Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-7 (2021) 108–113

Bayesian Frequency Estimation on Narrow Bands ⋆

Giorgio Picci * **Bin Zhu** **

* *Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131 Padova, Italy (e-mail: picci@dei.unipd.it)*
** *School of Intelligent Systems Engineering, Sun Yat-sen University, Waihuan East Road 132, 510006 Guangzhou, China (e-mail: zhub26@mail.sysu.edu.cn)*

Abstract:
In some recent works, the authors have proposed and developed an Empirical Bayes framework for frequency estimation. The unknown frequencies in a noisy oscillatory signal are modeled as uniform random variables supported on narrow frequency bands. The bandwidth and the relative band centers are known as hyperparameters which can be efficiently estimated using techniques from subspace identification. In the current paper, we examine carefully how the estimated frequency prior can be used to produce a Bayesian estimate of the unknown frequencies based on the same data (for hyperparameter estimation). To this end, we formulate the Bayesian Maximum A Posteriori (MAP) optimization problem and propose an iterative algorithm to compute its solution. Then, we do extensive simulations under various parameter configurations, showing that the MAP estimate of the frequencies are asymptotically close to the band centers of the frequency priors. These results provide an attractive link between the conventional Bayesian method and the Empirical Bayes method for frequency estimation, and in retrospect justify the use of the latter.

Keywords: Frequency estimation, multiband random signal, subspace identification, Bayesian MAP estimate.

1. INTRODUCTION

In this paper we study the identification of a class of quasi-periodic processes composed of random oscillations observed in additive noise. Data are modeled as trajectories of a process whose frequencies may deviate slightly about an unknown nominal value. In a Bayesian setting, it is then reasonable to model these frequencies as random variables which are noisy versions of some nominal frequency. Although estimation of quasi-periodic signals, often referred simply as *frequency estimation*, is a classical problem in signal processing with an enormous literature, an approach based on *Empirical Bayes* estimation proposed in this paper, following previous work in Picci and Zhu (2020), seems to have not been fully pursued in the literature. For a survey and some further bibliography on Empirical Bayes methods, we refer to Lehmann and Casella (1998, p. 262); Efron (2010, 2014); Chiuso (2015); Petrone et al. (2014); Aravkin et al. (2012). A general underlying motivation for the Empirical Bayes approach is that in some cases it has been proven to yield a mean squared error (MSE) which can even be smaller than that of maximum likelihood (Reinsel, 1985; Yuan et al., 2016).

The underlying model is a variant of the classical white-noise-corrupted sum of harmonic oscillations in that the hidden frequencies are modeled as randomly varying parameters. When frequencies are deterministic, such quasi-periodic stationary processes are well-known to be purely deterministic signals. This character persists when, as in the Bayesian setting of this paper, the frequencies are random but described by a natural class of a priori distributions. The prior for each hidden frequency is chosen as a parametric uniform distribution on a small frequency range of unknown width, which can be interpreted as an a priori confidence interval centered about some unknown nominal frequency. The width and the relative center frequencies are the *hyperparameters* of the prior which are estimated from data. This simple model seems to be reasonable for a variety of applications.

In this frame, we have shown that the estimation of the hyperparameters can be approached by a simple efficient *subspace algorithm* (Favaro and Picci, 2015; Picci and Zhu, 2019, 2020). This is in contrast with the standard marginal likelihood approach as considered for example in Lázaro-Gredilla et al. (2010); Aravkin et al. (2012). Our approach fully uses the special structure of the data process and need not involve numerical optimization. Such a subspace algorithm to estimate the prior is referred to as the first step in our Bayesian procedure for frequency estimation.

Assuming that the unknown hyperparameter for the center frequency has a true value, one can then prove consistency of the subspace estimation method which justifies our procedure in the framework of the traditional frequentist interpretation of the hyperparameter. In this paper, we shall pursue in more detail the second step of the (Empir-

⋆ B. Zhu was supported by the "Hundred-Talent Program" of the Sun Yat-sen University.

10.1016/j.ifacol.2021.08.343

ical) Bayes procedure, that is, the *Maximum A Posteriori* (MAP) frequency estimate based on the estimated prior. Our goal is to verify that the MAP estimate is asymptotically close to (and ideally coincides with) the true center frequency which is, on the other hand, consistently estimated by the subspace algorithm (Picci and Zhu, 2021).

This paper is organized as follows. In Sec. 2, we review the Empirical Bayes framework for frequency estimation which has emerged in recent literature. The structure of the signal covariance matrix is briefly described and its use for the hyperparameter estimation is also discussed. Then, Sec. 3 addresses the MAP Bayesian estimator of the random frequency based on the estimated prior. A linearization strategy to solve the resulting optimization problem is proposed. In the following Sec. 4, the Bayesian frequency estimator is tested via simulations under different data sizes and a wide range of signal-to-noise ratios. At last, Sec. 5 concludes the paper.

