\log N$. Slepian's asymptotic expressions for the eigenvalues, valid for $\theta = 0$, are also reported in Thomson (1982, p. 1059). Although we believe that analogous discrete-time estimates should hold, a formal proof is yet to be worked out. For our problem of frequency estimation, Theorem 1 is anyway sufficient.

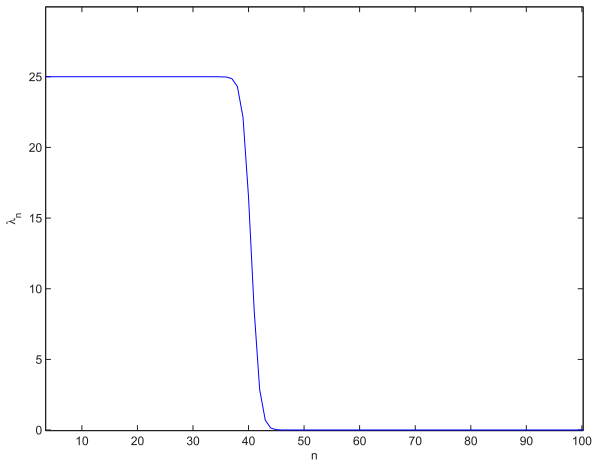

Fig. 1. Eigenvalues of the Sinc kernel covariance matrix, numerical rank ≈ 40.

The decay of eigenvalues is very fast, as it can be shown that the matrix R_N has only $o(N)$ eigenvalues that are strictly between 0 and 1, and for large sample size they can be reasonably neglected. For $\nu = 1$ the covariance matrix in (11) is just a scalar multiple of R_N via $\Sigma_N = \frac{\pi\sigma_0^2}{2W}R_N$. Clearly, the constant factor only rescales the eigenvalues. In particular, the assertion on the asymptotic numerical rank in Theorem 1 holds for Σ_N. Below we show some simulations of how the eigenvalues decay.

Fig. 1 shows the behavior of the eigenvalues μ_k of the Sinc kernel ($\theta = 0$) for $N = 1000, W/2\pi = 0.02, \sigma^2 = 1$ which yields a numerical rank approximately equal to 40. We can clearly see that for $n < 40$ the eigenvalues are all equal to the same constant while for $n > 40$ the μ_k's very quickly decrease to zero. The behavior of the eigenvalues of R_N is the same except that the normalization makes the μ_k all practically equal to one for $k < n$. In order to get the same normalization we just need to substitute μ_k with $W\mu_k/\pi$.

As for the modulated Sinc kernel, Fig. 2 shows the eigenvalues of a matrix Σ_N with the same values of N, W, and σ^2. One sees 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 $4NW/2\pi = 80$, i.e., twice the value of n for the sinc kernel. Moreover, the amplitudes of the eigenvalues for $k < n$ are half of those of the sinc kernel, for equal values of W. This follows from the symmetry of the spectrum and matches also the experimental findings of Khare (2006). In order to get the largest eigenvalues of the modulated Sinc kernel equal to one, a different normalization should be made by substituting μ_k with 2 $W\mu_k/\pi$. This agrees with the matrix rescaling described above.

5. Covariance estimation

For clarity of exposition, we shall now assume that $\nu = 1$ and neglect the subscript $_\ell$ altogether. The generalization to multiple sinusoids, i.e., $\nu > 1$, will be straightforward. In the case of one hidden frequency, we have rank $\Sigma_N \approx \frac{2W}{\pi}N$ according to Theorem 1. We see that the bandwidth W can be inferred from

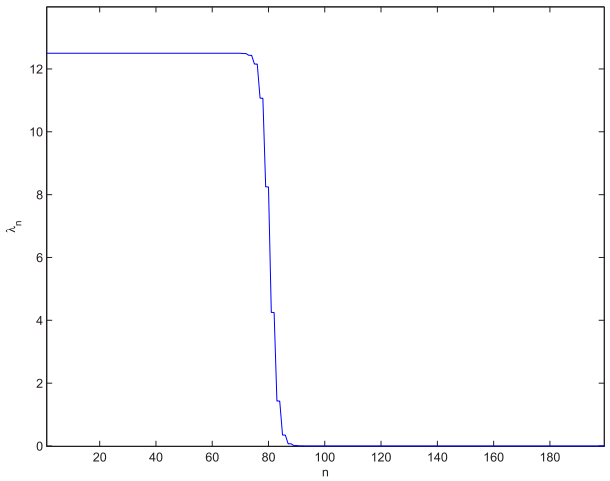

Fig. 2. Eigenvalues of the modulated Sinc kernel covariance matrix, numerical rank ≈ 80.

the asymptotic numerical rank of the signal covariance matrix Σ. Since our measurements come from the process $\mathbf{y}$, we start by estimating its covariance matrix $\Sigma_{\mathbf{y}}$.

A well-known difficulty in frequency estimation is that stationary random processes with periodic components, even when the frequencies are exactly known, are not ergodic. Nonergodicity means in particular that, when the sample size goes to infinity, the limit of the process sample covariance is *sample dependent*, that is, the limit sample covariance depends on the random amplitudes of its elementary oscillatory components, see e.g., Söderström and Stoica (1989, pp. 105–109). This lack of ergodicity is even more serious when the frequency is random. For this reason, one-sample-path estimation runs into difficulty and the standard approach in many practical situations is to consider estimation from cross-sectional or *panel data* (also called *snapshots*), as described in e.g., Hsiao (2014) and done in the DOA estimation. Cross-sectional frequency data can be the result of parallel measurements by multiple sensors which is quite common for example in testing of turbo, and in general rotating machines, but also in many directional signal processing and biomedical applications.

For the reasons above, we shall need to assume that our observed data consist of L strings of sample observations (snapshots), assumed for simplicity all of length N:

$$y_k(t) = a_k \cos(\omega_k t) + b_k \sin(\omega_k t) + w_k(t), \tag{18}$$

where $k = 1, \dots, L$, $t = 1, \dots, N$, $(a_k,\ b_k)$ are sample determinations of the random variables $(\mathbf{a},\ \mathbf{b})$, and the frequencies ω_k are sample determinations of the random variable $\boldsymbol{\omega}$ which is uniformly distributed on the fixed interval $[\theta - W,\ \theta + W]$. We assume that noises of different cross sections are independent. Furthermore, we assume that the random samples $[a_k, b_k, \omega_k]$ come from i.i.d. copies of $[\mathbf{a}, \mathbf{b}, \boldsymbol{\omega}]$, then the covariance matrix can be estimated by first subtracting the sample mean from the data, i.e.,

