MODEL SELECTION VIA MISSPECIFIED CRAMÉR-RAO BOUND MINIMIZATION

Nadav E. Rosenthal and Joseph Tabrikian

School of Electrical and Computer Engineering, Ben Gurion University of the Negev
Email: rosenthn@post.bgu.ac.il, joseph@bgu.ac.il

ABSTRACT

In many applications of estimation theory, the true data model is unknown, and a set of parameterized models are used to approximate it. This problem is encountered in learning systems, where the assumed model parameters are estimated using training data. One of the challenges in these problems is choosing the architecture used for the approximated model. Complex and high-order models with limited training data size may lead to overfitting, while simple and low-order models may lead to model misspecification. In this paper, we propose to use the misspecified Cramér-Rao bound (MCRB) as a criterion for model selection. The MCRB takes into account modeling errors due to both overfitting and model misspecification. The performance of the proposed approach is evaluated via simulations for model order selection in a linear regression problem. The proposed method outperforms the minimum description length and the Akaike information criterion.

***Index Terms*—** Model order selection, misspecified Cramér-Rao bound (MCRB), model misspecification, overfitting, MDL, AIC

1. INTRODUCTION

In many applications of statistical signal processing, such as, communications, radar, sonar, and speech processing, the data model is not perfectly known. In many cases, machine learning techniques are used in order to approximate the true model by a set of assumed models, frequently, parameterized with unknown parameters to be determined using training data. Despite of the popularity of these techniques, no solid theory for model selection in learning problems is available. The problem of model and order selection has been extensively investigated in the literature, where several basic approaches have been implemented in various applications. However, in these works, the main objective is minimization of the probability of error rather than the mean-squared-error (MSE) obtained after model selection.

One of the main challenges in machine learning is how to choose the model. For example, in neural networks, the model complexity is determined by the network architecture, number of layers, number of neurons, activation function, and other design parameters. For training data with limited size, complex and high-order models lead to overfitting, which may result in poor performance of the network. On the other hand, when using simple and low-order models, modeling misspecification occurs, leading to large estimation errors. Both cases can result in estimation errors via bias and/or variance. Methods for model order selection include Akaike information criterion (AIC) [1], Bayesian information criterion (BIC) [2], and minimum description length (MDL) principle [3–9]. Under Bayesian framework, BIC and MDL asymptotically minimize the probability of error. However, the contributions of over-estimation and under-estimation to estimation MSE are not equal, and thus, optimal model order selection in terms of minimum MSE does not necessarily provide minimum probability of error. Currently, the main approach for model selection in learning systems is trial and error in which validation data are used for performance evaluation of a given architecture. The goal of this paper is to provide theoretical tools to address the problem of model selection for parameter estimation, where the criterion of interest is the MSE.

Under the non-Bayesian framework, the Cramér–Rao bound (CRB) [10, 11] is widely used for performance evaluation and system design. This bound was derived under the assumption of perfect model specification, where the data distribution is perfectly known. The CRB and other bounds, which assume perfect knowledge of the model cannot be used as an optimization criterion for model selection, since they do not take into account the contribution of model errors which are misspecified. In [12, 13] White studied the asymptotic properties of the maximum-likelihood estimator (MLE) under misspecified models where the data samples are statistically independent (see also [14]). Continuing this theory, the pioneering work of Richmond and Horowitz in [15, 16] studied the effects of modeling misspecification on the MSE of the MLE and derived the misspecified CRB (MCRB). The MCRB extends the CRB to misspecified scenarios, depicting its impact on the estimation bias and covariance. The MCRB was applied to various signal processing problems in [17–19], and extended to constrained problems in [20, 21].

In this paper, we present an approach for model selection, in which the MCRB for parameter estimation is computed for all the candidate models. The selected model is chosen to minimize the MCRB. A case study of model order selection in linear regression with additive Gaussian noise is considered to demonstrate the performance of the proposed approach.

