A new kernel-based approach for spectral estimation

Mattia Zorzi

Abstract—The paper addresses the problem to estimate the power spectral density of an ARMA zero mean Gaussian process. We propose a kernel based maximum entropy spectral estimator. The latter searches the optimal spectrum over a class of high order autoregressive models while the penalty term induced by the kernel matrix promotes regularity and exponential decay to zero of the impulse response of the corresponding one-step ahead predictor. Moreover, the proposed method also provides a minimum phase spectral factor of the process. Numerical experiments showed the effectiveness of the proposed method.

I. INTRODUCTION

We consider the problem to estimate the power spectral density of an autoregressive and moving average (ARMA) zero mean Gaussian process. The latter has been successfully addressed by the so called “Tunable High REsolution” (THREE)-like methods for which a large body of literature is available [1], [2], [3], [4], [5], [6], [7], [8] as well as for the multidimensional case [9], [10], [11] and the case of dynamic networks [12], [13], [14], [15]. According to this approach, the output covariance of a bank of filters is used to extract information on the input power spectral density. More precisely, the estimated spectrum is the one maximizing a suitable entropy-like functional and matching the output covariance matrix. For a comprehensive analysis of the possible entropy-like functionals we refer to [16], [17]. In regard to the bank of filters, it has to be designed by the user and it fixes the model class in which we search the optimal spectrum [18]. In general the design of such a bank is very challenging especially in the case that no a priori information about the process is available.

In the simple case that the bank is constituted by \(n\) delays, then we obtain the maximum entropy (ME) estimator which is also known as Burg estimator [19]. The latter searches the optimal spectrum over the class of autoregressive (AR) processes of order \(n\). The selection of \(n\) (which is the unique parameter of the bank of filters) can be performed by resorting to complexity measures such as AIC and BIC, [20], [21].

Since an (stable) ARMA processes can be approximated by an high order process, one could use the ME spectral estimator with \(n\) large for ARMA process. However, the resulting estimator will be affected by high variance. In this paper we propose a kernel based ME spectral estimator which searches the optimal spectrum over the class of high order AR processes while the complexity (i.e. degrees of freedom) is controlled through a penalty term induced by kernel matrix. More precisely adopting the Bayesian perspective, the impulse response of the predictor (which is a “long” FIR) is modeled as a zero mean Gaussian vector with covariance matrix (i.e. kernel matrix) embedding impulse response regularity and the fact that it decays to zero in an exponential way. Then, we show that the estimated spectrum is given by solving a generalized version of the Yule-Walker equations whose solution leads to a minimum phase spectral factor of the process and thus the estimated spectrum is well defined. Such result is a natural extension of the proof by Stoica and Nehorai [22] for the usual Yule-Walker equations. Finally, the kernel matrix is not known: we propose an empirical Bayes approach to estimate it.

It is worth noting that spectral estimation can be performed using the kernel based method proposed in [23]. More precisely, it estimates the correlation function, however there is no theoretical guarantee that the corresponding spectrum is positive over the unit circle. An alternative is to consider kernel based prediction error methods (PEM), [24], [25]. However, it guarantees the predictor stability but it is not guaranteed the system stability [26], i.e. it is not guaranteed that the estimated spectrum is well defined.

The outline of the paper is as follows. In Section II we review the ME spectral estimator. In Section III we introduce the kernel based ME spectral estimator and in Section IV we propose an empirical Bayes approach to estimate the kernel matrix from the data. In Section V we provide a Fisherian interpretation of the proposed estimator. In Section VI we present some numerical experiments. Finally, in Section VII we draw the conclusions.

II. MAXIMUM ENTROPY SPECTRAL ESTIMATION

Let \(y = \{y_t, \ t \in \mathbb{Z}\}\) be a zero mean stationary Gaussian stochastic process. The latter is completely described by the (real-valued) power spectral density

\[
\Phi(e^{i\theta}) = \sum_{k \in \mathbb{Z}} r_k e^{i\theta k}, \quad \theta \in [0, 2\pi]
\]  

