IEEE copyright notice

POLYNOMIAL LMMSE ESTIMATION: A CASE STUDY

Stefan Uhlich, Benedikt Loesch and Bin Yang

Chair of System Theory and Signal Processing
Universität Stuttgart, Germany

ABSTRACT

This paper investigates the potential of the polynomial LMMSE estimation for nonlinear/nongaussian estimation problems. This is done by a case study: the estimation of the frequency of a sinusoidal signal with unknown amplitude and phase. We give analytical formulas to calculate the second order moments which are needed for the polynomial LMMSE estimation and we study the performance for varying orders of observations. Variable selection is used to identify the most relevant observations. It turns out that only a small number (less than one percent) of all available variables gives nearly the same mean squared error.

Index Terms— Linear MMSE estimation, Frequency estimation, Variable selection

1. INTRODUCTION

Linear minimum mean squared error (LMMSE) estimation is widely used in estimation theory. There are two reasons for this: One is that knowing the second order moments allows us to directly state the LMMSE estimator since a closed form solution exists (Wiener-Hopf equation). The other reason is its simple implementation and evaluation which are desirable for online estimation problems. However, the LMMSE estimator has the drawback that it will perform poorly if the estimation problem is nonlinear and/or nongaussian noise is present. One natural extension of the LMMSE principle for such problems is the use of a polynomial LMMSE (PLMMSE) estimator. It introduces an augmented observation vector that includes also higher order products of the available observations, see e.g. [1–3]. Similar to a Taylor series expansion, one hopes to obtain a good estimator with such a truncated Volterra series expansion. Especially for the case that there is no physical model for the nonlinearity, PLMMSE is still capable of modeling the nonlinearity. However, it is often not clear how large the required polynomial order should be to achieve a certain performance.

The aim of this paper is to investigate the potential of the PLMMSE estimator for a nonlinear estimation problem. As a case study, we investigate the task of estimating the frequency of a sinusoidal signal with unknown amplitude and phase in additive white Gaussian noise [4–7]. We choose this example since this nonlinear estimation problem allows us to give analytically the second order moments of the elements of the augmented observation vector instead of estimating them. Consequently, any limitation of the estimator performance is due to the PLMMSE approach and not to an uncertainty in the correlation matrices.

It is not our intention to show that the performance of PLMMSE is close to that of a maximum likelihood (ML) estimation which is easily feasible in this case. The former does not need any signal model while the latter relies on a perfect probabilistic signal model. We use this example only to study several issues of PLMMSE. Particularly, we would like to answer the following questions:

- How large is the performance gain by incorporating additional higher order observations?
- How does the condition number of the autocorrelation matrix depend on the polynomial order? What is the maximum polynomial order for a certain computation precision?
- How can we identify the relevant observations? How many relevant observations are there?

Following notations are used throughout this paper: $\underline{x}$ denotes a column vector, $\mathbf{X}$ a matrix and in particular $\mathbf{I}$ the identity matrix. The Kronecker product, trace operator, matrix transpose and euclidean norm are denoted by $\otimes$, $\mathrm{tr}\{.\}$, $(.)^T$ and $\|.\|$, respectively. $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ is the binomial coefficient.

2. POLYNOMIAL LMMSE ESTIMATION

The LMMSE estimation principle is well-known in signal processing, see e.g. [8]. It has the minimum mean squared error (MSE) among all possible linear estimators. Let $\underline{\theta}$ be the real-valued, unknown parameter vector that should be estimated from the observations $\underline{x} = \begin{bmatrix} x_1, \ldots, x_N \end{bmatrix}^T \in \mathbb{R}^N$. Minimizing the MSE $J = \mathrm{E}\|\underline{\theta} - \hat{\underline{\theta}}\|^2 = \mathrm{E}\|\underline{\theta} - \mathbf{W}\underline{x}\|^2$ yields the LMMSE estimator

$$\hat{\underline{\theta}} = \mathbf{W}\underline{x} = \mathbf{R}_{\theta x}\mathbf{R}_{xx}^{-1}\underline{x} \tag{1}$$