$$\tilde{y}_k(t) := y_k(t) - \frac{1}{N}\sum_{t=1}^{N} y_k(t)$$

and then doing a cross-sectional average

$$\hat{\Sigma}_{\mathbf{y},N} := \frac{1}{L}\sum_{k=1}^{L} \mathcal{Y}_k \mathcal{Y}_k^{\top}, \tag{19}$$

where $\mathcal{Y}_k = \begin{bmatrix}\tilde{y}_k(1) & \cdots & \tilde{y}_k(N)\end{bmatrix}^{\top}$ is a column vector of centered data. The procedure is asymptotically equivalent (for $L \to \infty$) to first computing the standard (biased) covariance estimator within each sample path (Stoica & Moses, 2005, Chapter 2),

$$\hat{\sigma}_k(\tau) := \frac{1}{N}\sum_{t=1}^{N-\tau} \tilde{y}_k(t+\tau)\tilde{y}_k(t),$$

constructing the sample *symmetric Toeplitz* estimate

$$\hat{\Sigma}_k := \mathrm{SymToep}\{\hat{\sigma}_k(0), \dots, \hat{\sigma}_k(N-1)\} \tag{20}$$

and then doing cross sectional average with respect to k to obtain $\hat{\Sigma}_{\mathbf{y},N}$ which is still symmetric Toeplitz (here the subscript N just refers to the dimension which is fixed). By the strong law of large numbers, we have

$$\hat{\Sigma}_{\mathbf{y},N} \to \Sigma_{\mathbf{y},N} \text{ as } L \to \infty \tag{21}$$

almost surely. Let $\hat{\lambda}_N$ be the smallest eigenvalue of $\hat{\Sigma}_{\mathbf{y},N}$. Then given (12) and Theorem 1, we have

$$\lim_{L,N\to\infty} \hat{\lambda}_N = \sigma^2_{\mathbf{w}}. \tag{22}$$

The limit here is understood as first letting $L \to \infty$ and then $N \to \infty$. In this sense we are able to build a consistent estimator of the signal covariance matrix Σ_N:

$$\hat{\Sigma}_N := \hat{\Sigma}_{\mathbf{y},N} - \hat{\lambda}_N I_N\,. \tag{23}$$

Thus a consistent estimator of the signal variance is given by

$$\hat{\sigma}_{\mathbf{x}}(0) := \hat{\sigma}_{\mathbf{y}}(0) - \hat{\lambda}_N, \tag{24}$$

since we have $\sigma_{\mathbf{y}}(0) = \sigma_{\mathbf{x}}(0) + \sigma^2_{\mathbf{w}}$ by (5).

Next, the rank of Σ_N (i.e., the *Slepian frequency*) may be approximated using Theorem 1 as

$$\mathrm{rank}(\Sigma_N) \simeq \frac{2W}{\pi}N \tag{25}$$

with an approximation error which roughly grows as $O(\log N)$. Hence the numerical rank of the sample signal covariance matrix $\hat{\Sigma}_N$, written as $\mathrm{rank}(\hat{\Sigma}_N)$ (not to be confused with the true rank of $\hat{\Sigma}_N$), can be estimated by locating the index at which its eigenvalues collapse to zero. From (25) we can then obtain an estimator of W:

$$\hat{W} := \frac{\pi}{2}\frac{\mathrm{rank}(\hat{\Sigma}_N)}{N} \to W \tag{26}$$

as $N \to \infty$. Unfortunately this estimator of W depends heavily on the estimate of the numerical rank whose computation is delicate and is not very reliable unless N is large.

6. A subspace approach to hyperparameter estimation

Consider now the general measurement model (2), with the signal $\mathbf{x}$ consisting of multiple sinusoids as in (3) satisfying all assumptions listed in Section 3. The covariance of $\mathbf{y}$ can then be computed similarly to that in (7). We have

$$\begin{aligned}\sigma_{\mathbf{y}}(\tau) &= \sigma(\tau) + \sigma^2_{\mathbf{w}}\delta(\tau, 0)\\ &= \frac{\sin W\tau}{W\tau}\sum_{\ell=1}^{\nu}\sigma^2_{0,\ell}\cos\theta_\ell\tau + \sigma^2_{\mathbf{w}}\delta(\tau,0)\\ &= \frac{\pi}{2W}\int_{-\pi}^{\pi} e^{i\omega\tau}\sum_{\ell=1}^{\nu}\sigma^2_{0,\ell}\,\chi_{S_\ell}(\omega)\frac{\mathrm{d}\omega}{2\pi} + \sigma^2_{\mathbf{w}}\delta(\tau,0)\end{aligned} \tag{27}$$

where the sum $\sum_{\ell=1}^{\nu}\sigma^2_{0,\ell}\,\chi_{S_\ell}(\omega)$ is a combination of indicator functions on the sets S_ℓ with possibly different weights $\sigma^2_{0,\ell}$. From the integral expression for the covariance function, we see

immediately that Theorem 1 is applicable, and the asymptotic rank of Σ_N is now $\frac{2\nu W}{\pi}N$. A rank estimator for the bandwidth W similar to (26) can be used since we have assumed that the supporting intervals for different frequencies have the same bandwidth. A more general situation with different W's can also be dealt with but it yields complicated formulas and will not be discussed here. Next, we will concentrate on the estimation of the center frequency vector $\theta := [\theta_1, \dots, \theta_\nu]^\top$.

Efficient estimation of the hyperparameters, assuming Gaussian additive noise, is generally based on maximum likelihood. See MacKay (1992, p. 429) for a general discussion of this point. The Gaussian *a posteriori* likelihood function based on the k-th snapshot of N data can be written as (Hannan & Deistler, 1988)

$$l_k(\theta, W) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det \Sigma_{\mathbf{y}}(\theta, W) - \frac{1}{2}\mathcal{Y}_k^\top \Sigma_{\mathbf{y}}(\theta, W)^{-1}\mathcal{Y}_k, \tag{28}$$

where $\mathcal{Y}_k$ is the vector introduced in (19) and $\Sigma_{\mathbf{y}}(\theta, W)$ is the theoretical *a posteriori* covariance matrix of $\mathcal{Y}_k$, with entries given in (27) which do not depend on the index k. By the independence of the sample paths, the log-likelihoods add to each other so that we end up with maximization of the function

$$l(\theta, W) = -\frac{L}{2}\log\det \Sigma_{\mathbf{y}}(\theta, W) - \sum_{k=1}^{L}\frac{1}{2}\mathcal{Y}_k^\top \Sigma_{\mathbf{y}}(\theta, W)^{-1}\mathcal{Y}_k \tag{29}$$

with respect to θ, W. This leads to the well-known unique maximizer, see e.g., Söderström and Stoica (1989, pp. 202–203), for the covariance matrix

$$\Sigma_{\mathbf{y}}(\theta, W) = \hat{\Sigma}_{\mathbf{y},N} \tag{30}$$

where $\hat{\Sigma}_{\mathbf{y},N}$ is defined in (19). Such an equation should be solved for the unknown hyperparameters (θ, W) appearing in the known structure (27). Note that this equation can be interpreted as resulting from the well-known *method of moments* which is the theoretical basis of Subspace Methods (Lindquist & Picci, 2015, Chap. 13). Since the equation is nonlinear, one may think of setting up at the outset an iterative solution scheme. However, these numerical algorithms very often converge only locally. In fact, the likelihood function is nonconvex and contains many flat regions. Therefore, brute-force optimization seems to be a hard task.