(1)

where \(r_k = \mathbb{E}[y_t y_{t+k}]\) is the \(k\)-th covariance lag. Notice that \(\Phi(e^{i\theta}) > 0\) for any \(\theta\). We consider the problem to estimate \(\Phi\) given a finite length sequence \(y_N := \{y_1, \ldots, y_N\}\) extracted from a realization of \(y\). Modern high resolution spectral estimators are given by maximizing a suitable entropy-like functional subject to some moment constraints, see [1]. Here, the model class of the process is characterized by a bank of filters, typically denoted by \(G(z)\), which has to be designed by the user. However, the choice of \(G(z)\) is difficult if no a priori information is available. In the special case...
that $G(z) = [1 \ z^{-1} \ dots \ z^{-n}]^T$, with $n \leq N$, and the objective function is the entropy rate, we obtain the ME estimator

$$
\hat{\Phi} = \arg\max_{\Phi} \int_{-\pi}^{\pi} \log \det \Phi(e^{i\theta}) \, d\theta
$$

subject to

$$
\int_{-\pi}^{\pi} \Phi(e^{i\theta}) e^{i\theta k} \, d\theta = \hat{r}_k, \ k = 0 \ldots n
$$

where $\hat{r}_k = \frac{1}{N} \sum_{t=1}^{N-k} y_{t+k}^2$ is the estimate of $r_k$ using data $y^N$. The latter searches the optimal spectrum over the class of AR models of order $n$

$$
b(z) y_t = e_t
$$

where $b(z) = \sum_{k=0}^{n} b_k z^{-k}$ and $e = \{e_t, t \in \mathbb{Z}\}$ is white Gaussian noise with unit variance. Indeed $\hat{\Phi} = (\hat{b}_{ML}(z) \hat{b}_{ML}(z))^{-1}$ where $\hat{b}_{ML}(z)$ is characterized as follows. If we define

$$
b = [b_0 \ \ dots \ \ b_n]^T,
$$

then $\hat{b}_{ML}(z)$ corresponds to the maximum likelihood (ML) estimator

$$
\hat{b}_{ML} = \arg\min_{b \in \mathbb{R}^{n+1}} \ell(y^N; b)
$$

where $\ell(y^N; b)$ is the so called Whittle log-likelihood

$$
\ell(y^N; b) := \frac{N-n}{2} \left( -\log(b^T vv^T b) + b^T \hat{\Sigma} b \right),
$$

$$
\hat{\Sigma} = \begin{bmatrix}
\hat{\rho}_0 & \hat{\rho}_1 & \ dots & \hat{\rho}_n \\
\hat{\rho}_1 & \hat{\rho}_0 & \ dots & \vdots \\
\vdots & \vdots & \ dots & \hat{\rho}_1 \\
\hat{\rho}_n & \dots & \hat{\rho}_1 & \hat{\rho}_0
\end{bmatrix},
$$

$v = [1 \ 0 \ dots \ 0]^T$. Notice that, $\hat{\Sigma}$ is Toeplitz and positive definite with high probability. It is not difficult to prove that the solution of (5) is $\hat{b}_{ML} = a/\sqrt{a_0}$ where $a = [a_0 \ a_1 \ dots \ a_n]^T$ is given by solving the Yule-Walker equations

$$
a = \hat{\Sigma}^{-1} v.
$$

It is well known that $\hat{b}_{ML}(z)$ has zeros inside the unit circle and thus $\hat{b}_{ML}(z)^{-1}$ is a minimum phase spectral factor of $\hat{\Phi}$. Moreover, $\hat{\Phi} > 0$, i.e. the spectrum is well defined. Although in (2) the bank of filters is very simple, we need to choose $n$, that is the order of the AR model. The latter step is typically addressed by resorting to complexity measures such as AIC and BIC [20], [21]. However, the selection of $n$ is not trivial and may compromise the optimality properties, in particular for shorter data records; moreover, it corresponds to a combinatorial optimization problem. In what follows we propose a kernel based maximum entropy estimator which searches the optimal over a large model class (i.e. high order AR models) while the complexity is controlled by a regularization term induced by the kernel matrix.

### III. Spectral estimation using a Gaussian prior

We address the problem to estimate the power spectral density $\Phi$ of a zero mean ARMA Gaussian stationary process $y$. Let $w(z)$ be a minimum phase rational spectral factor of $\Phi$, then

$$
w(z)^{-1} y_t = e_t
$$

where $e_t$ is white Gaussian noise with variance equal to one. Since $w(z)^{-1}$ is rational, we can rewrite it using the Laurent expansion:

$$
w(z)^{-1} = \sum_{k=0}^{\infty} b_k z^{-k}.
$$

Notice that $-b_0^{-1} b_k$, with $k \geq 1$, is the impulse response of the one-step ahead predictor (see equation (16) below). Since $w(z)$ is minimum phase, we have that the sequence $b_k$, with $k \geq 0$, belongs to $\ell_1$, i.e. $\sum_{t=0}^{\infty} |b_k| < \infty$. This implies that

$$
w(z)^{-1} \approx b(z) := \sum_{k=0}^{n} b_k z^{-k}
$$

where $n$ is sufficiently large. In other words, we can always approximate an ARMA process with an high order AR process. Let $b$ be defined as in (4). Given the dataset $y^N$, then we could estimate it by solving (5). However, the corresponding estimator will be affected by high variance, i.e. the complexity of the model is not well controlled.