where $\mathbf{R}_{\theta x} = \mathrm{E}[\underline{\theta}\,\underline{x}^T]$ is the cross-correlation matrix and $\mathbf{R}_{xx} = \mathrm{E}[\underline{x}\,\underline{x}^T]$ the (auto-)correlation matrix. The minimum MSE for (1) is $J_{\min} = \mathrm{E}\|\underline{\theta}\|^2 - \mathrm{tr}\{\mathbf{R}_{\theta x}\mathbf{R}_{xx}^{-1}\mathbf{R}_{\theta x}^T\}$. Compared to the MMSE estimator, the LMMSE estimator will be clearly suboptimal in general. In [9], Balakrishnan analyzed the problem how to identify that the MMSE estimator is of a specific (polynomial) form. If we have, for example, a joint Gaussian probability density function $p(\underline{\theta}, \underline{x})$, then the optimum MMSE estimator is linear in $\underline{x}$ and therefore equivalent to the LMMSE estimator. This is known as the Bayesian Gauss-Markov theorem [8].

However, in most cases the LMMSE will have a limited performance. One possibility to overcome this is the use of an augmented observation vector $\underline{y}$ instead of $\underline{x}$ which also includes products of elements in $\underline{x}$. If we e.g. include all elements up to the quadratic terms of $\underline{x}$, then $\underline{y}$ has the form $\underline{y} = \begin{bmatrix} 1 & \underline{x}^T & \mathrm{Rem}\{\underline{x} \otimes \underline{x}\}^T \end{bmatrix}^T$ where $\mathrm{Rem}\{.\}$ is an operator that removes all redundant elements of its argument, i.e.

$$\mathrm{Rem}\left\{\begin{bmatrix} \cdots & x_1x_2 & x_2x_1 & \cdots \end{bmatrix}^T\right\} = \begin{bmatrix} \cdots & x_1x_2 & \cdots \end{bmatrix}^T.$$

The highest sum of exponents for a product that occurs in the augmented observation vector $\underline{y}$ is denoted as polynomial order D. In [1–3], this idea was used with $D = 2$ and is called a linear-quadratic or polynomial estimator.

In general, $\underline{y}$ has the form

$$\underline{y} = \begin{bmatrix} 1 & \underline{x}^T & \mathrm{Rem}\{\underline{x} \otimes \underline{x}\}^T & \mathrm{Rem}\{\underline{x} \otimes \underline{x} \otimes \underline{x}\}^T & \cdots \end{bmatrix}^T \tag{2}$$

and (1) for the PLMMSE becomes $\hat{\underline{\theta}} = \mathbf{W}\underline{y} = \mathbf{R}_{\theta y}\mathbf{R}_{yy}^{-1}\underline{y}$ with

978-1-4244-2710-9/09/$25.00

$\mathbf{R}_{\theta y} = \mathrm{E}[\theta \underline{y}^T]$ and $\mathbf{R}_{yy} = \mathrm{E}[\underline{y}\,\underline{y}^T]$. The idea of PLMMSE is to obtain a better estimator with higher orders of $\underline{x}$. The length of the augmented observation vector $\underline{y}$ is $L = \sum_{d=0}^{D} \binom{N+d-1}{d}$ which quickly increases with D and N. Therefore, D and N can only take moderate values. In the next section, we introduce the frequency estimation problem we will investigate as a case study.

3. PLMMSE FREQUENCY ESTIMATION

To evaluate the performance of the PLMMSE approach, we consider the following frequency estimation problem: Given are observations $x_1, \ldots, x_N$ that stem from the signal model

$$x_n = A\cos(\theta n + \phi) + z_n, \quad n = 1, \ldots, N \tag{3}$$

where A is an unknown amplitude, θ an unknown frequency and ϕ an unknown phase. They are assumed to be uniformly distributed random variables with $A \sim \mathcal{U}(A_{\min}, A_{\max}), \theta \sim \mathcal{U}(\theta_{\min}, \theta_{\max}),\ (0 \leq \theta_{\min} < \theta_{\max} \leq \pi)$ and $\phi \sim \mathcal{U}(-\pi, \pi)$. z_n is a white Gaussian, zero-mean noise process with variance σ^2. A, θ, ϕ and z_n are all independent of each other. Note that this model also includes the case of no a priori information about θ as we can set $\theta_{\min} = 0$ and $\theta_{\max} = \pi$.