2. SIGNAL MODEL

In this paper, bold lowercase symbols like $\mathbf{x}$ denote random quantities, while italic lowercase letters like $\boldsymbol{x}$ denote vectors with numerical entries.

Consider the following noisy measurements of a quasi-periodic signal

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{w}(t), \quad t \in \mathbb{Z} \tag{1}$$

where t represents time and

$$\mathbf{x}(t) := \sum_{\ell=1}^{\nu} \mathbf{a}_\ell \cos(\boldsymbol{\omega}_\ell t) + \mathbf{b}_\ell \sin(\boldsymbol{\omega}_\ell t) \tag{2}$$

is the random oscillatory signal, and $\mathbf{w}$ is additive white noise. The angular frequencies $\boldsymbol{\omega}_\ell$ are unknown but their number ν is assumed to be given. In addition, we make the following assumptions:

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

Since we are considering only *real* processes, the spectral content of the processes can be thought of as concentrated on the half interval $[0, \pi]$. Hence, it does not cause any loss of generality to assume that each random frequency $\boldsymbol{\omega}_\ell$ takes value in $[0, \pi]$. We shall let $\boldsymbol{\omega} := [\boldsymbol{\omega}_1 \ \dots \ \boldsymbol{\omega}_\nu]^\top$ and denote by $\mathbf{a}$, $\mathbf{b}$ two similarly arranged amplitude vectors. Given the frequencies $\boldsymbol{\omega}$, estimation of the amplitudes $\mathbf{a}$, $\mathbf{b}$ is just a standard linear estimation problem. Therefore, we shall mainly concentrate in the problem of frequency estimation.

We impose that each component $\boldsymbol{\omega}_\ell$ of the random vector $\boldsymbol{\omega}$ follows a uniform distribution on the frequency band $[\theta_\ell - W_\ell, \theta_\ell + W_\ell]$ such that the symmetrized sets w.r.t. the origin

$$S_\ell := [\theta_\ell - W_\ell, \theta_\ell + W_\ell] \cup [-\theta_\ell - W_\ell, -\theta_\ell + W_\ell], \quad \ell = 1, \dots, \nu \tag{3}$$

do not overlap. Here $0 \le \theta_\ell \le \pi$ is called a *center frequency* and $0 \le W_\ell < |\theta_\ell|$ a *bandwidth*. Note that the inequality for W_ℓ implies that the two intervals composing S_ℓ in (3) do not intersect. We can collect the center frequencies into a vector $\boldsymbol{\theta} := [\theta_1, \dots, \theta_\nu]^\top$. For technical reasons, we further assume that the bandwidths are identical for different components, i.e., $W_1 = \cdots = W_\nu = W$. In the literature, both $\boldsymbol{\theta}$ and W are called *hyperparameters* of the *a priori* distribution for the frequency vector $\boldsymbol{\omega}$.

It turns out that under the model assumptions above, both processes $\mathbf{y}$ and $\mathbf{x}$ are stationary. In fact, let us define the covariance functions

$$\Sigma(t, s) := \mathbb{E}\{\mathbf{y}(t)\mathbf{y}(s)\}, \quad K(t, s) := \mathbb{E}\{\mathbf{x}(t)\mathbf{x}(s)\}. \tag{4}$$

Then after some standard calculations, we have

$$\begin{aligned} \Sigma(t, s) &= K(t, s) + \sigma_{\mathbf{w}}^2 \delta(t, s) \\ &= \sum_{\ell=1}^{\nu} \sigma_\ell^2 \, \mathbb{E}(\cos \boldsymbol{\omega}_\ell \tau) + \sigma_{\mathbf{w}}^2 \delta(\tau, 0) \\ &= \frac{\sin W\tau}{W\tau} \sum_{\ell=1}^{\nu} \sigma_\ell^2 \cos \theta_\ell \tau + \sigma_{\mathbf{w}}^2 \delta(\tau, 0) \\ &= \frac{\pi}{2W} \int_{-\pi}^{\pi} e^{i\omega\tau} \sum_{\ell=1}^{\nu} \sigma_\ell^2 \chi_{S_\ell}(\omega) \frac{\mathrm{d}\omega}{2\pi} + \sigma_{\mathbf{w}}^2 \delta(\tau, 0), \end{aligned} \tag{5}$$

where the new variable $\tau := t - s$, $\delta(t, s)$ is the Kronecker symbol, and χ_{S_ℓ} is the indicator function on the set S_ℓ given in (3), namely

$$\chi_{S_\ell}(\omega) = \begin{cases} 1 \text{ for } \omega \in S_\ell, \\ 0 \text{ for } \omega \in [-\pi, \pi] \setminus S_\ell. \end{cases} \tag{6}$$