In order to counteract this aspect, we adopt the Bayesian perspective. More precisely, we model $b$ as a Gaussian random vector with zero mean and covariance matrix $\lambda K$. $\lambda > 0$ is called scale factor and $K > 0$ is called kernel matrix. Therefore, its probability density is:

$$
p(b) = \frac{1}{\sqrt{(2\pi)^{n+1} \det K}} \exp \left( -\frac{1}{2\lambda} b^T K^{-1} b \right).
$$

The, the negative log-likelihood of $y^N$ and $b$ is

$$
\ell(y^N, b) = -\log p(y^N | b) = - \log p(y^N | b) - \log p(b)
$$

$$
= \frac{N-n}{2} \left( -\log(b^T vv^T b) + b^T \hat{\Sigma} b \right)
$$

$$
+ \frac{1}{2} \log \det(\lambda K) + \frac{1}{2\lambda} b^T K^{-1} b + c_1
$$

where $c_1$ is a constant term not depending on $b$ and $K$. Assume that we have a preliminary estimate of $b_0$, say $\hat{b}_0$, then

$$
\log(b^T vv^T b) = \log b_0^2 = 2 \log b_0 \approx 2 \log \hat{b}_0 + \frac{2}{b_0} (b_0 - \hat{b}_0)
$$

$$
= 2 \log \hat{b}_0 + \frac{2}{b_0} (b^T v - \hat{b}_0)
$$

accordingly

$$
\ell(y^N, b) \approx \frac{N-n}{2} \left( -\frac{2}{b_0} b^T v + b^T \hat{\Sigma} b \right)
$$

$$
+ \frac{1}{2} \log \det(\lambda K) + \frac{1}{2\lambda} b^T K^{-1} b + c_2
$$

535
where $c_2$ is a constant term not depending on $b$ and $K$. It is worth noting that we can rewrite the AR model in (3) as
\begin{equation}
y_t = a(z)y_t + u_t
\end{equation}
where $a(z) = 1 - b_0^{-1}b(z)$ and $u_t$ is white noise with variance equal to $b_0^{-2}$. As suggested in [27], a simple an effective way to estimate $b_0^{-2}$ is to consider a low order AR model and solve (2). Then, notice that adding the constant term $(N-n)v^T \hat{\Sigma}^{-1}v/2b_0^2$ in (15) we obtain
\begin{equation}
\ell(y_N, b) \approx \frac{1}{2} \|\tilde{v} - \Phi b\|^2 + \frac{1}{2} \log \det(\lambda K) + \frac{1}{2\lambda} b^T K^{-1} b + c_3
\end{equation}
where
\begin{equation}
\tilde{v} = \frac{\sqrt{N-n}}{b_0^{-1} v_L, \quad \Phi = \sqrt{N-n}L^T, \quad \hat{\Sigma} = LL^T
\end{equation}
c_3 is a constant term not depending on $b$ and $K$. Notice that $L$ is an arbitrary square root decomposition of $\hat{\Sigma}$, e.g. the Cholesky decomposition. Accordingly, the ML estimator for $b$ is
\begin{equation}
b_{ML} = \arg\min_b \ell(y_N, b) = \arg\min_b \|\tilde{v} - \Phi b\|^2 + \lambda^{-1} \|b\|^2_{K^{-1}}.
\end{equation}
Therefore, $b_{ML}$ is the solution to a regularized Least squares problem. In particular, it is not difficult to prove that
\begin{equation}
b_{ML} = \lambda K \Phi^T (\lambda K \Phi \Phi^T + I)^{-1}\tilde{v} = b_0^{-1} (\hat{\Sigma} + (N-n)\lambda K)^{-1} v
\end{equation}
Notice that $b_{ML}$ does not depends on the particular decomposition $\hat{\Sigma} = LL^T$. It is worth noting that the role of the penalty term $\lambda^{-1} \|b\|^2_{K^{-1}}$ is to reduce the variance of the estimator since $n$ is chosen large. In case that $\lambda \to \infty$, i.e. there is no a priori information on $b$, then we obtain the usual maximum entropy problem in (5).

It remains to design the kernel matrix $K$ that encodes the a priori information on $b(z)$. For instance, we want that $b_{ML}$ corresponds to a polynomial $b_{ML}(z)$ whose zeros are inside the unit circle. Moreover, $b_0$ should decay to zero in an exponential way as $k$ increases because the latter are the coefficients of the Laurent expansion of a transfer function whose poles are inside the unit circle. In what follows we propose two kernel matrices satisfying such requirements. The latter has been already used in kernel based PEM methods in a successful way, see e.g. [24], [25].