2. PROBLEM STATEMENT AND DEFINITIONS

Let $\mathbf{x} \in \mathbb{R}^J$ be a J-dimensional random vector representing the observation vector, and $g(\mathbf{x};\boldsymbol{\varphi})$ denote the true probability density function (PDF) of $\mathbf{x}$, which belongs to a family $\mathcal{G}$ of distributions. $\boldsymbol{\varphi} \in \mathbb{R}^s$ is an unknown deterministic vector which is the parameter of interest. In cases where the true PDF $g(\mathbf{x};\boldsymbol{\varphi})$ is not known, it can be approximated with a parameterized PDF known up to a set of parameters to be determined in a learning stage. Let $\mathcal{O}^{(m)} = \{f^{(m)}_{\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)})} : \boldsymbol{\theta}^{(m)} \in \Omega^{(m)}\}$ be the mth model, where $m \in \mathcal{M}$. This model is defined by the set of parametric PDFs $f^{(m)}_{\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)})}$, with parameter space $\Omega^{(m)}$ and $\boldsymbol{\varphi}^{(m)} : \Omega^{(m)} \to \mathbb{R}^s$, a continuously differentiable mapping. Generally, m denotes the model index, which is not necessarily the model order. A model class is defined by the collection of models $\{\mathcal{O}^{(m)}\}_{m\in\mathcal{M}}$. Note that $g(\mathbf{x};\boldsymbol{\varphi})$ is not necessarily included in the model class, that is, there may not exist any

$m^* \in \mathcal{M}$ and $\boldsymbol{\theta}^{(m*)} \in \Omega^{(m*)}$ such that $g(\mathbf{x};\boldsymbol{\varphi}) = f^{(m*)}_{\boldsymbol{\varphi}^{(m*)}(\boldsymbol{\theta}^{(m*)})}$. For deriving an estimator of $\boldsymbol{\varphi}$, we approximate the data model by $f^{(m)}_{\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)})}$. Let $\boldsymbol{\varphi}^{(m)}(\hat{\boldsymbol{\theta}}^{(m)}(\mathbf{x}))$ be an estimator of $\boldsymbol{\varphi}$ with order m by $\mathbf{x}$, which for simplicity is referred by $\hat{\boldsymbol{\varphi}}^{(m)}$. This problem statement can be used to formulate a large class of learning systems, where the estimate of $\boldsymbol{\varphi}$ represents the output of the system and $\boldsymbol{\theta}^{(m)}$ stands for the machine or network parameters.

The estimation error of $\boldsymbol{\varphi}$ under the mth model is defined by:

$$\mathbf{e}(m) = \boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}^{(m)}, \tag{1}$$

and the corresponding mean-squared-error (MSE) matrix of $\boldsymbol{\varphi}$ with model order m is:

$$\mathbf{MSE}(m) = \mathrm{E}_g\left[\left(\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}^{(m)}\right)\left(\boldsymbol{\varphi} - \hat{\boldsymbol{\varphi}}^{(m)}\right)^T\right], \tag{2}$$

where $\mathrm{E}_g[\cdot]$ represents expectation with PDF $g(\mathbf{x};\boldsymbol{\varphi})$. In order to determine the optimal model, we will derive a lower bound on the MSE under each model and minimize it with respect to (w.r.t.) the model. We will use the MCRB for estimation of $\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)})$ $\forall m \in \mathcal{M}$. For the derivation of the MCRB, we assume that the regularity conditions used in [15, 17, 22, 23] are satisfied $\forall m \in \mathcal{M}$, i.e. for all the models in the class $\{\mathcal{O}^{(m)}\}_{m\in\mathcal{M}}$. Thus, the pseudo-true parameter vector $\boldsymbol{\theta}_0^{(m)}$ [17] is defined for every model order $m \in \mathcal{M}$. As described earlier, we intend to estimate $\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)})$. Therefore, the concept of misspecified (MS) unbiasedness of a function of $\boldsymbol{\theta}^{(m)}$ from Definition 3.1 in [23] is used here. The MCRB for estimation of continuously differentiable function of $\boldsymbol{\theta}^{(m)}$ was derived in Theorem 4.1 in [23] and will be used in the next section.