We shall instead take inspiration from (30) to propose a *subspace-based* approach. Of course subspace methods are restricted to *finite dimensional* linear models. For a fixed N, we may and shall here assume that the truncated $N \times N$ covariance matrix Σ_N of the process $\mathbf{x}$ has precisely rank $n := \frac{2\nu W}{\pi}N$. As discussed in Section 4.1, for N large this is a reasonable approximation. Hence for each N the problem can be phrased as the subspace identification of a finite dimensional process of rank n approximating the *a posteriori* message $\mathbf{x}$.

In terms of (a strict-sense) approximation of random signals, we are just looking for a jointly stationary zero-mean process $\mathbf{z}$, spanning a finite-dimensional subspace $\mathbf{H}(\mathbf{z}) \subset \mathbf{H}(\mathbf{x})$ *of dimension* n. Both $\mathbf{x}$ and $\mathbf{z}$ will occasionally be written as doubly infinite column vectors. Any such process $\mathbf{z}$ must also be purely deterministic. Abusing the terminology, we shall say that it is of rank n, even if saying that it has *dimension* n would be more appropriate. Note that $\mathbf{z}$ is uniquely determined by any finite string of random variables $\{\mathbf{z}(t)\}_{t\in I}$ induced on an interval I of length $N \geq n$. This follows from the statement in Lindquist and Picci (2015, pp. 276–277) which is reported below for completeness.

Lemma 1. *Any p.d. process* $\mathbf{z}$ *of rank* n *can be represented for all* $t \in \mathbb{Z}$ *by a state–space model (i.e., a stochastic realization) of the form*

$$\boldsymbol{\xi}(t+1) = A\boldsymbol{\xi}(t) \tag{31a}$$

$$\mathbf{z}(t) = c\,\boldsymbol{\xi}(t) \tag{31b}$$

where $\boldsymbol{\xi}(t) = [\,\xi_1(t), \xi_2(t), \dots, \xi_n(t)\,]^\top$ *is a* n*-dimensional basis vector spanning the Hilbert space* $\mathbf{H}(\mathbf{z}_N)$ *linearly generated by the* $N \geq n$ *random variables of the set* $\{\mathbf{z}(s) : t \geq s \geq t - N + 1\}$*,* A *is a* $n \times n$ *unitary matrix and* c *is a* n*-dimensional row vector such that the pair* (c, A) *is observable.*

Proof. See Lindquist and Picci (2015, p. 277). □

This linear state–space realization provides a representation of $\mathbf{z}(t)$ as a deterministic autoregression of order n and hence extends in time the finite family of random variables $\{\mathbf{z}(s)\}$, generators of $\mathbf{H}(\mathbf{z}_N)$, to a stationary p.d. process $\mathbf{z}$ defined on the whole time domain $\mathbb{Z}$. Since this realization is uniquely determined by the space $\mathbf{H}(\mathbf{z}_N)$ modulo a choice of basis, it follows that the process $\mathbf{z}$ is also uniquely determined by the finite dimensional subspace $\mathbf{H}(\mathbf{z}_N)$. Hence all covariances $\gamma(\tau) = \mathbb{E}\mathbf{z}(t+\tau)\mathbf{z}(t)$ are also uniquely defined and determine all the entries of the infinite covariance matrix of the process. It is worth remembering that the function $\gamma(\tau)$ must be a periodic function of τ containing n oscillatory modes and that the frequencies of these n modes are exactly determined by the eigenvalues of the unitary matrix A, which must belong to the unit circle of the complex plane.

Given the cross sectional measurements (18) of size $L \times N$, we summarize our subspace algorithm below:

(1) Compute $\hat{\Sigma}_{\mathbf{y},N}$, an estimate of the covariance matrix of $\mathbf{y}$, using (19);
(2) Subtract the smallest eigenvalue to obtain the signal covariance estimate $\hat{\Sigma}_N$, see (23),
(3) Estimate the numerical rank n of the signal covariance matrix and then estimate the bandwidth W via a formula like (26);
(4) Do eigen-decomposition to $\hat{\Sigma}_N$, keep the largest n eigenvalues, and call the $N \times n$ matrix of corresponding eigenvectors H_N;
(5) Let $k = N - 1$, and let now $\downarrow H_k$ be the matrix H_k deprived of its first row and $\uparrow H_k$ be the same matrix deprived of its last row. Solve the *shift-invariance* equation

$$\uparrow H_k A = \downarrow H_k$$

by the Procrustes algorithm (Golub & van Loan, 2013) to get a unique orthogonal $n \times n$ matrix A;
(6) Compute the eigenvalues of A, and extract their phase angles (between $-\pi$ and π);
(7) Run a clustering algorithm, e.g., k-means, on the phase angles, and take the centers of final clusters as estimates of the center frequencies.

As declared in point (3) the bandwidth W can be estimated from the asymptotic numerical rank of the covariance matrix. Given finite data strings, however, some ad-hoc schemes are usually needed. One possibility could be to estimate n by maximizing the ratio

$$\underset{k\in\{1,\dots,N-1\}}{\operatorname{argmax}} \frac{\lambda_k^2(\hat{\Sigma}_{\mathbf{y},N})}{\lambda_{k+1}^2(\hat{\Sigma}_{\mathbf{y},N})}, \tag{32}$$

where $\lambda_k(\hat{\Sigma}_{\mathbf{y},N})$ denotes the kth eigenvalue of the estimated covariance matrix $\hat{\Sigma}_{\mathbf{y},N}$ of the observation process, arranged in nonincreasing order. Intuitively, the maximum should be attained at the beginning of the flat region in the eigen-plot, see e.g., Figs. 1 and 2.