Since all the covariances depend only on the time difference τ, stationarity indeed follows. In the following, we will write $\Sigma(\tau)$ and $K(\tau)$ in place of $\Sigma(t, s)$ and $K(t, s)$, respectively. Notice that by the last equality in (5), we can write the covariance of $\mathbf{x}$ as the Fourier coefficient of the function

$$\Phi_{\mathbf{x}}(\omega) := \frac{\pi}{2W} \sum_{\ell=1}^{\nu} \sigma_\ell^2 \, \chi_{S_\ell}(\omega). \tag{7}$$

The latter is a weighted sum of indicator functions and is known as the *spectral density* of the process $\mathbf{x}$. Since the indicator function (6) can be viewed as the spectrum of a *bandlimited white noise* within the frequency band S_ℓ, the process $\mathbf{x}$ can be thought of as a sum of independent bandlimited noises, each having variance σ_ℓ^2.[1] Given the assumption that the union $\cup_{\ell=1}^{\nu} S_\ell$ is a strict subset of $[-\pi, \pi]$, we can conclude that $\mathbf{x}$ 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).

2.1 Properties of the Covariance Matrix

In practice, we often observe sample paths of finite lengths from the process $\mathbf{y}$. With an abuse of notation, it is convenient to collect the observed random variables into a column vector

$$\mathbf{y}_N := [\mathbf{y}(t), \mathbf{y}(t+1), \dots, \mathbf{y}(t+N-1)]^\top. \tag{8}$$

[1] Our setting here is somehow reminiscent of Multiple Kernel methods as in Hoffmann et al. (2008); Bach et al. (2004).

Consider then the $N \times N$ covariance matrix $\mathbf{\Sigma}_N := \mathbb{E}\{\mathbf{y}_N \mathbf{y}_N^\top\}$. Due to the independence of the noise to the signal, we have the additive structure

$$\mathbf{\Sigma}_N = \mathbf{K}_N + \sigma_{\mathbf{w}}^2 I_N, \tag{9}$$

where $\mathbf{K}_N$ is defined as the covariance matrix of the signal string

$$\mathbf{x}_N := [\mathbf{x}(t), \mathbf{x}(t+1), \dots, \mathbf{x}(t+N-1)]^\top, \tag{10}$$

namely

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

which is symmetric Toeplitz as a consequence of stationarity.

As shown in the papers Favaro and Picci (2015); Picci and Zhu (2019, 2020), the eigenvalues of $\mathbf{K}_N$ (arranged in nonincreasing order) initially remain constant, and then starting from a certain index (known as the *Slepian frequency* in the theory of Prolate Spheroidal Wave Functions), decay extremely fast to zero. This property proves to be important for the estimation of the hyperparameters $(\boldsymbol{\theta}, W)$ from the noisy measurements. To avoid repetition, the reader may consult the aforementioned papers for some eigenplots of $\mathbf{K}_N$ which resemble the shape of a step. Next, we present a formal argument for such decay of eigenvalues, alternative to the proof given in Picci and Zhu (2020), yet more general since now we have allowed the signal variances σ_ℓ^2 to be different.

Proposition 1. The covariance matrix $\mathbf{K}_N$ asymptotically, i.e., as $N \to \infty$, has rank equal to $2\nu W N/\pi$.

Proof. The spectral density $\Phi_{\mathbf{x}}(\omega)$ in (7) of the signal process $\mathbf{x}$ has support in the union $S := \cup_{\ell=1}^{\nu} S_\ell$. By a famous theorem of Szegö for the eigenvalue distribution of Toeplitz matrices (Grenander and Szegö (1958), see also e.g., Gray (2006)), we have the relation

$$\lim_{N\to\infty} \frac{\operatorname{rank}(\mathbf{K}_N)}{N} = \frac{m(S)}{2\pi} = \frac{4\nu W}{2\pi}$$

where $m(\cdot)$ denotes the Lebesgue measure of a set. In other words, the fraction of positive eigenvalues of $\mathbf{K}_N$ in N is asymptotically equal to the fraction of the spectral support of the signal in the whole frequency domain. The assertion of the proposition then follows immediately. □

2.2 Hyperparameter Estimation

From Proposition 1, we see that the bandwidth W can be directly recovered from the rank of $\mathbf{K}_N$ when the matrix size N is sufficiently large. The band centers $\boldsymbol{\theta}$, on the other hand, can be reliably estimated using a subspace method. We quote the following result from Picci and Zhu (2020) without a proof.