3. MODEL SELECTION VIA MCRB MINIMIZATION

In this section, we review the MCRB and use it for model selection. The MSE from (2) can be rewritten as:

$$\begin{aligned}\mathbf{MSE}(m) &= \mathrm{E}_g\left[\left(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)} + \boldsymbol{\varphi}_0^{(m)} - \hat{\boldsymbol{\varphi}}^{(m)}\right)\right.\\ &\quad\left.\cdot\left(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)} + \boldsymbol{\varphi}_0^{(m)} - \hat{\boldsymbol{\varphi}}^{(m)}\right)^T\right]\\ &= \left(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right)\left(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right)^T\\ &\quad + \left(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right)\mathrm{E}_g\left[\left(\boldsymbol{\varphi}_0^{(m)} - \hat{\boldsymbol{\varphi}}^{(m)}\right)^T\right]\\ &\quad + \mathrm{E}_g\left[\left(\boldsymbol{\varphi}_0^{(m)} - \hat{\boldsymbol{\varphi}}^{(m)}\right)\right]\left(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right)^T\\ &\quad + \mathrm{E}_g\left[\left(\hat{\boldsymbol{\varphi}}^{(m)} - \boldsymbol{\varphi}_0^{(m)}\right)\left(\hat{\boldsymbol{\varphi}}^{(m)} - \boldsymbol{\varphi}_0^{(m)}\right)^T\right]\end{aligned} \tag{3}$$

in which the first equality is derived by adding and subtracting the deterministic pseudo-true parameter vector $\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}_0^{(m)})$, referred by $\boldsymbol{\varphi}_0^{(m)}$ for simplicity. Assuming that $\hat{\boldsymbol{\varphi}}^{(m)}$ is a MS-unbiased estimator of $\boldsymbol{\varphi}_0^{(m)}$ [23], i.e.

$$\mathrm{E}_g\left[\hat{\boldsymbol{\varphi}}^{(m)} - \boldsymbol{\varphi}_0^{(m)}\right] = \mathbf{0}, \tag{4}$$

the mixed terms in (3) vanish. Thus, we obtain:

$$\begin{aligned}\mathbf{MSE}(m) &= \left(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right)\left(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right)^T\\ &\quad + \mathrm{E}_g\left[\left(\hat{\boldsymbol{\varphi}}^{(m)} - \boldsymbol{\varphi}_0^{(m)}\right)\left(\hat{\boldsymbol{\varphi}}^{(m)} - \boldsymbol{\varphi}_0^{(m)}\right)^T\right].\end{aligned} \tag{5}$$

For minimization of the MSE bound w.r.t. $m \in \mathcal{M}$, we define the following scalar function as a measure of performance:

$$tr\left(\mathbf{MSE}(m)\right) = \left\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right\|^2 + \mathrm{E}_g\left[\left\|\hat{\boldsymbol{\varphi}}^{(m)} - \boldsymbol{\varphi}_0^{(m)}\right\|^2\right]. \tag{6}$$

By substituting the MCRB from Theorem 4.1 in [23] for the second term, the trace of the MSE with order m is bounded by:

$$tr(\mathbf{MSE}(m)) \geq B^{(m)}, \tag{7}$$

where

$$B^{(m)} \triangleq \left\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right\|^2 + tr\left(\dot{\boldsymbol{\varphi}}_0^{(m)}\mathbf{MCRB}(\boldsymbol{\theta}_0^{(m)})\left(\dot{\boldsymbol{\varphi}}_0^{(m)}\right)^T\right) \tag{8}$$

with $\mathbf{MCRB}(\boldsymbol{\theta}_0^{(m)})$ as defined in [17] and the (p,q)th element of the matrix $\dot{\boldsymbol{\varphi}}_0^{(m)}$ are given by:

$$\left[\dot{\boldsymbol{\varphi}}_0^{(m)}\right]_{p,q} \triangleq \left.\frac{\partial\varphi_p^{(m)}(\boldsymbol{\theta}^{(m)})}{\partial\theta_q^{(m)}}\right|_{\boldsymbol{\theta}_0^{(m)}}. \tag{9}$$

The bound in (8) is composed of the norm squared of the bias and the trace of the lower bound on the covariance of $\hat{\boldsymbol{\varphi}}^{(m)}$.