A. Diagonal (DI) kernel

We consider the kernel matrix defined as
\begin{equation}
K_{DI} = \text{diag}(\beta, \beta^2, \beta^3 \ldots \beta^{n+1})
\end{equation}
where $\beta \in (0, 1)$. In Figure 1 (top panel) we show five realizations of $b \sim N(0, \lambda K_{DI})$ with $\lambda = 1$ and $\beta = 0.85$. As we see, in all these realizations $b_k$ decays in an exponential way. Indeed, for any realization of $b$ we have that $\|b\|^2_{K^{-1}} < \infty$ almost surely which implies that $b_k$ decays in an exponential way almost surely.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1}
\caption{Top panel. Five realizations of $b$ using the Diagonal kernel. Bottom panel. Five realizations of $b$ using the Tuned-Correlated kernel.}
\end{figure}

\begin{proposition}
Let $b_{ML}$ be the estimate using $K_{DI}$. If the corresponding polynomial $b_{ML}(z)$ is nonnull, then it has zeros inside the unit circle.
\end{proposition}
\begin{proof}
By (19) we have that
\begin{equation}
(\hat{\Sigma} + R)b_{ML} = b_0^{-1} v
\end{equation}
where
\begin{equation}
R = ((N-n)\lambda K)^{-1} := \text{diag}(r_1, r_2 \ldots r_{n+1})
\end{equation}
and $r_k = ((N-n)\lambda \beta^k)^{-1}$. Thus, $r_{k+1} \geq r_k > 0$. Let $\xi$ be a zero of $b_{ML}(z)$. If $\xi = 0$, then it is inside the unit circle and we have finished. Otherwise we have $\xi \neq 0$. Define
\begin{equation}
u_k = \begin{cases}
\tilde{b}_0, & k = 1 \\
\tilde{b}_{k-1} + \tilde{b}_{k-1}, & k = 2 \ldots n
\end{cases}
\end{equation}
where $b_{ML} = [\tilde{b}_0 \tilde{b}_1 \ldots \tilde{b}_n]^T$. Accordingly,
\begin{equation}
u = \tilde{v} + b_{ML}
\end{equation}
where $u = [u_1 u_2 \ldots u_n] 0^T$ and $\tilde{u} = [0 u_1 u_2 \ldots u_n]$. Using (24), we have
\begin{equation}
u^T (\hat{\Sigma} + R) u = [\tilde{v}^T u^* + \tilde{u}^* (\hat{\Sigma} + R) \tilde{u} + b_{ML}^T (\hat{\Sigma} + R) b_{ML} + 2 \mathbb{R}[\tilde{u}^* (\hat{\Sigma} + R) \tilde{u}]
\end{equation}
\begin{equation}
= [\tilde{v}^T u^* + \tilde{u}^* (\hat{\Sigma} + R) \tilde{u} + b_{ML}^T (\hat{\Sigma} + R) b_{ML} + 2 \tilde{u}^* \mathbb{R}[\tilde{u}^T \tilde{u}]
\end{equation}
where in the last equality we exploited (21). Notice that $\tilde{u}^T \tilde{u} = 0$, accordingly the last term on the right hand side is equal to zero. Moreover $\nu^T \tilde{u} = 0$, because $\tilde{u}$ is a Toeplitz matrix. Therefore,
\begin{equation}
u^T (\hat{\Sigma} + R) u = [\tilde{v}^T u^* + \tilde{u}^* R \tilde{u} + b_{ML}^T (\hat{\Sigma} + R) b_{ML}].
\end{equation}
Finally notice that
\begin{equation}
\tilde{u}^* R \tilde{u} = \sum_{k=1}^n r_{k+1} |u_k|^2 \geq \sum_{k=1}^n r_k |u_k|^2 = u^* Ru.
\end{equation}
Using the latter inequality in (26), we obtain
\begin{equation}
u^T (\hat{\Sigma} + R) u \geq [\tilde{v}^T u^* (\hat{\Sigma} + R) u + b_{ML}^T (\hat{\Sigma} + R) b_{ML}]
\end{equation}
and thus
\begin{equation}
1 - |\tilde{v}|^2 u^* (\hat{\Sigma} + R) u \geq b_{ML}^T (\hat{\Sigma} + R) b_{ML}.
\end{equation}
The right hand side in (29) is positive because \( \hat{\Sigma} + R > 0 \) and \( \hat{b}_{ML} \neq 0 \) (otherwise \( \hat{b}_{ML}(z) = 0 \)). Moreover, \( u'(\hat{\Sigma} + R)u > 0 \) because \( u \neq 0 \) (otherwise \( \hat{b}_{ML} = 0 \)). Accordingly, we have \( 1 - |\xi|^2 > 0 \), that is \( \xi \) is inside the unit circle.