We are only interested in the PLMMSE estimation of the angular frequency θ as A and ϕ can be easily estimated if the frequency is known [8]. The PLMMSE estimator for this problem is $\hat{\theta} = \underline{w}^T \underline{y}$ where $\underline{w}$ is the solution of the Wiener-Hopf equation $\underline{w}^T \mathbf{R}_{yy} = \underline{r}_{\theta y}^T$. To calculate the correlation matrix $\mathbf{R}_{yy}$ and the cross-correlation row vector $\underline{r}_{\theta y}^T = \mathrm{E}[\theta \underline{y}^T]$, we need higher order moments of the form

$$\mathrm{E}[x_{n_1}^{m_1} x_{n_2}^{m_2} \ldots x_{n_k}^{m_k}] \text{ and } \mathrm{E}[\theta x_{n_1}^{m_1} x_{n_2}^{m_2} \ldots x_{n_k}^{m_k}] \tag{4}$$

where all n_i and n_j are pairwise different time instances. In the appendix, a formula is given to calculate these higher order moments. It turns out that all moments in (4) are zero if $m = m_1 + \ldots + m_k$ is odd. This is due to the symmetric probability density function of z_n and the integration over ϕ.

Because (4) are zero for all odd numbers of m, $\underline{w}^T \mathbf{R}_{yy} = \underline{r}_{\theta y}^T$ can be written as a set of two independent linear equation systems. Since $\underline{r}_{\theta y}^T = \begin{bmatrix} b_0 & \underline{0}^T & \underline{b}_2^T & \underline{0}^T & \cdots \end{bmatrix}$ and

$$\mathbf{R}_{yy} = \mathrm{E}[\underline{y}\,\underline{y}^T] = \begin{bmatrix} A_{00} & \underline{0}^T & \underline{A}_{20}^T & \underline{0}^T & \cdots \\ \underline{0} & \mathbf{A}_{11} & \mathbf{0} & \mathbf{A}_{31}^T & \cdots \\ \underline{A}_{20} & \mathbf{0} & \mathbf{A}_{22} & \mathbf{0} & \cdots \\ \underline{0} & \mathbf{A}_{31} & \mathbf{0} & \mathbf{A}_{33} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

where $\underline{y}$ has the form of (2), these two linear equation systems are

$$\begin{bmatrix} w_0 & \underline{w}_2^T & \cdots \end{bmatrix} \begin{bmatrix} A_{00} & \underline{A}_{20}^T & \cdots \\ \underline{A}_{20} & \mathbf{A}_{22} & \cdots \\ \vdots & \vdots & \end{bmatrix} = \begin{bmatrix} b_0 & \underline{b}_2^T & \cdots \end{bmatrix} \tag{5a}$$

for the even part and

$$\begin{bmatrix} \underline{w}_1^T & \underline{w}_3^T & \cdots \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{31}^T & \cdots \\ \mathbf{A}_{31} & \mathbf{A}_{33} & \cdots \\ \vdots & \vdots & \end{bmatrix} = \begin{bmatrix} \underline{0}^T & \underline{0}^T & \cdots \end{bmatrix} \tag{5b}$$

for the odd part where $\underline{w}^T = \begin{bmatrix} w_0 & \underline{w}_1^T & \underline{w}_2^T & \underline{w}_3^T & \cdots \end{bmatrix}$. Eq. (5b) has the solution $\underline{w}_1 = \underline{0}, \underline{w}_3 = \underline{0}, \ldots$ and all odd order elements of the augmented observation vector $\underline{y}$ have no contribution to the PLMMSE estimation. Hence, they are neglected below in our case study.