We propose to use the MCRB as a criterion for model selection, i.e.

$$m^* = \arg\min_{m\in\mathcal{M}}\left\{B^{(m)}\right\}. \tag{10}$$

Note that $\forall m \in \mathcal{M}$, due to the miss-knowledge of $g(\mathbf{x};\boldsymbol{\varphi})$ and $\boldsymbol{\varphi}$, the MCRB used in (10) cannot be directly computed, and an approximation of the bound and the unknown parameters is required. We will address this problem for linear regression problems in the next section. After determining the model which minimizes the MCRB, the ML estimator for the parameter-of-interest can be obtained as follows:

$$\hat{\boldsymbol{\varphi}}_{ML}^{(m*)} = \boldsymbol{\varphi}^{(m*)}\left(\arg\max_{\boldsymbol{\theta}^{(m*)}}\left\{f^{(m*)}\left(\mathbf{x};\boldsymbol{\varphi}^{(m*)}\left(\boldsymbol{\theta}^{(m*)}\right)\right)\right\}\right). \tag{11}$$

4. TEST CASE: MODEL ORDER SELECTION IN LINEAR REGRESSION

Linear regression is commonly used in various applications of estimation theory. The observation vector $\mathbf{x}$ is assumed to depend on the regressor matrix $\mathbf{H}^{(m)}$ and the parameter vector $\boldsymbol{\theta}^{(m)}$, with additive noise vector $\mathbf{v}$ by:

$$\mathbf{x} = \mathbf{H}^{(m)}\boldsymbol{\theta}^{(m)} + \mathbf{v}, \tag{12}$$

where $\mathbf{H}^{(m)} \in \mathbb{R}^{J\times(m+1)}$, $\boldsymbol{\theta}^{(m)} \in \mathbb{R}^{m+1}$, $\mathbf{v} \in \mathbb{R}^J$ and $m \in \mathcal{M}$. In this problem setting, the model index, m, denotes the model order. The vector $\mathbf{v}$ is assumed to be Gaussian with known covariance matrix, $\sigma^2\mathbf{I}_J$, where $\mathbf{I}_J$ denotes the identity matrix of size J. We define the model class $\{\mathcal{O}^{(m)}\}_{m\in\mathcal{M}}$ with $\mathcal{M} = \{0,\ldots,M\}$ and the assumed PDF with order m as follows:

$$f^{(m)}(\mathbf{x};\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)})) = N(\mathbf{H}^{(m)}\boldsymbol{\theta}^{(m)}, \sigma^2\mathbf{I}_J), \tag{13}$$

where $\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)}) = \mathbf{H}^{(m)}\boldsymbol{\theta}^{(m)}$. The PDF of the data model is given by:

$$g(\mathbf{x};\mathbf{H}^{(P)}\boldsymbol{\theta}_t) = N(\mathbf{H}^{(P)}\boldsymbol{\theta}_t, \sigma^2\mathbf{I}_J). \tag{14}$$

The parameter of interest is defined with another regressor matrix $\tilde{\mathbf{H}}^{(P)}$ to predict $\boldsymbol{\varphi} = \tilde{\mathbf{H}}^{(P)}\boldsymbol{\theta}_t$, using the estimator $\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)}) = \tilde{\mathbf{H}}^{(m)}\boldsymbol{\theta}^{(m)}$.

Under the mth model, the contribution of the bias term in (8) is given by

$$\left\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0^{(m)}\right\|^2 = \left\|\tilde{\mathbf{H}}^{(P)}\boldsymbol{\theta}_t - \tilde{\mathbf{H}}^{(m)}\boldsymbol{\theta}_0^{(m)}\right\|^2, \tag{15}$$

where the pseudo-parameter of order m, which is defined as the parameter $\boldsymbol{\theta}_0^{(m)}$ that minimizes the KLD, is:

$$\boldsymbol{\theta}_0^{(m)} = \left(\mathbf{H}^{(m)T}\mathbf{H}^{(m)}\right)^{-1}\mathbf{H}^{(m)T}\mathbf{H}^{(P)}\boldsymbol{\theta}_t\ . \tag{16}$$