It is worth noting the proof above is in the same spirit of the one in [22] for the Yule-Walker equations in (8).

**B. Tuned-Correlated (TC) kernel**

We consider the kernel matrix \( K_{TC} \) whose entry in position \((t,s)\) is defined as

\[
[K_{TC}]_{t,s} = \beta^{\max(t,s)} - \beta^{t+s+2}, \quad t,s = 1, \ldots, n + 1
\]

(30)

where \( \beta \in (0,1) \). It is worth noting that \( K_{TC} \) in (30) is a modified version of the one proposed in [25]. Indeed, the latter is defined as

\[
[K_{TC}]_{t,s} = \beta^{\max(t,s)}, \quad t,s = 1, \ldots, n + 1.
\]

(31)

However, for \( n \) large the two kernel matrices \( K_{TC} \) and \( \hat{K}_{TC} \) are very similar. In Figure 1 (bottom panel) we show five realizations of \( b \sim N(0, \lambda K_{DI}) \) with \( \lambda = 1 \) and \( \beta = 0.85 \). As we see, in all these realizations \( b_k \) decays in an exponential way. Moreover, these impulse responses are more smooth than the ones using the diagonal kernel. Also in this case, for any realization of \( b \) we have that \( b_k \) decays in an exponential way almost surely. It is not difficult to prove that

\[
K_{TC} = (FDF^T)^{-1}
\]

(32)

where

\[
F = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
-1 & 1 & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots \\
0 & \cdots & \ddots & 1
\end{bmatrix}
\]

(33)

and

\[
D = \frac{1}{\beta - \beta^2} \text{diag}\left(\frac{1}{\beta}, \frac{1}{\beta^2}, \ldots, \frac{1}{\beta^{n-1}}, \frac{1}{\beta^n}\right)
\]

(34)

**Proposition 2:** Let \( \hat{b}_{ML} \) be the estimate using \( K_{TC} \). If the corresponding polynomial \( \hat{b}_{ML}(z) \) is nonnull, then it has zeros inside the unit circle.

**Proof:** Equation (21) still holds with \( R = FDF^T \) where \( D = \text{diag}(d_1, d_2, \ldots, d_{n+1}) \)

(35)

and \( d_k = ((N-n)\lambda \beta^{k-1} - (\beta - \beta^2))^{-1} \) for \( k = 1, \ldots, n + 1 \). Thus, \( d_{k+1} \geq d_k > 0 \) for \( k = 1, \ldots, n \). Let \( \xi \) be a zero of \( \hat{b}_{ML}(z) \). If \( \xi = 0 \), then it is inside the unit circle and we have finished. Otherwise we have \( \xi \neq 0 \). We define \( u \) and \( \tilde{u} \) as in the proof of Proposition 1. Accordingly, (24) holds. Using the same arguments in the proof of Proposition 1, we obtain (26). Then, we have

\[
\tilde{u}^* R \tilde{u} = d_1|u_1|^2 + \sum_{k=1}^{n} d_{k+1}|u_k - u_{k+1}|^2 + d_{n+1}|u_n|^2
\geq d_1|u_1|^2 + \sum_{k=1}^{n} d_k|u_k - u_{k+1}|^2 + d_{n+1}|u_n|^2
\geq \sum_{k=1}^{n} d_k|u_k - u_{k+1}|^2 = u^* R u
\]

and thus \( \tilde{u}^* R \tilde{u} \geq u^* R u \). Accordingly, using the same arguments in the proof of Proposition 1 we obtain (29) and thus \( \xi \) is inside the unit circle.

**IV. Hyperparameters estimation**

The maximum likelihood estimator of Section III depends on \( \eta := [\lambda \beta]^T \) which are called hyperparameters in the machine learning community. \( \eta \) is not known and has to be estimated form the data \( y^N \). According to the empirical Bayes perspective, [28], an estimate of \( \eta \) is given by maximizing the probability density of \( y^N \) under the model AR of order \( n \) in (3) with \( b \sim N(0, \lambda K_{\beta}) \) where we made explicit the fact that \( K \) depends on \( \beta \). In our case, such probability density is

\[
p(y^N) = c_4 \int_{\mathbb{R}^n+1} \exp(-\ell(y^N, b)) \, db
\]

(36)

where \( c_4 \) is a positive constant term not depending on \( b \) and \( \eta \). Accordingly, an estimator for \( \eta \) is

\[
\hat{\eta} = \text{argmax}_{\eta} \ell(y^N)
\]

(37)

s.t. \( \lambda > 0 \), \( 0 < \beta < 1 \)

where \( \ell(y^N) = -\log p(y^N) \) is the so called negative log-marginal likelihood function. In order to find an analytical expression for \( \ell(y^N) \) we need the following results.