Proposition 2. For N large enough, let $n \approx 2\nu W N/\pi$. Then there is an $n \times n$ orthogonal matrix A and an n-dimensional row vector c such that the random oscillatory signal $\mathbf{x}$ in (2) can be represented by the system

$$\boldsymbol{\xi}(t+1) = A\boldsymbol{\xi}(t) \tag{12a}$$

$$\mathbf{x}(t) = c\,\boldsymbol{\xi}(t) \tag{12b}$$

where $\boldsymbol{\xi}(t) = [\,\xi_1(t),\, \xi_2(t),\, \dots, \xi_n(t)\,]^\top$ is an n-dimensional basis vector spanning the Hilbert space $\mathbf{H}(\mathbf{x})$ linearly generated by the N random variables of the set $\{\mathbf{x}(s) : t \geq s \geq t-N+1\}$.

This representation suggest the use of a *subspace algorithm* to estimate the center-frequency vector $\boldsymbol{\theta}$. To implement the procedure we start from an estimate of the covariance matrix $\mathbf{K}_N$ extracted from the signal covariance $\mathbf{\Sigma}_N$ in (9). This last estimate is computed given multiple snapshots of measurements (1). Notice that multiple snapshots are necessary due to the fact that sinusoidal signals with random frequencies are not ergodic. Then, a state-space realization (12) is estimated from $\mathbf{K}_N$, and the band centers can be recovered from the phase angles of the eigenvalues of A which have unit moduli. A detailed discussion can be found in Picci and Zhu (2020) where some simulation results are also reported, showing good performance of the subspace estimator. It can actually be shown that the procedure is statistically consistent in the sense that for $N \to \infty$ the estimated finite rank purely deterministic approximate process (12) converges to the a posteriori process (2). We refer to Picci and Zhu (2021) for technical details

3. BAYESIAN ESTIMATION

Assume now that we have a consistent estimate of the parameters of the prior, in particular of the center frequencies $\{\theta_\ell\}$. The question is what this estimate has to do with (say) the Bayesian MAP estimate [2] of the random angular frequency $\boldsymbol{\omega}$, computed from the relative posterior distribution. Is there any reason why the MAP estimate should coincide, at least asymptotically, with the center frequencies of the prior?

In Sec. 4 we shall provide experimental evidence that in our setting the inherent optimization problem leads to a MAP estimate of $\boldsymbol{\omega}$ which is practically indistinguishable from the subspace estimate of the center frequency $\boldsymbol{\theta}$. This fact is verified experimentally but should be also evident from the theoretical analysis which follows. For the purpose of simplifying the presentation, we only describe the procedure in the case of one snapshot of data.

The MAP estimator of $\boldsymbol{\omega}$ is obtained by maximizing the log of the unnormalized posterior distribution of $\boldsymbol{\omega}$ given N observations [3] $\boldsymbol{y} := [y(t) \dots y(t-N)]^\top$, neglecting the denominator $p(y)$ which does not depend on the parameters. The prior for one frequency is

$$p(\omega_\ell \mid \theta_\ell, W) = \frac{1}{2W}\, \chi_{[\theta_\ell - W,\, \theta_\ell + W]}$$

and since the intervals do not overlap we have independence and the overall prior of $\boldsymbol{\omega}$ is the product of the priors for each $\boldsymbol{\omega}_\ell$ so that, recalling that the noise is Gaussian i.i.d., we have

$$\hat{\boldsymbol{\omega}}^{\mathrm{MAP}} = \underset{\omega\in[0,\pi]^\nu}{\operatorname{argmax}} \left\{ -\frac{1}{2\sigma_{\mathbf{w}}^2} \|\boldsymbol{y} - V(\boldsymbol{\omega})\boldsymbol{u}\|^2 + \sum_\ell \log p(\omega_\ell \mid \theta_\ell, W) \right\}$$

[2] MAP is known to be the best estimate in a variety of norms.

[3] The estimation from multiple snapshots data can be dealt with in a similar way even in case of unequal measurement error variances.

with $V(\boldsymbol{\omega}) = [C(\boldsymbol{\omega})\ S(\boldsymbol{\omega})]$ where

$$C(\boldsymbol{\omega}) = \begin{bmatrix} \cos\omega_1 & \dots & \cos\omega_\nu \\ \vdots & \ddots & \vdots \\ \cos\omega_1 N & \dots & \cos\omega_\nu N \end{bmatrix} := [\boldsymbol{c}(\omega_1)\ \dots\ \boldsymbol{c}(\omega_\nu)]$$

$$S(\boldsymbol{\omega}) = \begin{bmatrix} \sin\omega_1 & \dots & \sin\omega_\nu \\ \vdots & \ddots & \vdots \\ \sin\omega_1 N & \dots & \sin\omega_\nu N \end{bmatrix} := [\boldsymbol{s}(\omega_1)\ \dots\ \boldsymbol{s}(\omega_\nu)]$$

and $\boldsymbol{u} = [a_1\ \dots\ a_\nu\ b_1\ \dots\ b_\nu]^\top := [\boldsymbol{a}\ \boldsymbol{b}]^\top$.