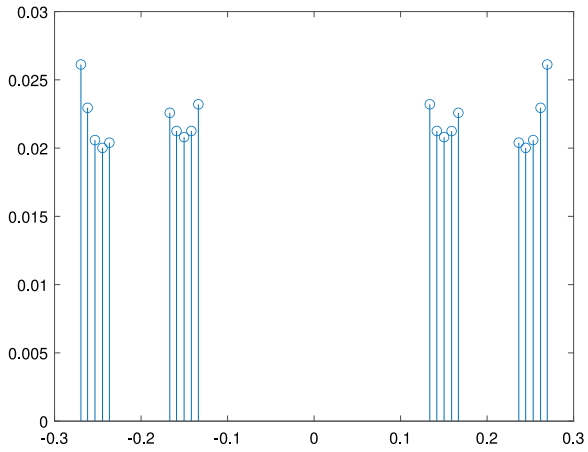

Fig. 3. Discrete spectrum estimate with two hidden frequencies. The true hyperparameters are $[\theta_1, \theta_2, W] = 2\pi \times [0.1499, 0.2524, 0.0155]$ and the estimated band centers are $\hat{\theta} = 2\pi \times [0.1503, 0.2532]$.

In the last step of this subspace algorithm, the center of each cluster may be obtained by simply taking the average of all the points in the cluster. This yields the estimate

$$\hat{\theta}_\ell = \frac{1}{n_\ell} \sum_{k=1}^{n_\ell} \varphi_{k,\ell} \qquad \ell = 1, \dots, \nu \tag{33}$$

where n_ℓ is the number of phase points in each cluster of positive phases.

Fig. 3 shows the discrete spectrum of the process $\mathbf{z}(t)$, output of the approximate state–space model (31), in one simulation trial in the case of $L = 100$. The horizontal axis is scaled to represent the frequency in Hz. In this particular trial, the true hyperparameters are $[\theta_1, \theta_2, W] = 2\pi \times [0.1499, 0.2524, 0.0155]$, and the estimated band centers are $\hat{\theta} = 2\pi \times [0.1503, 0.2532]$. One can see that the Dirac deltas indeed cluster around the true center frequencies inside the supporting interval.

6.1. The question of consistency

We shall give for granted that a subspace identification procedure for finite dimensional stochastic systems, assuming a true model exists, provides a consistent estimate of the "external" description (e.g., the covariance or spectrum) of a finite dimensional model describing a finite rank p.d. process $\mathbf{z}$; see Favaro and Picci (2012) and Lindquist and Picci (2015, Sec. 13.4) for a general proof of consistency of subspace algorithms. In this finite-dimensional setting consistency can be seen just as a consequence of the (exact) stochastic realization procedure underlying any subspace algorithm. Note however that here the "true" process $\mathbf{x}$ may in general be described (in the wide sense) by an *infinite-dimensional* linear stochastic model or by a nonrational spectral density function of which any identified finite dimensional model (31) can only provide, for each finite N, an approximation (in some sense yet to be ascertained). Hence the question of statistical consistency should be posed as that of discovering if, and in what sense, the sequence of identified models (31) may converge when $N \to \infty$, to a "true" description of the infinite dimensional signal.

Each n-dimensional realization (31) defines a p.d. process $\mathbf{z}$, output of the identified finite-dimensional model (31), of rank n. We shall denote the *infinite* covariance matrix of this p.d. process by the boldface symbol $\boldsymbol{\Gamma}$, notation specialized to $\boldsymbol{\Gamma}(n)$ when studying its dependence on the dimension n (or equivalently, on the sample size N). Clearly $\boldsymbol{\Gamma}(n)$ is a positive semidefinite infinite symmetric Toeplitz matrix of rank n. In this way we may construct a sequence of "approximations" of $\boldsymbol{\Sigma}$, each having a finite rank n. The symbol (spectral density) of $\boldsymbol{\Gamma}(n)$, say $\varphi_n(e^{i\theta})$, is the sum of $2n$ Dirac pulses supported on $[-\pi, \pi]$.

We would like to investigate in what sense, if any, the sequence of Toeplitz matrices $\boldsymbol{\Gamma}(n)$ could be considered an approximation of $\boldsymbol{\Sigma}$ or equivalently, $\varphi_n(e^{i\theta})$ could be considered an approximation of the symbol $\varphi(e^{i\theta})$. It seems evident that this last property could only be established in a weak sense, say for arbitrary test functions $\psi(e^{i\theta})$ continuous on the unit circle one should have

$$\int \psi(e^{i\theta})\varphi_n(e^{i\theta})\psi(e^{i\theta})^* \to \int \psi(e^{i\theta})\varphi(e^{i\theta})\psi(e^{i\theta})^* \tag{34}$$

as $n \to \infty$. An equivalent question can be posed in terms of L^2 approximation of the stationary process $\mathbf{x}$ by the rank n p.d. process $\mathbf{z}$ of (31). As we shall see this problem should also be naturally formulated in a weak sense.

7. Finite rank approximation of random processes

To begin with, suppose we want to approximate a N-dimensional zero-mean random vector $\mathbf{y}$ (not to be confused with the output process of our model) having a positive definite covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$, by another N-dimensional vector say $\hat{\mathbf{y}}$ having covariance Σ_n of rank $n < N$. To make the problem well-posed we shall require that the approximation $\hat{\mathbf{y}}$ should generate a linear n-dimensional subspace of $\mathbf{H}(\mathbf{y})$ which means that $\hat{\mathbf{y}}$ can be represented as a linear function of $\mathbf{y}$, say

$$\hat{\mathbf{y}} := M\mathbf{y}$$

where $M \in \mathbb{R}^{N\times N}$ has rank n. Motivated by the discussion in the previous subsection, let us consider the following approximation problem:

Problem 3. Find a matrix $M \in \mathbb{R}^{N\times N}$ of rank n, solving the following minimum problem

$$\min_{\text{rank}(M)=n} \mathbb{E}\{\|\mathbf{y} - M\mathbf{y}\|^2\}. \tag{35}$$

Note that an equivalent geometric formulation is to look for an optimal n-dimensional subspace of $\mathbf{H}(\mathbf{y})$ onto which $\mathbf{y}$ should be projected in order to minimize the approximation error variance. Let us stress that this is quite different from the usual least squares approximation problem which amounts to projecting onto a *given subspace*.

Theorem 4. *The solution of the approximation problem (35) has the form*

$$M = U_n U_n^\top, \tag{36}$$

where U_n is a $N \times n$ matrix whose columns can be chosen as the first n normalized eigenvectors of Σ, ordered in the descending magnitude ordering of the corresponding first n eigenvalues, collected in the diagonal matrix Λ_n.

The proof is deferred to Appendix.

Observe that $\hat{\mathbf{y}} = U_n U_n^\top \mathbf{y}$ is just the first *n-Principal Components Approximation* of $\mathbf{y}$. In fact it is well-known that the PCA vector $\hat{\mathbf{y}}$ can be expressed as a linear transformation acting on $\mathbf{y}$ (Hotelling, 1936). This result confirms in particular that the truncated PCA expansion is optimal in the sense that it provides the best M and the best approximation subspace for the criterion (35). This characterization has been also found by a different technique studying subspace approximation problems; see e.g., Yang (1995).