4. EVALUATION OF THE PLMMSE

In this section, we evaluate the performance of the PLMMSE estimator. We consider the following three important issues: *Estimator risk*, *condition number* and *variable selection* corresponding to the three questions in Sec. 1.

Poly. order D	Length of $\underline{y}$	$\kappa(\mathbf{R}_{yy})$ for (a)	$\kappa(\mathbf{R}_{yy})$ for (b)
2	56	1.48×10^{4}	7.07×10^{1}
4	771	5.87×10^{8}	3.36×10^{3}
6	5776	3.04×10^{13}	2.53×10^{5}
8	30086	2.56×10^{18}	2.93×10^{7}

Table 1. Condition number κ of $\mathbf{R}_{yy}$

4.1. Estimator Risk

The risk of an estimator derived from a quadratic loss function is [10] $R(\theta) = \mathrm{E}_{\underline{x}|\theta}[(\theta - \hat{\theta}(\underline{x}))^2] = \int (\theta - \hat{\theta}(\underline{x}))^2 p(\underline{x}|\theta) d\underline{x}$. It is the quadratic loss $(\theta - \hat{\theta})^2$ averaged over the distribution of the measurements $\underline{x}$ conditioned on a fixed θ. Fig. 1 and Fig. 2 show a contour plot of the estimator risk in dB as a function of the true angular frequency θ and the signal-to-noise ratio (SNR) $\frac{\mathrm{E}[A^2]}{2\sigma^2}$ for $N = 10$ observations and varying polynomial order D with $\theta_{\min} = 0, \theta_{\max} = \pi$. We choose the amplitude of the sinusoid to vary uniformly within $A_{\min} = 1$ and $A_{\max} = 10$. Note that the PLMMSE estimator with $D = 0$ corresponds to the a priori estimator $\hat{\theta} = \mathrm{E}[\theta] = \frac{\pi}{2}$. This is also the PLMMSE estimate for a low SNR and we included it in both plots in dotted lines. The results show that an increase of D mostly reduces the estimator risk as expected. However, we can also see that the improvement becomes smaller with increasing polynomial order.

For comparison, Fig. 2 shows the performance of the ML algorithm [8] for the same input data. To maximize the log-likelihood function for all three unknowns A, θ and ϕ, we first performed a grid search within $[A_{\min} \ldots A_{\max}, \theta_{\min} \ldots \theta_{\max}, -\pi \ldots \pi]$. The found maximum is then used as initial value for the second step which uses the Newton algorithm to find the maximum. Clearly, the ML performs much better as it has a broader valley which is important for frequency estimation. Additionally, we need to know the SNR for the calculation of the correlation matrix $\mathbf{R}_{yy}$ and the cross-correlation vector $\underline{r}_{\theta y}^T$. The ML estimator does not exploit this knowledge.

It is obvious that the ML estimator for this nonlinear problem performs better than the PLMMSE estimator as expected. However, this comparison is not fair since the ML estimation assumes to know the probabilistic signal model perfectly. This is not necessary for PLMMSE that only needs to estimate $\mathbf{R}_{yy}$ and $\mathbf{R}_{\theta y}$ from the observations. The price for the lack of knowledge about the signal model is a worse estimation performance.

4.2. Condition Number

To obtain the coefficient vector $\underline{w}$, we have to solve the Wiener-Hopf equation $\underline{w}^T \mathbf{R}_{yy} = \underline{r}_{\theta y}^T$. The condition number κ of $\mathbf{R}_{yy}$ gives a bound on the accuracy of the solution $\underline{w}$ we obtain. Table 1 gives the values of κ for two cases: (a) the same signal as in Sec. 4.1 and (b) exact as in (a) except that $A = 1$ is now deterministic.