Now for each $\ell = 1, \dots, n$, the log of the prior for ω_ℓ is $-\infty$ outside the interval $J_\ell := [\theta_\ell - W,\ \theta_\ell + W]$ and equal to $\log \frac{1}{(2W)^\nu}$ inside. Hence the MAP estimator of $\boldsymbol{\omega}$ can be found by solving the constrained minimization problem

$$\hat{\boldsymbol{\omega}}^{\mathrm{MAP}} = \underset{\omega}{\operatorname{argmin}} \left\{ \frac{1}{2\sigma_{\mathbf{w}}^2} \|\boldsymbol{y} - V(\boldsymbol{\omega})\boldsymbol{u}\|^2 + \nu \log(2W) \right\}$$
$$\text{subject to}: \omega_\ell \in J_\ell, \quad l = 1, \dots, \nu. \tag{13}$$

Suppose that $\hat{\boldsymbol{\theta}}$ and $\hat{W}$ are our subspace estimates of the hyperparameters of the prior. Since these are consistent as discussed in the previous section, substituting these estimates for the true values leads to an asymptotically equivalent optimization problem. Here W appears as a nuisance parameter which shall be fixed to the estimated width $\hat{W}$. The Bayes MAP estimate of $\boldsymbol{\omega}$ can then in principle be compute by minimizing the quadratic criterion $\|\boldsymbol{y} - V(\boldsymbol{\omega})\boldsymbol{u}\|^2$ subject to the fixed deterministic constraint J: an hypercube in $\mathbb{R}^\nu$ centered in $\hat{\boldsymbol{\theta}}$ of edge length $2\hat{W}$.

The minimization problem (13) can then equivalently be interpreted as the *Maximum Likelihood* estimation of a *deterministic angular frquency* $\boldsymbol{\omega}$ ranging on the compact set J. On this set the likelihood function is smooth and, according to standard statistical theory, the estimate must be consistent, that is converging for $N \to \infty$ to some "true value" $\boldsymbol{\omega}_0$ which has generated the observations, and asymptotically efficient.

For a finite data set, problems of the type (13) have in general several local minima. However, because of the compact feasible set constraint $\boldsymbol{\omega} \in J$, the solution must stay in a small neighborhood of the center frequency. Also, the squared norm term in (13) depends on ν sinusoidal functions of $\boldsymbol{\omega}$ and hence, for a small enough W there are no equivalent values of the frequency $\boldsymbol{\omega}$ leading to the same value of the cost. The function has generically a unique minimum.

We now propose a local-search algorithm for the problem (13) by using the a priori estimate $\hat{\boldsymbol{\theta}}$ as a starting point. Since the subspace estimate $\hat{\boldsymbol{\theta}}$ asymptotically tends to the center frequency, for large N we are allowed to identify $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$. As a first preliminary step, solve a least squares problem minimizing $\|\boldsymbol{y} - V(\hat{\boldsymbol{\theta}})\boldsymbol{u}\|^2$ to get an estimate [4] of the amplitude vector $\boldsymbol{u}$ and use the estimated amplitude vector,

$$\hat{\boldsymbol{u}} = [V(\hat{\boldsymbol{\theta}})^\top V(\hat{\boldsymbol{\theta}})]^{-1} V(\hat{\boldsymbol{\theta}})^\top \boldsymbol{y}$$

in place of $\boldsymbol{u}$ in the formulas.

[4] The estimate can also be justified based on a *noninformative prior* as in Zacharias et al. (2013).

Let $\tilde{\boldsymbol{y}}(\hat{\boldsymbol{\theta}}) := \boldsymbol{y} - V(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{u}}$ and introduce the deviation $\tilde{\boldsymbol{\omega}} := \boldsymbol{\omega} - \hat{\boldsymbol{\theta}}$. The gradient of V with respect to $\boldsymbol{\omega}$ computed at $\hat{\boldsymbol{\theta}}$, is an array of 2ν rectangular $N \times \nu$ gradient matrices of the form

$$\nabla V(\hat{\boldsymbol{\theta}}) = \Big[\nabla_{\omega_1}\boldsymbol{c}(\hat{\theta}_1), \dots, \nabla_{\omega_\nu}\boldsymbol{c}(\hat{\theta}_\nu), \nabla_{\omega_1}\boldsymbol{s}(\hat{\theta}_1), \dots, \nabla_{\omega_\nu}\boldsymbol{s}(\hat{\theta}_\nu)\Big] \tag{14}$$

where each matrix entry has only the k-th column nonzero, equal (in Matlab notation) to

$$\nabla_{\omega_k}\boldsymbol{c}(\hat{\theta}_k)[:,k] = -D_N\boldsymbol{s}(\hat{\theta}_k),\ \nabla_{\omega_k}\boldsymbol{s}(\hat{\theta}_k)[:,k] = D_N\boldsymbol{c}(\hat{\theta}_k), \tag{15}$$