It is immediate to check that the variance matrix of $\hat{\mathbf{y}}$ has rank n since

$$\mathbb{E}\hat{\mathbf{y}}\hat{\mathbf{y}}^\top = U_n U_n^\top \mathbb{E}\mathbf{y}\mathbf{y}^\top U_n U_n^\top = U_n \operatorname{diag}\{\lambda_1, \dots, \lambda_n\} U_n^\top \tag{37}$$

given that

$$\begin{aligned} &U_n^\top \mathbb{E}\mathbf{y}\mathbf{y}^\top U_n \\ =&U_n^\top U_N \operatorname{diag}\{\lambda_1, \dots, \lambda_n, \lambda_{n+1}, \dots, \lambda_N\} U_N^\top U_n \end{aligned} \tag{38}$$

and $U_n^\top U_N = [I_n\ O]$ picks the $n \times n$ submatrix of Λ with the first n eigenvalues. This expression holds for arbitrary N and hence for arbitrary n. Naturally one should keep in mind that the eigenvector matrices now depend on N but the eigenvalues stay fixed.

7.1. Extension to infinite dimension

In analogy to Problem 3, let us now consider the approximation of our *a posteriori* signal $\mathbf{x}$ and ask if there is a stationary process $\mathbf{z}$ spanning a subspace $\mathbf{H}(\mathbf{z}) \subset \mathbf{H}(\mathbf{x})$ of dimension n, which minimizes an approximation criterion of the type (35). If such a process exists we shall call it a *n-Principal Components (n-PC) approximation* of $\mathbf{x}$.

Let $I = [t, t+1, \dots, t+N-1]$ be a finite subinterval of the time domain $\mathbb{Z}$ of length $N \geq n$ and consider the finite random vectors $\mathbf{x}_N := [\mathbf{x}(t)\ \dots\ \mathbf{x}(t+N-1)]^\top$ and a candidate vector $\mathbf{z}_N := [\mathbf{z}(t)\ \dots\ \mathbf{z}(t+N-1)]^\top$ which is extracted from a yet undefined candidate stationary process $\mathbf{z}$. Due to stationarity the second order properties of these vectors are invariant with respect to the particular time t. We may represent $\mathbf{z}_N$ as $\mathbf{z}_N = M\mathbf{x}_N$ for some $N \times N$ matrix M of rank n so that $\mathbf{H}(\mathbf{z}_N) \subset \mathbf{H}(\mathbf{x}_N)$ has dimension n. We want to find such a $\mathbf{z}_N$ which minimizes the norm $\mathbb{E}\{\|\mathbf{x}_N - \mathbf{z}_N\|^2\}$. This minimization problem is analogous to that of Problem 3 where now $\hat{\mathbf{y}}$ is substituted by $\mathbf{z}_N$. By Theorem 4 the solution vector, say $\hat{\mathbf{x}}_N = \hat{M}\mathbf{x}_N$, is a linear function of $\mathbf{x}_N$ which must span a subspace $\mathbf{H}(\hat{\mathbf{x}}_N) \subset \mathbf{H}(\mathbf{x})$ of dimension n. Hence it determines by stationary extension (Lemma 1) a purely deterministic process $\hat{\mathbf{x}}$ such that $\mathbf{H}(\hat{\mathbf{x}}) = \mathbf{H}(\hat{\mathbf{x}}_N)$ has finite dimension n. Since $\mathbf{H}(\hat{\mathbf{x}}) = \mathbf{H}(\mathbf{x}_N) \subset \mathbf{H}(\mathbf{x})$, this process satisfies our requirements. By stationarity $\hat{\mathbf{x}}$ is invariant with respect to translations of the interval I provided its length N is fixed. Then $\hat{\mathbf{x}}$ is a n-PC approximation of $\mathbf{x}$. Below we state a key relation linking this n-PC approximation to the subspace procedure. The result is quite general and although here we shall only refer to the $\mathbf{x}$ of (3), it apparently holds for a more general class of p.d. message processes.

Theorem 5. *For $N > n$, the subspace identification algorithm based on the (true) truncated covariance Σ_N provides a n-PC approximation of $\mathbf{x}$.*

Proof. To understand in what sense the algorithm may provide an approximation of the true *a posteriori* signal $\mathbf{x}$, we shall re-examine the subspace procedure in terms of the *limit true covariances* (i.e., not in terms of their sample estimates). Consider the $N \times N$-truncation of the matrix $\boldsymbol{\Sigma}$, introduced in (11) which for each N has a positive point spectrum, say

$$\mathfrak{S}_N := \{\sigma_{N,1}, \dots, \sigma_{N,N}\}$$

where and the eigenvalues are ordered in decreasing magnitude. In our problem all Σ_N's are nonsingular so that the eigenvalues are strictly positive for all N.

An approximate model is defined by extracting a rank $n < N$ factorization of Σ_N. In terms of random variables, the subspace identification procedure starts from the spectral decomposition

$$\Sigma_N = U_N \Lambda_N U_N^\top, \qquad \Lambda_N = \operatorname{diag}\{\sigma_{N,1}, \dots, \sigma_{N,N}\}$$

and discards the eigenvalues of index larger than n to get an approximate rank n factorization

$$\hat{\Sigma}_N = U_n \Lambda_n U_n^\top, \qquad \Lambda_n = \operatorname{diag}\{\sigma_{N,1}, \dots, \sigma_{N,n}\}$$

where $U_N = \begin{bmatrix} U_n & \tilde{U}_n \end{bmatrix}$ with $U_n \in \mathbb{R}^{N \times n}$ having orthonormal columns and

$$\|\tilde{U}_n \tilde{\Lambda}_n \tilde{U}_n^\top\|_F^2 = \sum_{k=n+1}^{N} \sigma_{N,k}^2.$$

Here $\tilde{\Lambda}_n$ is the diagonal matrix with discarded eigenvalues of index larger than n and the subscript $_F$ denotes the Frobenius norm. Let now $\downarrow U_n$ be the matrix U_n deprived of its first row and $\uparrow U_n$ be the same matrix deprived of its last row. Apply the shift-invariance property

$$\uparrow U_n A = \downarrow U_n$$

and the Procrustes algorithm (Golub & van Loan, 2013) to get a unique orthogonal $n \times n$ matrix A, which leads to the Observability structure

$$U_n = \begin{bmatrix} c \\ cA \\ \vdots \\ cA^{N-1} \end{bmatrix}$$

having rank n by construction. This procedure leads to a stochastic system of the same structure as (31) where the state vector $\boldsymbol{\xi}$ has uncorrelated components of variance Λ_n. The output process has a covariance function $\gamma(t-s) = cA^t \Lambda_n A^{-s} c^\top = cA^{t-s} \Lambda_n c^\top$ since A and Λ_n commute as $A\Lambda_n A^* = \Lambda_n$ (by Lyapunov).