In both cases, the condition number of $\mathbf{R}_{yy}$ increases for an increasing polynomial order D. In particular, in case (a) and if $D \geq 8$, we can no longer trust the results even using double precision floating-point calculation because $\mathbf{R}_{yy}$ is ill conditioned. This contradicts the observation from Sec. 4.1 that we need a large value D to obtain a satisfactory MSE. The reason for a large condition number κ is the strong correlation, i.e. information redundancy among some of the elements of the augmented observation vector $\underline{y}$. This redundancy motivates variable selection which we consider in the next section.

4.3. Variable Selection

The number of elements L in the augmented observation vector $\underline{y}$ increases rapidly with an increase in the polynomial order D or the number of observations N. Therefore, the advantage of the LMMSE

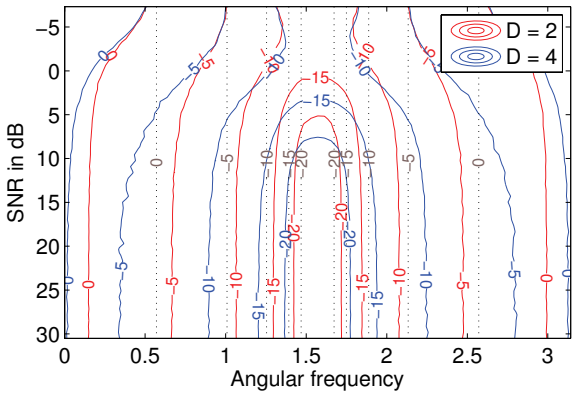

Fig. 1. Estimator risk in dB for $D = 0, 2, 4$ with $N = 10$

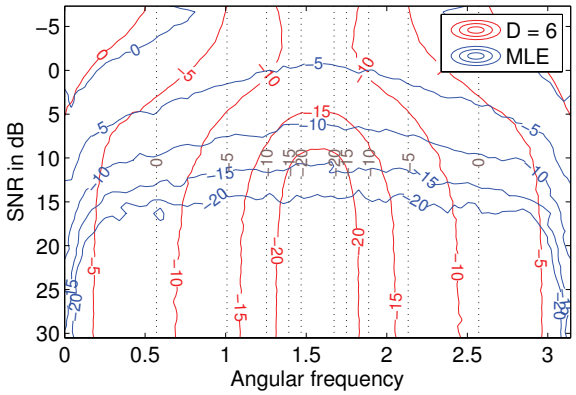

Fig. 2. Estimator risk in dB for $D = 0, 6$ and ML with $N = 10$

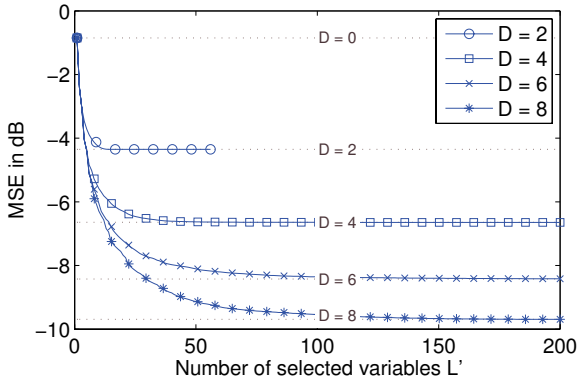

Fig. 3. MSE in dB after using SFFS variable selection

estimator, namely its simple calculation, is lost if D or N are large. Also the condition number of the correlation matrix $\mathbf{R}_{yy}$ limits the maximal value of D. On the other hand, only a polynomial order that is large enough enables us to obtain an acceptable MSE. A solution to this problem is to use only the relevant elements in $\underline{y}$ because not all elements in $\underline{y}$ have the same contribution to the PLMMSE estimation. The key step to identify these relevant elements in $\underline{y}$ is called variable selection. It allows us to reduce the number of used elements in $\underline{y}$ and still retain a good estimator performance.