where $D_N = \mathrm{diag}\{1, 2, \dots, N\}$. Hence $\nabla\{V(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{u}}\}$ turns out to be a linear combination of these 2ν, $N \times \nu$ matrices, properly combined by the corresponding components of the vector $\boldsymbol{u} \in \mathbb{R}^{2\nu}$. By this operation the zero columns are superseded and the linear combination leads to a $N \times \nu$ matrix made by linearly combining the 2ν nonzero column vectors in (14) to form a final matrix which we denote $\mathbf{M}(\hat{\boldsymbol{\theta}})$. For $\nu = 1$ we have for example $\boldsymbol{u} = [a\ b]^\top$ and

$$M(\hat{\theta}) = D_N(-\boldsymbol{s}(\hat{\theta})a + \boldsymbol{c}(\hat{\theta})b) \in \mathbb{R}^{N \times 1}.$$

With this gradient calculation established, we proceed to approximate (13) by a constrained local linear Least Squares minimization

$$\underset{\tilde{\boldsymbol{\omega}}}{\text{minimize}}\ \left\{ \|\tilde{\boldsymbol{y}} - \mathbf{M}(\hat{\boldsymbol{\theta}})\,\tilde{\boldsymbol{\omega}}\|^2 \right\}$$
$$\text{subject to: } |\tilde{\omega}_\ell| \le \hat{W} \quad (\text{equivalent to } \omega_l \in J_\ell) \tag{16}$$

for $\ell = 1, \dots, \nu$. The solution can be refined iteratively by an algorithm of the form

$$\tilde{\boldsymbol{\omega}}(k+1) = [\mathbf{M}(\boldsymbol{\omega}(k))^\top \mathbf{M}(\boldsymbol{\omega}(k))]^{-1} \times \mathbf{M}(\boldsymbol{\omega}(k))^\top \tilde{\boldsymbol{y}}(\boldsymbol{\omega}(k))\ \ k = 1, 2, , \dots \tag{17}$$

where at each step $\boldsymbol{\omega}(k) := \tilde{\boldsymbol{\omega}}(k) + \hat{\boldsymbol{\theta}}$ is substituted back in place of $\boldsymbol{\omega}(k-1)$ or, initially, of $\hat{\boldsymbol{\theta}}$ in the expression of the gradient. The scheme is initialized for $k = 0$ setting $\boldsymbol{\omega}(0) = \hat{\boldsymbol{\theta}}$ and then stopping when the difference $\tilde{\boldsymbol{\omega}}(k+1) = \boldsymbol{\omega}(k+1) - \boldsymbol{\omega}(k)$ becomes small enough. It requires to check at each step if $|\tilde{\omega}_\ell| \le \hat{W}$ otherwise the estimator should be re-initialized. Alternatively, we may try to keep $\|\tilde{\omega}\|$ small by adding a ridge penalty term $\lambda(k)\|\tilde{\boldsymbol{\omega}}(k)\|^2$ with $\lambda(k) \to 0$ for k large (in order to retain consistency), to the least squares formulation. This may in fact also make the computation of the inverse better conditioned.

Remark 1. The reasoning above can be extended to include multiple snapshots of data in a straightforward manner. Since the conditional likelihood function for each snapshot multiplies given the hidden frequencies, the squared-norm term in the objective function of (13) becomes $\|\mathcal{Y} - V(\boldsymbol{\omega})\mathcal{U}\|_{\mathrm{F}}^2$, where $\mathcal{Y}$ and $\mathcal{U}$ are matrices whose columns are the data and the amplitude vectors, respectively, and the subscript $_{\mathrm{F}}$ denotes the Frobenius norm. A similar linearization scheme can be devised to solve the enlarged optimization problem.

4. SIMULATIONS

In this section, we present simulation results showing that the Bayesian MAP estimate described in Sec. 3 works very well in the case of *two* hidden frequencies ($\nu = 2$). We do Monte-Carlo simulations under different parameter configurations. The main parameters are three: the signal

length N, the number of snapshots (cross sections) L, and the signal-to-noise ratio (SNR). The latter is defined as $20\log_{10}(\sigma/\sigma_{\mathbf{w}})$ where for simplicity we have taken the signal variances $\sigma_1^2 = \cdots = \sigma_\nu^2 = \sigma^2$. The key steps in each trial are outlined below:

(1) Randomly generate the hyperparameters (θ_1, θ_2, W). More precisely, first the bandwidth W is drawn from the uniform distribution in $2\pi \times [0.01, 0.05]$, and then the center frequencies θ_1 and θ_2 are drawn from $U[W, \pi - W]$ such that $|\theta_1 - \theta_2| > 2W$ and the supporting intervals for the two frequencies do not overlap.
(2) Given the frequency hyperparameters, generate L independent measurement sequences (snapshots) $\boldsymbol{y}$ of length N according to the model (1) and the assumptions after it.
(3) Given the multiple snapshots of data, run the Empirical Bayes procedure for hyperparameter estimation described in Subsec. 2.2 to compute the estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{W})$.
(4) Given the estimated hyperparameters (Step 3) and the *multiple* snapshots of measurements (Step 2), run the algorithm for Bayesian MAP estimation explained in Sec. 3 to compute the frequency estimate $\hat{\boldsymbol{\omega}}^{\text{MAP}}$ and compare it with the true band center $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