**Proposition 3:** Let \( h : \mathbb{R}^{n+1} \to \mathbb{R} \) be such that its Hessian matrix \( \partial^2 h(b)/\partial b^2 = H \) is constant and positive definite. Then,

\[
\int_{\mathbb{R}^{n+1}} \exp(-h(b)) \, db = \frac{1}{\sqrt{\det(2\pi H)}} \exp(-h(b^0))
\]

(38)

where \( b^0 \) minimizes \( h(b) \), see [29].

In our case \( h(b) = \ell(y^N, b) \) and

\[
H = \frac{\partial^2 \ell(y^N, b)}{\partial b^2} = \Phi^T \Phi + \lambda^{-1} K_{\beta}^{-1}
\]

(39)

which is constant and positive definite. Moreover, \( b^0 = \hat{b}_{ML} \) and it is not difficult to see that

\[
\ell(y^N, \hat{b}_{ML}) = \frac{1}{2} \tilde{v}^T (\lambda \Phi K_{\beta} \Phi^T + I)^{-1} \tilde{v} + \frac{1}{2} \log \det(\lambda K_{\beta}) + c_2.
\]

(40)

Then, we can apply Proposition 3, thus

\[
p(y^N) = \frac{1}{\sqrt{\det(2\pi (\Phi^T \Phi + \lambda^{-1} K_{\beta}^{-1}))}} \exp(-\ell(y^N, \hat{b}_{ML}))
\]

(41)

and

\[
\ell(y^N) = \frac{1}{2} \log \det(\lambda \Phi K_{\beta} \Phi^T + I) + \frac{1}{2} \tilde{v}^T (\lambda \Phi K_{\beta} \Phi^T + I)^{-1} \tilde{v} + \frac{1}{2} \log \det(\lambda K_{\beta}) + c_5
\]

(42)

where \( c_5 \) is a constant term not depending on \( \eta \).
The resulting spectral estimation procedure can be summarized in the following steps:

1) Compute a preliminary estimate \( \hat{b}_0 \) of \( b_0 \) from \( y^N \) using a low order AR model
2) Set \( n \) large and such that \( n \leq N \)
3) Compute the Toeplitz matrix \( \hat{\Sigma} \) in (7) from \( y^N \)
4) Compute a square root of \( \hat{\Sigma} \)
5) Compute \( \Phi \) and \( \bar{v} \) as in (17)
6) Compute \( \hat{\beta} = [\lambda, \hat{\beta}]^T \) which minimizes (42)
7) \( \hat{b}_{ML} = \lambda \hat{K}_p \Phi^T(\lambda \hat{K}_p \Phi^T + 1)^{-1} \bar{v} \).

It is worth noting that (42) is not a convex function. However, there is a large body of literature dealing with the global minimization problem of functions of type (42) under constraints \( \lambda > 0 \) and \( 0 < \beta < 1 \), see e.g. [30].

V. DEGREES OF FREEDOM

Problem (18) has also an interpretation in the Fisherian perspective, i.e. when \( b \) is understood as an unknown but fixed vector. In this case, (18) is a regularized ML estimator of \( b \). Since (18) is a regularized Least squares problem, the corresponding degrees of freedom of the estimated model are defined as [31]:

\[
df(\eta) = \text{tr}[(\Phi^T \Phi + (\lambda K_p)^{-1})^{-1} \Phi^T]
= \text{tr}[(\hat{\Sigma} + ((N-n)\lambda K_p)^{-1})^{-1} \hat{\Sigma}].
\] (43)

It is well known that \( df(\eta) \in \mathbb{R} \) (and nonnegative) measures the complexity of the model: the higher \( df(\eta) \) is, the more complex the model is. In the special case that \( \lambda = \infty \), i.e. no regularization is present, we have:

\[
df(\eta) = \text{tr}[(\hat{\Sigma}^{-1} \hat{\Sigma})] = n + 1
\] (44)

that is, as expected, the complexity is an integer number corresponding to the number of parameters to estimate for the AR model. We conclude that regularization allows to control the complexity, through \( \eta \), in a finer way than the case without regularization.

VI. EXAMPLE

We test the performance of the proposed kernel based spectral estimator using the DI and TC kernel with the ME spectral estimator. An alternative way to estimate the spectrum is to use a kernel-based PEM method introduced in [24]. The latter estimates an high order AR model by minimizing the one-step ahead prediction error of the model plus the penalty term induced by the DI or the TC kernel. Note that this method does not guarantees that the \( b(z) \) has zeros inside the unit circle, and thus that the estimated power spectral density is well defined. In what follows we will use the following notation for the spectral estimators:

- ME denotes the ME estimator equipped with the BIC criterium for selecting the optimal \( n \) over the set \( \{1, 2 \ldots 50\} \)
- ME-DI denotes the kernel based ME estimator using the DI kernel
- ME-TC denotes the kernel based ME estimator using the TC kernel

For all the kernel based methods we set \( n = 50 \).