The task of variable selection for our problem can be stated as follows: Use a reduced-size augmented observation vector $\underline{y}' = \mathbf{P}\underline{y} \in \mathbb{R}^{L'}$ and find the optimal selection matrix $\mathbf{P}_{\text{opt}} \in \mathbb{R}^{L' \times L}$ $(L' < L)$ among all possible selection matrices $\mathbf{P} \in \mathbb{R}^{L' \times L}$ of the same size. The selection matrices $\mathbf{P}$ reduce the number of variables from L to L' and are composed of the unit row vectors of the surviving variables. The corresponding MSE for a particular $\mathbf{P}$ is

$$J_{\min}(\mathbf{P}) = \mathrm{E}\|\underline{\theta}\|^2 - \mathrm{tr}\{\mathbf{R}_{\theta y}\mathbf{P}^T(\mathbf{P}\mathbf{R}_{yy}\mathbf{P}^T)^{-1}\mathbf{P}\mathbf{R}_{\theta y}^T\} \quad (6)$$

with $J_{\min}(\mathbf{P}_{\text{opt}}) \leq J_{\min}(\mathbf{P})$ for all $\mathbf{P}$ of the same size.

Beside an exhaustive search to find the optimum $\mathbf{P}_{\text{opt}}$, two different approaches can be distinguished. The first possibility is to use sequential variable selection. In each step, a variable is added to or removed from the active set. Probably the most prominent ones are *branch & bound* algorithms that exploit the monotonicity of $J_{\min}$ to find the optimal variable set [11, 12], Matching pursuit [13] and the *sequential floating forward selection* (SFFS) algorithm introduced in [14, 15]. The second class of approaches is to use a regularized least squares approach, as is done in [16] where the lasso is introduced.

We use SFFS to select the relevant observations in $\underline{y}$. It is well known that the SFFS has a good trade-off between its computational complexity and the quality of the selected variable subset [17]. In each iteration, a new variable is added to the subset of selected variables (forward step) and conditionally the least significant variables are excluded (backward step). Backward steps are applied as long as the resulting subset is better than the previous subset with the same number of variables. This conditional exclusion step can be motivated by the fact that each new included variable may carry information that was already present in other variables inside the selected subset. Thus, the old variables can be taken out and the redundancy is reduced without losing too much estimation performance. As the optimization criterion (6) involves the inversion of a matrix where only one row and column was added or deleted during a SFFS forward/backward step, the block-matrix inversion lemma can be used to compute $(\mathbf{P}\mathbf{R}_{yy}\mathbf{P}^T)^{-1}$ recursively as shown in [11].

Fig. 3 shows the result of the variable selection. The dotted lines indicate the mean squared error if all variables for a given polynomial order D are used. Clearly, we see that only a small number of variables is needed to achieve nearly the same MSE. For example, if $D = 8$, 200 out of 30086 variables (less than 1%) are sufficient. This indicates that most of the elements in the augmented vector $\underline{y}$ are redundant and that we can use variable selection to identify them.

5. CONCLUSION

The problem of estimating an unknown parameter using a polynomial LMMSE estimator was investigated in this paper. It is shown in a case study of frequency estimation that only a sufficiently large polynomial order allows a good estimation performance if we do not know the signal model at all. A high polynomial order, however, results in two problems: high computational complexity due to a large length of $\underline{y}$ and ill condition of $\mathbf{R}_{yy}$ due to strong correlation of some elements of $\underline{y}$. In order to combat these problems, the SFFS algorithm was considered. It is shown that the SFFS yields almost the same performance with only a small percentage of selected elements from $\underline{y}$.

As we have to learn $\mathbf{R}_{yy}$ and $\mathbf{R}_{\theta y}$ in practice, a second important case study is to investigate the influence of estimation errors on the overall MSE. Simulation results during our studies revealed that the higher order moments are more sensitive to the number of samples used to estimate the correlation matrices and thus also the number of available training data limits the maximum polynomial order. Another important issue, which we currently do not consider, is the comparison of the PLMMSE approach with other regression approaches, e.g. Kernel methods. We would like to investigate these two open points in a follow-up paper.