Let now $\mathbf{z}_N := U_n \boldsymbol{\xi}$. This random vector has covariance matrix

$$\Gamma_N := \mathbb{E}\mathbf{z}_N \mathbf{z}_N^\top = U_n \Lambda_n U_n^\top, \tag{39}$$

which is an approximate SVD factorization of rank n of Σ_N. In fact,

$$\Sigma_N = \Gamma_N + \tilde{U}_n \tilde{\Lambda}_n \tilde{U}_n^\top, \tag{40}$$

Therefore in view of the identity of the expressions (39) and (37), $\mathbf{z}_N$ has the same covariance structure of $\hat{\mathbf{x}}_N$ and can be considered an equivalent n-PC approximation of $\mathbf{x}_N$. □

The question now is to understand in what sense this approximation can get close to $\mathbf{x}$ as $n \to \infty$. Since we are now studying the behavior of the n-PC approximation of $\mathbf{x}$ when the dimension n varies, we shall attach a superscript to $\hat{\mathbf{x}}$ and denote it by $\hat{\mathbf{x}}^n$; likewise we shall do for its covariance matrix, which will be denoted $\boldsymbol{\Gamma}(n)$. From what we have seen so far, for each n there is a n-PC approximation of $\mathbf{x}$ having a well defined (infinite) covariance matrix $\boldsymbol{\Gamma}(n)$ of rank n.

Theorem 6. *The sequence $\{\boldsymbol{\Gamma}(n)\}$ converges weakly to $\boldsymbol{\Sigma}$ as n diverges to ∞, that is*

$$\psi^\top [\boldsymbol{\Sigma} - \boldsymbol{\Gamma}(n)] \psi \to 0 \qquad as \quad n \to \infty$$

for all functions (infinite row sequences) $\psi^\top$ having finite support in an interval $I \subset \mathbb{Z}$.

Proof. Consider the $N \times N$ truncation (with $N \geq n$) of the (infinite) covariance matrix $\boldsymbol{\Gamma}(n)$ of the n-PC approximation process $\hat{\mathbf{x}}^n$ of $\mathbf{x}$ and denote it $\boldsymbol{\Gamma}_N(n)$. By analogy to (37) this $N \times N$ matrix is the rank n SVD approximation of Σ_N and has the structure

$$\Gamma_N(n) = U_N^n \operatorname{diag}\{\sigma_{N,1}, \dots, \sigma_{N,n}\} (U_N^n)^\top \tag{41}$$

where $\sigma_{N,k}$ are the first n eigenvalues of the $N \times N$-truncation of $\boldsymbol{\Sigma}$. These eigenvalues are well-defined for all finite N. The $N \times n$ eigenvector matrices U_N^n depend on N and their dimension obviously grows with N.

By a well-known property of the SVD, see e.g., Golub and van Loan (2013, Chap. 2), the variance matrix $\Gamma_N(n)$ of $\hat{\mathbf{x}}_N^n$ is the symmetric $N \times N$ matrix of rank n which has minimum distance from that of $\mathbf{x}_N$ in the Frobenius norm. This in turn implies that the (infinite) covariance matrix $\boldsymbol{\Gamma}(n)$ is the symmetric positive operator of rank n for which

$$\begin{aligned}\psi^\top(\boldsymbol{\Sigma} - \boldsymbol{\Gamma}(n))\psi &= \psi_N^\top(\Sigma_N - \Gamma_N(n))\psi_N \\ &\le \|\psi_N\|^2 \sum_{k=n+1}^{N} \sigma_{N,k}\end{aligned} \tag{42}$$

for all functions ψ having support in an interval $I \subset \mathbb{Z}$ of length $N \ge n$. We shall let $n \to \infty$ while still keeping $N \ge n$ (so obviously we have $N \to \infty$ as well).

Recall now that $\boldsymbol{\Sigma}$ is a bounded linear operator in ℓ^2 and let $m(\varphi)$ be the essential infimum of its spectral density $\varphi(e^{i\theta})$. Since for our model this function is zero on a set of positive Lebesgue measure we have $m(\varphi) = 0$. We quote the following fundamental fact from Grenander and Szegő (1958, p. 65) and Gray (2006, Corollary 6, p. 58).

Theorem 7 (*Szegő*). *Let $\varphi(e^{i\theta})$ be the symbol of the infinite covariance matrix $\boldsymbol{\Sigma}$, then*

$$\lim_{N\to\infty} \min_k \sigma_{N,k} = m(\varphi). \tag{43}$$

Therefore if the eigenvalues are listed in descending order, one has

$$\lim_{N\to\infty} \sigma_{N,N} = 0\,. \tag{44}$$

Since $\sigma_{N,N}$ tends to zero by Theorem 7, so does any finite sequence $\{\sigma_{N,k};\ n+1 \le k \le N\}$ and likewise should do their sum. Therefore the first member in (42) converges to zero for all ψ as $N \to \infty$. □

Remark 8. Contrary to all submatrices Σ_N, the infinite covariance operator $\boldsymbol{\Sigma}$ may not have eigenvalues (nor corresponding eigenvectors) and consequently the idea of PC approximation does not apply directly to the full (infinite) matrix. For this reason the approximation and the convergence results may not hold in a strong sense.

8. Approximation in the spectral domain

We have shown that the theoretical subspace algorithm described in Section 7.1 provides a n-PC approximation converging weakly to the "true" process $\mathbf{x}$. Next we want to study this approximation in the spectral domain.