In the first experiment we generate a dataset \( y^N \) with \( N = 500 \) from the ARMA process

\[
y_t = \sqrt{2} \frac{(z - z_0)(z - \bar{z}_0)}{(z - p_0)(z - \bar{p}_0)} e_t
\] (45)

where \( z_0 = 0.85e^{0.52} \) and \( p_0 = 0.98e^{0.482} \). Each estimator returns a polynomial \( \hat{b}(z) \), then the corresponding spectral estimator is \( \hat{\Phi}(z) = (\hat{b}(z) \hat{b}(z^{-1}))^{-1} \). Figure 2 depicts the estimated spectra versus the true spectrum. As we can see ME-DI and ME-TC performs better than ME. Finally, the PEM method provides the worst performance. Regarding the degrees of freedom we have: \( df_{ME} = \hat{n}_{BIC} + 1 = 8 \) \( df_{ME-DI}(\hat{\eta}) = 10, 8131 \) and \( df_{ME-TC}(\hat{\eta}) = 8, 8943 \) where \( \hat{n}_{BIC} \) is the optimal AR order according to the BIC criterium.

In the second experiment we consider a Monte Carlo study composed by 100 runs. In each run we generate a dataset \( y^N \) with \( N = 500 \) of an ARMA process \( y_t = w(z)e_t \) where \( w(z) \) is a minimum phase spectral factor with 3 zeros and poles. Each poles has modulus equal to 0.98 while the phase is drawn form a uniform distribution over \([0, \pi]\). Each zero has modulus equal to 0.85 while the phase is chosen randomly (according to a uniform distribution) but sufficiently close to the one of a pole. More precisely, if \( \vartheta_z \) denotes the phase of the zero, then there exists a pole with phase \( \vartheta_p \) such that \(|\vartheta_z - \vartheta_p| \leq 0.06\). Then, for each estimator we evaluate the reconstruction error which is defined as:

\[
e = \frac{\int_{-\pi}^{\pi} ||\Phi(e^{j\vartheta}) - \Phi(e^{j\vartheta})||^2 d\vartheta}{\int_{-\pi}^{\pi} ||\Phi(e^{j\vartheta})||^2 d\vartheta}
\] (46)

where \( \Phi \) denotes the true spectrum. Figure 3 shows the
Fig. 3: Boxplots of the reconstruction error of the spectral estimators in the Monte Carlo experiment. Red crosses represent the outliers.

boxplots of the reconstruction error. As we can see, ME-DI and ME-TC provide the best performance while the PEM methods provide the worst performance. Finally, in all the runs the kernel based PEM methods returned \( \hat{b}(z) \) with zeros inside the unit circle, so that we did not need to use the refined procedure in [26] to approximate \( \hat{b}(z) \) with a minimum phase filter. However choosing the poles of \( w(z) \) with absolute value equal to 0.995, the kernel based PEM methods sometimes returned \( \hat{b}(z) \) non-minimum phase while our methods always returned \( \hat{b}(z) \) minimum phase.

VII. CONCLUSION

In this paper we have introduced a kernel based ME spectral estimator. We have proposed two kernel matrices embedding the property that the impulse response of the one-step ahead predictor should have: regularity and exponential decay to zero. Moreover, we showed that the estimated spectrum is given by solving a generalized version of the Yule-Walker equation which leads also a minimum phase spectral factor of the process. Finally, numerical experiments showed the effectiveness of the proposed method in respect to the ME spectral estimator and kernel based PEM methods.

REFERENCES

[1] C. Byrnes, T. T. Georgiou, and A. Lindquist, “A new approach to spectral estimation: A tunable high-resolution spectral estimator,” IEEE Transactions on Signal Processing, vol. 48, no. 11, pp. 3189–3205, 2000.

[2] T. T. Georgiou and A. Lindquist, “Kullback-leibler approximation of spectral density functions,” IEEE Transactions on Information Theory, vol. 49, no. 11, pp. 2910–2917, 2003.

[3] A. Ferrante, M. Pavon, and F. Ramponi, “Hellingier versus kullback-leibler multivariable spectrum approximation,” IEEE Transactions on Automatic Control, vol. 53, no. 4, pp. 954–967, 2008.

[4] A. Ferrante, C. Masiero, and M. Pavon, “Time and spectral domain relative entropy: A new approach to multivariate spectral estimation,” IEEE Transactions on Automatic Control, vol. 57, no. 10, pp. 2561–2575, 2012.