$$\mathrm{E}[\theta^b x_{n_1}^{m_1}\cdots x_{n_k}^{m_k}] = \mathrm{E}\left[\theta^b \prod_{i=1}^{k}\sum_{l_i=0}^{m_i}\binom{m_i}{l_i} A^{l_i}\cos^{l_i}(\theta n_i+\phi) z_{n_i}^{m_i-l_i}\right]$$
$$= \sum_{l_1=0}^{m_1}\cdots\sum_{l_k=0}^{m_k}\binom{m_1}{l_1}\cdots\binom{m_k}{l_k}\mathrm{E}\left[A^{l_1}\cdots A^{l_k}\right]\mathrm{E}\left[\theta^b\prod_{i=1}^{k}\cos^{l_i}(\theta n_i+\phi)\right]\mathrm{E}\left[z_{n_1}^{m_1-l_1}\right]\cdots\mathrm{E}\left[z_{n_k}^{m_k-l_k}\right] \quad (7)$$

$$\mathrm{E}\left[\theta^b\prod_{i=1}^{k}\cos^{l_i}(\theta n_i+\phi)\right] = \begin{cases} 0 & l \text{ odd} \\ \frac{1}{2^l}\underbrace{\sum_{r_1=0}^{l_1}\cdots\sum_{r_k=0}^{l_k}}_{\sum_{i=1}^{k}(2r_i-l_i)=0}\binom{l_1}{r_1}\cdots\binom{l_k}{r_k}\mathcal{S}_b\left(\sum_{i=1}^{k} n_i(2r_i-l_i)\right) & l \text{ even} \end{cases} \quad (11)$$

APPENDIX: HIGHER ORDER MOMENTS

For the calculation of the optimal weight vector $\underline{w}$, we need higher order moments of the form $\mathrm{E}[\theta^b x_{n_1}^{m_1} x_{n_2}^{m_2}\ldots x_{n_k}^{m_k}]$ with $m = m_1 + \cdots + m_k$ and $b \in \{0,1\}$. $b = 0$ gives the elements of $\mathbf{R}_{yy}$ and $b = 1$ the elements of $\underline{r}_{\theta y}^T$. Using the binomial formula $(\alpha+\beta)^m = \sum_{l=0}^{m}\binom{m}{l}\alpha^l\beta^{m-l}$, we obtain (7) which is shown at the top of this page. Note that we exploited the independence of A, θ, ϕ and z_{n_i} for all n_i with $i = 1,\ldots,k$.

After this expansion, we have a sum of terms in (7) that are composed of the following elements:

$$\mathrm{E}[z_{n_i}^{m_i-l_i}] = \begin{cases} 0 & m_i - l_i \text{ odd} \\ \frac{\sigma^{m_i-l_i}}{2^{(m_i-l_i)/2}}\frac{(m_i-l_i)!}{(\frac{m_i-l_i}{2})!} & m_i - l_i \text{ even} \end{cases} \quad (8)$$

$$\mathrm{E}\left[A^{l_1}\cdots A^{l_k}\right] = \mathrm{E}[A^l] = \frac{1}{l+1}\frac{A_{\max}^{l+1} - A_{\min}^{l+1}}{A_{\max} - A_{\min}} \quad (9)$$

where we used $l = l_1 + \cdots l_k$. In addition,

$$\mathrm{E}\left[\theta^b\prod_{i=1}^{k}\cos^{l_i}(\theta n_i+\phi)\right]$$
$$= \frac{1}{2^l}\mathrm{E}\left[\theta^b\prod_{i=1}^{k}\left(e^{j\theta n_i}e^{j\phi} + e^{-j\theta n_i}e^{-j\phi}\right)^{l_i}\right]$$
$$= \frac{1}{2^l}\mathrm{E}\left[\theta^b\prod_{i=1}^{k}\sum_{r_i=0}^{l_i}\binom{l_i}{r_i}e^{j\theta n_i(2r_i-l_i)}e^{j\phi(2r_i-l_i)}\right]$$
$$= \frac{1}{2^l}\mathrm{E}\left[\theta^b\sum_{r_1=0}^{l_1}\cdots\sum_{r_k=0}^{l_k}\binom{l_1}{r_1}\cdots\binom{l_k}{r_k}\times e^{j\theta\sum_{i=1}^{k} n_i(2r_i-l_i)}e^{j\phi\sum_{i=1}^{k}(2r_i-l_i)}\right]. \quad (10)$$