From Lindquist and Picci (2015, Chap. 3), the processes $\mathbf{x}$ and $\hat{\mathbf{x}}^n$ have a spectral representation with random spectral measures $dZ(e^{i\theta})$ and $dZ_n(e^{i\theta})$ such that

$$\begin{aligned}\mathbb{E}dZ(e^{i\theta})dZ(e^{i\theta})^* &= \varphi(e^{i\theta})\frac{d\theta}{2\pi}, \\ \mathbb{E}dZ_n(e^{i\theta})dZ_n(e^{i\theta})^* &= \varphi_n(e^{i\theta})\frac{d\theta}{2\pi}\,.\end{aligned}$$

where the second expression is formally written as a *bona fide* spectral density in spite of the fact that $\varphi_n(e^{i\theta})$ is a sum of Dirac delta functions. Letting $\hat{\psi}(e^{i\theta}) := \sum_{k=0}^{N-1} \psi(k)e^{i\theta k}$, one has

$$\begin{aligned}\psi^\top \boldsymbol{\Sigma}\psi = \mathbb{E}\left[\sum_{k=0}^{N-1}\psi(k)\mathbf{x}(k)\right]^2 &= \mathbb{E}\left[\int_{-\pi}^{\pi}\hat{\psi}(e^{i\theta})dZ(e^{i\theta})\right]^2 \\ &= \int_{-\pi}^{\pi}\hat{\psi}(e^{i\theta})\varphi(e^{i\theta})\hat{\psi}(e^{i\theta})^*\frac{d\theta}{2\pi},\end{aligned}$$

and similarly for $\psi^\top\boldsymbol{\Gamma}(n)\psi$ we have

$$\begin{aligned}\psi^\top\boldsymbol{\Gamma}(n)\psi = \mathbb{E}\left[\sum_{k=0}^{N-1}\psi(k)\hat{\mathbf{x}}(k)\right]^2 &= \mathbb{E}\left[\int_{-\pi}^{\pi}\hat{\psi}(e^{i\theta})dZ_n(e^{i\theta})\right]^2 \\ &= \int_{-\pi}^{\pi}\hat{\psi}(e^{i\theta})\varphi_n(e^{i\theta})\hat{\psi}(e^{i\theta})^*\frac{d\theta}{2\pi},\end{aligned}$$

so that

$$\psi^\top[\boldsymbol{\Gamma}(n) - \boldsymbol{\Sigma}]\psi = \int_{-\pi}^{\pi}\hat{\psi}(e^{i\theta})[\varphi_n(e^{i\theta}) - \varphi(e^{i\theta})]\hat{\psi}(e^{i\theta})^*\frac{d\theta}{2\pi}$$

and (34) follows from Theorem 6. □

Note also that, because of the orthogonality $\psi^\top(\mathbf{x} - \hat{\mathbf{x}}^n) \perp \psi^\top\hat{\mathbf{x}}^n$ (which follows from the relation (A.1)), the difference (42) can be rewritten as $\mathbb{E}[\psi^\top(\mathbf{x} - \hat{\mathbf{x}}^n)]^2$ which also must tend to zero when $n \to \infty$ for all functions ψ having support in an interval $I \subset \mathbb{Z}$ of length $N \ge n$. Therefore we can say that $\hat{\mathbf{x}}^n$ *converges weakly to* $\mathbf{x}$. In particular, we can take the support of ψ to be a single point $t \in \mathbb{Z}$, and the above result implies that $\hat{\mathbf{x}}^n(t) \to \mathbf{x}(t)$ in mean square for all $t \in \mathbb{Z}$.

Remark 9. As the reader may have noticed, this proof of consistency uses only certain special properties of the random-frequency model (3), namely the fact that the spectral density of $\mathbf{x}$, $\varphi(e^{i\theta})$, is a bounded function which is zero on a set of positive Lebesgue measure. In this sense the argument does apply to a larger class of p.d. message signals (and to a larger class of positive definite infinite Toeplitz matrices). An analysis of this class of signals goes however beyond the scope of this paper.

9. Conclusions

We have studied consistency of an approximate subspace identification procedure for the infinite dimensional *a posteriori* model of a frequency estimation problem in an Empirical Bayesian formulation. By first imposing a natural uniform prior probability density on the unknown frequencies, the estimation of the hyperparameters of the *a priori* distribution can be accomplished by a sequence of subspace identification techniques. The key idea is to exploit the special structure of the covariance matrix of the *a posteriori* process which has been discovered by exploring classical results on energy concentration in deterministic signal processing. The convergence of the spectrum of the subspace estimates to the (nonrational) spectral density of the *a posteriori* process is proven by a weak-sense approximation theorem for Toeplitz spectra.

Appendix. Proof of Theorem 4

As usual, minimizing the square distance in (35) requires that the approximation $M\mathbf{y}$ should be uncorrelated with the approximation error, namely

$$\mathbf{y} - M\mathbf{y} \perp M\mathbf{y}, \tag{A.1}$$

which is equivalent to

$$M\Sigma - M\Sigma M^\top = 0.$$

Introducing a square root $\Sigma^{1/2}$ of Σ and defining $\hat{M} := \Sigma^{-1/2}M\Sigma^{1/2}$, this condition is seen to be equivalent to

$$\hat{M} = \hat{M}\hat{M}^\top$$

which just says that $\hat{M}$ must be *symmetric and idempotent* (i.e. $\hat{M} = \hat{M}^2$), in other words an *orthogonal projection* from $\mathbb{R}^N$ onto some n-dimensional subspace. Hence M must have the following structure

$$M = \Sigma^{1/2}\Pi\,\Sigma^{-1/2}, \quad \Pi = \Pi^2, \quad \Pi = \Pi^\top \tag{A.2}$$

where Π is an orthogonal projection matrix of rank n. Let $\Lambda := \mathrm{diag}\{\lambda_1, \dots, \lambda_N\}$ and $\Sigma = U\Lambda U^\top$ be the spectral decomposition of Σ in which U is an orthogonal matrix of eigenvectors. We can, for example, pick as a square root of Σ the matrix $\Sigma^{1/2} := U\Lambda^{1/2}$.

Now, no matter how $\Sigma^{1/2}$ is chosen, the random vector $\mathbf{e} := \Sigma^{-1/2}\mathbf{y}$ has orthonormal components. Hence using (A.2) the cost function of our minimum problem can be rewritten as

$$\begin{aligned}&\mathbb{E}\{\|\mathbf{y} - M\mathbf{y}\|^2\} = \mathbb{E}\{\|\Sigma^{1/2}\mathbf{e} - \Sigma^{1/2}\Pi\,\Sigma^{-1/2}\mathbf{y}\|^2\}\\ =&\mathbb{E}\{\|\Sigma^{1/2}(\mathbf{e} - \Pi\,\mathbf{e})\|^2\} = \mathbb{E}\{\|\Lambda^{1/2}(\mathbf{e} - \Pi\,\mathbf{e})\|^2\}\\ =&\mathbb{E}\,(\mathbf{e} - \Pi\,\mathbf{e})^\top \Lambda\,(\mathbf{e} - \Pi\,\mathbf{e})\\ =&\,\mathrm{tr}\left[\Lambda\,\mathbb{E}(\mathbf{e} - \Pi\,\mathbf{e})(\mathbf{e} - \Pi\,\mathbf{e})^\top\right]\end{aligned} \tag{A.3}$$

where $\mathrm{tr}\,A := \sum a_{kk}$ is the trace of A. Our minimum problem can therefore be rewritten as

$$\min_{\mathrm{rank}(\Pi)=n} \mathrm{tr}\{\Lambda\Pi^\perp\}$$

where $\Pi^\perp := I - \Pi$ is the orthogonal projection matrix onto the orthogonal complement of the subspace $\mathrm{Im}\,\Pi$.