In the first set of Monte-Carlo simulations, we fix the signal length $N = 100$ and SNR $= 15$ dB, and change the number L of cross sections. Each Monte-Carlo simulation consists of 100 trials. The relative errors[5] of $\hat{\boldsymbol{\omega}}^{\text{MAP}}$ with respect to the true center-frequency vector $\boldsymbol{\theta}$ under each parameter configuration are shown in Fig. 1. In each box, the red line indicates the median of the 100 errors, while the upper and lower bounds of the box represent the 75% and 25% percentiles, respectively. The dashed lines (called "whiskers") extend to the 95% and 5% percentiles of the errors, and the red crosses are considered as outliers. It appears that the MAP estimate of the frequencies is close to the true band centers with a cumulative relative error below 6%. Moreover, the estimation accuracy improves as more snapshots of data are available.

We repeat the simulation with the same SNR and change N to 500 and 1000, and the results are depicted in Figs. 2 and 3, respectively. It is evident that the cumulative relative errors reduce as N increases (i.e., more data are available). In the case of $N = 1000$, the relative errors are within 1%, which means that the MAP estimate $\hat{\boldsymbol{\omega}}^{\text{MAP}}$ is very close to the true band centers $\boldsymbol{\theta}$. Notice that these new data sets are quite large, and we have used them only to investigate the asymptotic behavior of the Bayesian MAP estimate (see our conjecture in Remark 2). We do not suggest that this corresponds to any realistic settings.

Next, we test our algorithm versus the SNR. For example, we fix $N = 100$, $L = 200$, and do Monte-Carlo simulations of 50 trials under different SNRs. The relative errors in each Monte-Carlo simulation are averaged and plotted in Fig. 4. The average relative errors are still quite small even in the low SNR regime. Moreover, the flatness of the curves seems to indicate that the Bayesian MAP estimate is very robust against noise.

[5] The relative error here is defined as $\|\hat{\boldsymbol{\omega}}^{\text{MAP}} - \boldsymbol{\theta}\|/\|\boldsymbol{\theta}\|$.

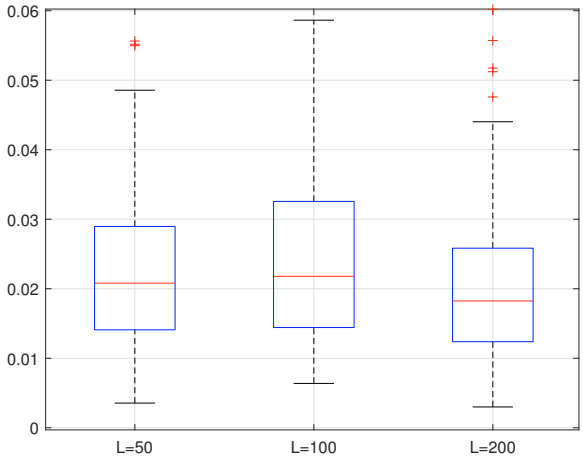

Fig. 1. Relative estimation errors of two hidden frequencies $(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ using the Bayesian MAP method in Sec. 3 versus the number L of cross sections while $N = 100$ and SNR $= 15$ dB.

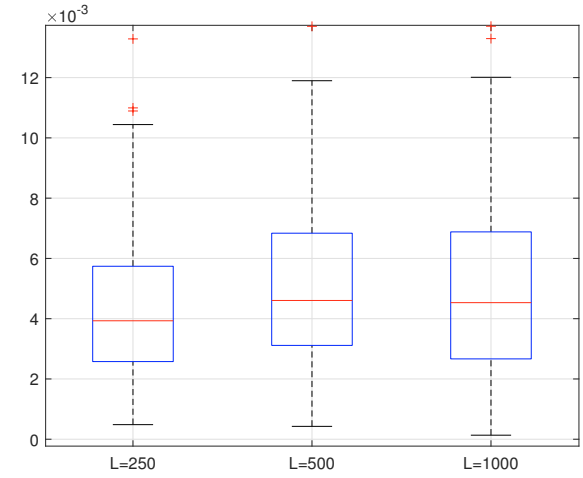

Fig. 2. Relative estimation errors of two hidden frequencies $(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ using the Bayesian MAP method in Sec. 3 versus the number L of cross sections while $N = 500$ and SNR $= 15$ dB.