When we calculate the expectation in the last line of (10) with respect to ϕ, it will be zero if $\sum_{i=1}^{k}(2r_i - l_i) \neq 0$. This is the case for all elements in (10) if $l = l_1 + \cdots + l_k$ is odd. If l is even, there will be some $\sum_{i=1}^{k}(2r_i - l_i) = 0$ and the higher order moment is unequal to zero. Hence, (10) can be simplified to (11) shown at the top of this page where $\mathcal{S}_b(n) = \mathrm{Re}\{\mathrm{E}[\theta^b e^{j\theta n}]\}$ is given by

$$\mathcal{S}_0(n) = \begin{cases} 1 & n = 0 \\ \frac{\sin(\theta_{\max}n)-\sin(\theta_{\min}n)}{n(\theta_{\max}-\theta_{\min})} & \text{otherwise} \end{cases} \quad \text{and} \quad (12a)$$

$$\mathcal{S}_1(n) = \begin{cases} \frac{\theta_{\max}+\theta_{\min}}{2} & n = 0 \\ \frac{\cos(\theta_{\max}n)-\cos(\theta_{\min}n)}{n^2(\theta_{\max}-\theta_{\min})} + \frac{\sin(\theta_{\max}n)\theta_{\max}-\sin(\theta_{\min}n)\theta_{\min}}{n(\theta_{\max}-\theta_{\min})} & \text{otherwise} \end{cases}. \quad (12b)$$

REFERENCES

[1] B. Picinbono and P. Devaut, "Optimal linear-quadratic systems for detection and estimation," *IEEE Trans. Information Theory*, vol. 34, no. 2, pp. 304–311, 1988.

[2] P. Bondon, "Polynomial estimation of the amplitude of a signal," *IEEE Trans. Information Theory*, vol. 40, no. 3, pp. 960–965, 1994.

[3] C. Therrien and W. Jenkins, "New insights in the analysis of polynomial adaptive filters," *IEEE Digital Signal Processing Workshop Proceedings,*, pp. 382–385, 1996.

[4] D. C. Rife and R. R. Boorstyn, "Single-tone parameter estimation from discrete-time observations," *IEEE Trans. Information Theory*, vol. 20, no. 5, pp. 591–598, Sep. 1974.

[5] R. Kenefic and A. Nuttall, "Maximum likelihood estimation of the parameters of a tone using real discrete data," *IEEE Journal of Oceanic Engineering*, vol. 12, no. 1, 1987.

[6] H. W. Fung, A. C. Kot, K. H. Li, and K. C. Teh, "Parameter estimation of a real single tone from short data records," *Signal Processing*, vol. 84, no. 3, pp. 601–617, 2004.

[7] E. Jacobsen and P. Kootsookos, "Fast, accurate frequency estimators," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 123–125, 2007.

[8] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*. Prentice-Hall, 1993.

[9] A. Balakrishnan, "On a characterization of processes for which optimal mean-square systems are of specified form," *IEEE Trans. Information Theory*, vol. 6, no. 4, pp. 490– 500, 1960.

[10] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*. Addison-Wesley, 1990.

[11] P. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Computers*, vol. 26, no. 9, pp. 917–922, 1977.

[12] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 900–912, Jul. 2004.

[13] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[14] P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion," *Pattern Recognition - Conference B: Computer Vision*, vol. 2, pp. 279 – 283, Oct. 1994.

[15] P. Pudil, K. Fuka, K. Beranek, and P. Dvorak, "Potential of artificial intelligence based feature selection methods in regression models," *Computational Intelligence and Multimedia Applications*, pp. 159–163, 1999.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc., Series B.*, vol. 58, no. 1, pp. 267–288, 1996.

[17] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, Feb. 1997.