Since the eigenvalues are in a decreasing order, i.e., $\lambda_1 \geq \cdots \geq \lambda_N$, one sees that the minimum of this function of Π is reached when Π projects onto the subspace spanned by the first n coordinate axes. In other words, $\Pi_{\mathrm{optimal}} = \mathrm{diag}\{I_n,\ O\}$ the minimum being $\lambda_{n+1} + \cdots + \lambda_N$. It is then evident that

$$M = U\Lambda^{1/2}\,\Pi_{\mathrm{optimal}}\,\Lambda^{-1/2}U^\top = U_n U_n^\top$$

Naturally, multiplying U_n by any $n \times n$ orthogonal matrix does not change the result. □

References

Akhiezer, N., & Glazman, I. M. (1961). *Theory of Linear Operators in Hilbert Space Vol I.* New York: Fredrik Ungar Pub. Co..

Aravkin, A., Burke, J. V., Chiuso, A., & Pillonetto, G. (2012). On the estimation of hyperparameters for empirical Bayes estimators: Maximum marginal likelihood vs minimum MSE. In *Proc. of the 16th IFAC Symposium on System Identification* (pp. 125–132). Brussels, Belgium.

Chiuso, A. (2016). Regularization and Bayesian learning in dynamical systems: Past, present and future. *Annual Reviews in Control*, *41*, 24–38.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge U.P..

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science*, *29*(2), 285–301.

Favaro, M., & Picci, G. (2012). Consistency of subspace methods for signals with almost-periodic components. *Automatica*, *48*(3), 514–520.

Favaro, M., & Picci, G. (2015). A Bayesian non-parametric approach to frequency estimation. In *Proc. SYSID 2015, vol. 48, no. 28* (pp. 478–483). IFAC-Papers On Line.

Golub, G., & van Loan, C. (2013). *Matrix Computations* (4th ed.). Baltimore: Johns Hopkins University Press.

Gray, R. M. (2006). Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, *2*(3), 155–239.

Grenander, U., & Szegő, G. (1958). *Toeplitz Forms and Their Applications*. University of California Press.

Hannan, E. J., & Deistler, M. (1988). *The Statistical Theory of Linear Systems*. New York: John Wiley.

Hartman, P., & Wintner, A. (1954). The spectra of Toeplitz matrices. *American Journal of Mathematics*, *76*, 867–882.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–377.

Hsiao, C. (2014). *Econometric Society Monographs*, *Analysis of Panel Data* (3rd ed.). (54), Cambridge University Press.

Khare, K. (2006). Bandpass sampling and bandpass analogues of prolate spheroidal functions. *Signal Processing*, *86*(7), 1550–1558.

Landau, H. J., & Pollak, H. (1961). Prolate spheroidal wave functions, Fourier analysis and uncertainty II. *The Bell System Technical Journal*, *40*, 65–84.

Landau, H. J., & Widom, H. (1980). Eigenvalue distribution of time and frequency limiting. *Journal of Mathematical Analysis and Applications*, *77*(2), 469–481.

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

Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer Texts in Statistics.

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

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, *4*, 415–447.

Picci, G., & Zhu, B. (2019). Bayesian frequency estimation. In *The 17th European Control Conference* (pp. 848–853). IEEE.

Picci, G., & Zhu, B. (2020). An empirical Bayes approach to frequency estimation. arXiv preprint: https://arxiv.org/abs/1910.09475.

Picci, G., & Zhu, B. (2021). Bayesian frequency estimation on narrow bands. In *Proc. 2021 SYSID Symposium, IFAC-papersonline, vol. 54, no. 7* (pp. 108–113). Padova, Italy.

Reinsel, G. C. (1985). Mean squared error properties of empirical Bayes estimators in a multivariate random effects general linear model. *Journal of the American Statistical Association*, *80*(391), 642–650.

Slepian, D. (1978). Prolate spheroidal wave functions, Fourier analysis and uncertainty V: The discrete case. *The Bell System Technical Journal*, *57*(5), 1371–1430.

Slepian, D., & Pollak, H. O. (1961). Prolate spheroidal wave functions, Fourier analysis and uncertainty–I. *Bell Labs Technical Journal*, *40*(1), 43–63.

Söderström, T., & Stoica, P. (1989). *System Identification*. New York: Prentice Hall.

Stoica, P., & Moses, R. L. (2005). *Spectral Analysis of Signals*. New Jersey: Prentice-Hall.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, *70*(9), 1055–1096.

Yang, B. (1995). Projection approximation subspace tracking. *IEEE Transactions on Signal Procesing*, *43*, 95–107.

Yuan, M., Wan, C., & Wei, L. (2016). Superiority of empirical Bayes estimator of the mean vector in multivariate normal distribution. *Science China Mathematics*, *59*, 1175–1186.

Zacharias, D., Wirfält, P., Jansson, M., & Chatterjee, S. (2013). Line spectrum estimation with probabilistic priors. *Signal Processing*, *93*(11), 2969–2974.

Giorgio Picci received the Dr.Eng. degree from the University of Padua, Padua, Italy, in 1967. Currently, he is Professor Emeritus with the Department of Information Engineering, University of Padua, Padua, Italy. He has held several long-term visiting appointments with various American, European, Japanese, and Chinese universities among which Brown University, MIT, the University of Kentucky, Arizona State University, the Royal Institute of Technology, Stockholm, Sweden, Kyoto University, and Washington University, St. Louis, MO, USA. He has been contributing to systems and control mostly in the area of modeling, estimation, and identification of stochastic systems and published over 150 papers and edited three books in this area. He has been involved in various joint research projects with industry and state agencies.

Dr. Picci is a life Fellow of IEEE, a Fellow of IFAC and a foreign member of the Swedish Royal Academy of Engineering Sciences. He has been chairman of the IFAC Technical Committee on Stochastic Systems, past member of the EUCA council, project manager of the Italian team for the Commission of the European Communities Network of Excellence System Identification (ERNSI) and general coordinator of the Commission of European Communities IST project RECSYS, in the fifth Framework Program.

Bin Zhu was born in Changshu, Jiangsu, China in 1991. He received the B.Eng. degree from Xi'an Jiaotong University, Xi'an, China in 2012 and the M.Eng. degree from Shanghai Jiao Tong University, Shanghai, China in 2015, both in control science and engineering. In 2019, he obtained a Ph.D. degree in information engineering from University of Padova, Padova, Italy, and he had a one-year postdoc position in the same university. Since December 2019, he has been working at the School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, China, as an assistant professor.

His current research interest includes spectral analysis, frequency estimation, and sparsity-promoting techniques for signal processing and machine learning.