Remark 2. The above simulations all empirically support the conjecture mentioned in Picci and Zhu (2020), which states that for $N \to \infty$ and small enough W, the Bayesian estimate $\hat{\boldsymbol{\omega}}^{\text{MAP}}$ should converge to the true center frequency $\boldsymbol{\theta}$.

5. CONCLUSIONS

We have formulated the problem of frequency estimation in an Empirical Bayesian framework by first imposing a natural uniform prior probability density on the unknown frequency. In this way the estimation of the hyperparameters of the a priori distribution can be accomplished via essentially linear techniques of subspace identification, exploiting the special structure of the covariance matrix of the posterior process. Using the estimated hyperparameters one can adapt the prior to the data and this leads

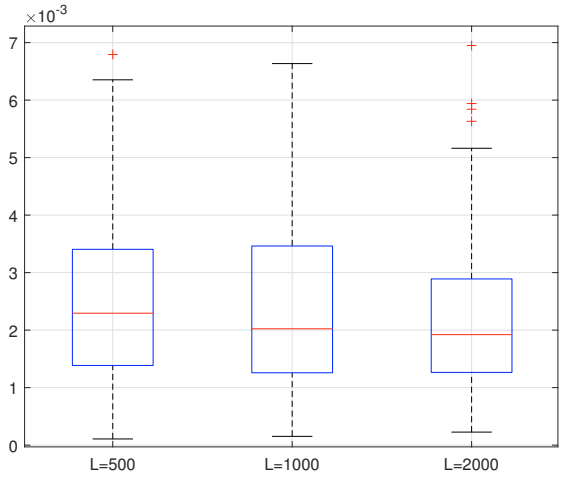

Fig. 3. Relative estimation errors of two hidden frequencies $(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ using the Bayesian MAP method in Sec. 3 versus the number L of cross sections while $N = 1000$ and SNR = 15 dB.

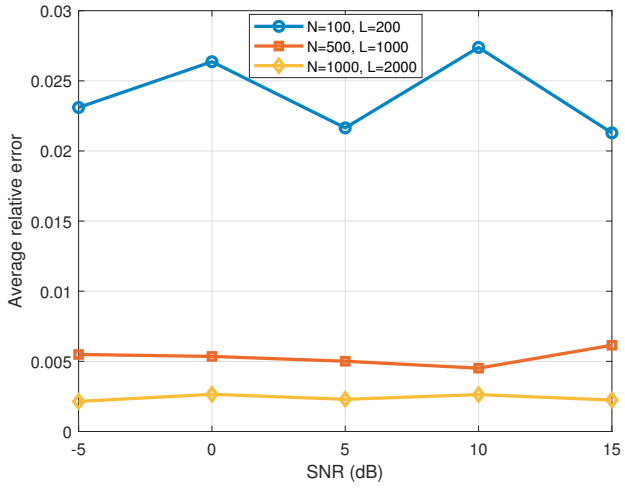

Fig. 4. Average relative estimation error in one Monte-Carlo simulation of two hidden frequencies $(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ using the Bayesian MAP method in Sec. 3 versus the SNR with three data sets of different sizes.

to Bayesian estimates which are asymptotically maximum likelihood and therefore the best possible in a variety of metrics. Extensive simulations indicate that the Bayesian MAP estimate of the frequencies given multiple snapshots of data is very close to the true band centers of the prior distribution, which empirically justifies our Empirical Bayes philosophy for the frequency estimation problem.

REFERENCES

Aravkin, A., Burke, J.V., Chiuso, A., and 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*, 125–132. Brussels, Belgium.

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

Chiuso, A. (2015). Regularization and Bayesian learning in dynamical systems: Past, present and future (invited paper). In *Proc. SYSID 15*. Beijing, China.

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

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

Favaro, M. and Picci, G. (2015). A Bayesian nonparametric approach to frequency estimation. In *Proc. SYSID 2015*, volume 48(28), 478–483. IFAC-Papers On Line.

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

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

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

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

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation, Second Ed.* Springer Texts in Statistics.

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

Petrone, S., Rizzitelli, S., Russeau, J., and Scricciolo, C. (2014). Empirical Bayes methods in classical and Bayesian inference. *Metron*, 72, 201–215.

Picci, G. and Zhu, B. (2020). An empirical Bayes approach to frequency estimation. arXiv:1910.09475v2 [eess.SP].

Picci, G. and Zhu, B. (2021). Approximation of stationary processes and Toeplitz spectra. Accepted for presentation at the MTNS Symposium, arXiv:2009.05758 [math.PR].

Picci, G. and Zhu, B. (2019). Bayesian frequency estimation. In *the 17th European Control Conference, ECC 2019, Naples, Italy*, 848–853. IEEE.

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.

Yuan, M., Wan, C., and 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., and Chatterjee, S. (2013). Line spectrum estimation with probabilistic priors. *Signal Processing*, 93(11), 2969–2974.