[5] M. Zorzi and A. Ferrante, “On the estimation of structured covariance matrices,” Automatica, vol. 48, no. 9, pp. 2145–2151, 2012.

[6] B. Zhu and G. Baggio, “On the existence of a solution to a spectral estimation problem à la byrnes–georgiou–lindquist,” IEEE Transactions on Automatic Control, vol. 64, no. 2, pp. 820–825, 2018.

[7] M. Zorzi, “Rational approximations of spectral densities based on the alpha divergence,” Math. Control Signals Systems, vol. 26, no. 2, pp. 259–278, 2014.

[8] M. Pavon and A. Ferrante, “On the geometry of maximum entropy problems,” SIAM Review, vol. 55, no. 3, pp. 415–439, 2013.

[9] T. T. Georgiou, “Relative entropy and the multivariable multidimensional moment problem,” IEEE Transactions on Information Theory, vol. 52, no. 3, pp. 1052–1066, 2006.

[10] A. Ringh, J. Karlsson, and A. Lindquist, “Multidimensional rational covariance extension with approximate covariance matching,” SIAM Journal on Control and Optimization, vol. 56, no. 2, pp. 913–944, 2018.

[11] A. Ringh, J. Karlsson, and A. Lindquist, “Multidimensional rational covariance extension with applications to spectral estimation and image compression,” SIAM Journal on Control and Optimization, vol. 54, no. 4, pp. 1950–1982, 2016.

[12] E. Avventi, A. Lindquist, and B. Wahlberg, “ARMA identificaion of graphical models,” IEEE Transactions on Automatic Control, vol. 58, no. 5, pp. 1167–1178, 2012.

[13] D. Alpago, M. Zorzi, and A. Ferrante, “Identification of sparse reciprocal graphical models,” IEEE Control Systems Letters, vol. 2, pp. 659–664, Oct 2018.

[14] M. Zorzi, “Empirical Bayesian learning in AR graphical models,” Automatica, vol. 109, p. 108516, 2019.

[15] M. Zorzi, “Graphical model selection for a particular class of continuous-time processes,” Kybernetika, vol. 55, no. 5, pp. 782–801, 2019.

[16] M. Zorzi, “A new family of high-resolution multivariate spectral estimators,” IEEE Transactions on Automatic Control, vol. 59, no. 4, pp. 892–904, 2014.

[17] M. Zorzi, “Multivariate spectral estimation based on the concept of optimal prediction,” IEEE Transactions on Automatic Control, vol. 60, no. 6, pp. 1647–1652, 2015.

[18] M. Zorzi, “An interpretation of the dual problem of the THREE-like approaches,” Automatica, vol. 62, pp. 87–92, 2015.

[19] T. J. Ulrych and T. N. Bishop, “Maximum entropy spectral analysis and autoregressive decomposition,” Reviews of Geophysics, vol. 13, no. 1, pp. 183–200, 1975.

[20] H. Akaike, “A new look at the statistical model identification,” IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716–723, 1974.

[21] G. Schwarz et al., “Estimating the dimension of a model,” The annals of statistics, vol. 6, no. 2, pp. 461–464, 1978.

[22] P. Stoica and A. Nehorai, “On stability and root location of linear prediction models,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 35, no. 6, pp. 582–584, 1987.

[23] G. Bottegal and G. Pillonetto, “Regularized spectrum estimation using stable spline kernels,” Automatica, vol. 49, no. 11, pp. 3199–3209, 2013.

[24] G. Pillonetto, A. Chiuso, and G. D. Nicolao, “Prediction error approaches,” Automatica, no. 47, pp. 291–305, 2011.

[25] T. Chen, H. Ohlsson, and L. Ljung, “On the estimation of transfer functions, regularizations and gaussian processes revisited,” Automatica, vol. 48, no. 8, pp. 1525–1535, 2012.

[26] G. Pillonetto and A. Chiuso, “Identification of stable linear systems via the sequential stabilizing spline algorithm,” IFAC-PapersOnLine, vol. 51, no. 15, pp. 19–24, 2018.

[27] G. C. Goodwin, M. Gevers, and B. Ninness, “Quantifying the error in estimated transfer functions with application to model order selection,” IEEE Transactions on Automatic Control, vol. 37, no. 7, pp. 913–928, 1992.

[28] J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning. Springer series in statistics New York, 2001.

[29] N. G. De Bruijn, Asymptotic methods in analysis, vol. 4. Courier Corporation, 1981.

[30] T. Chen, M. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto, “System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques,” IEEE Transactions on Automatic Control, 2014.

[31] R. J. Tibshirani, Degrees of freedom and model search, Statistica Sinica, pp. 1265–1296, 2015.