The MCRB for estimation of $\boldsymbol{\theta}^{(m)}$ [17] is given by:

$$\mathbf{MCRB}(\boldsymbol{\theta}_0^{(m)}) = \sigma^2\left(\mathbf{H}^{(m)T}\mathbf{H}^{(m)}\right)^{-1}, \tag{17}$$

and $\dot{\boldsymbol{\varphi}}_0^{(m)} = \tilde{\mathbf{H}}^{(m)}$. Note that for the regularity conditions to hold, $\mathbf{H}^{(m)T}\mathbf{H}^{(m)}$ is required to be non-singular for $m \in \mathcal{M}$, thus $J \geq M+1$.

Finally, the MSE bound under the mth model can be obtained by substitution of (15) and (17) into (8):

$$\begin{aligned} B^{(m)} &= \left\|\tilde{\mathbf{H}}^{(P)}\boldsymbol{\theta}_t - \tilde{\mathbf{H}}^{(m)}\boldsymbol{\theta}_0^{(m)}\right\|^2 \\ &\quad + tr\left(\tilde{\mathbf{H}}^{(m)}\sigma^2\left(\mathbf{H}^{(m)T}\mathbf{H}^{(m)}\right)^{-1}\tilde{\mathbf{H}}^{(m)T}\right). \end{aligned} \tag{18}$$

This bound is parameter-dependent. Thus, evaluation of the bound requires substitution of the unknown parameters with their estimate. Generally, $\tilde{\mathbf{H}}^{(P)}\boldsymbol{\theta}_t$ can be substituted using validation samples, and $\boldsymbol{\theta}_0^{(m)}$ can be substituted by the ML estimator based on training samples $\hat{\boldsymbol{\theta}}_{ML}^{(m)}(\mathbf{x})$, which satisfies $\mathrm{E}_g\left[\hat{\boldsymbol{\theta}}_{ML}^{(m)}(\mathbf{x}) - \boldsymbol{\theta}_0^{(m)}\right] = \mathbf{0}$.

5. NUMERICAL RESULTS

A simple example of the model presented in the previous section is polynomial regression, where $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_J]$ denotes the vector of time instances and the elements of $\mathbf{H}^{(m)}$ are defined by:

$$[\mathbf{H}^{(m)}]_{j,k+1} = \tau_j^k, \qquad k = 0, \ldots, m, \ \ j = 1, \ldots, J. \tag{19}$$

The vector $\boldsymbol{\theta}^{(m)}$ is composed of the polynomial coefficients. The true parameter $\boldsymbol{\theta}_t$ is unknown deterministic, where the true order P of the polynomial is unknown as well. We are interested in optimal estimation of polynomial extrapolation point, for an arbitrary time instance $\tau_{ext} \in \mathbb{R}$, given time instances $\boldsymbol{\tau}$ and observation vector $\mathbf{x}$. In this case, $\varphi = \tilde{\mathbf{h}}^{(P)T}\boldsymbol{\theta}_t$, where $[\tilde{\mathbf{h}}^{(P)}]_{k+1} = \tau_{ext}^k, \quad k = 0, \ldots, P$. Our proposed criterion for model selection and estimation is to retrieve m^* in (10) and estimate $\hat{\boldsymbol{\theta}}_{ML}^{(m*)}(\mathbf{x})$ with training samples $\{\mathbf{x}, \boldsymbol{\tau}\}$ at extrapolation point τ_{ext}, based on the model:

$$x_{ext} = \tilde{\mathbf{h}}^{(P)T}\boldsymbol{\theta}_t + v, \tag{20}$$

where v is a zero-mean Gaussian random variable with known variance σ^2. Then, the ML estimator of the extrapolation point φ is:

$$\varphi^{(m*)}\left(\hat{\boldsymbol{\theta}}_{ML}^{(m*)}(\mathbf{x})\right) = \tilde{\mathbf{h}}^{(m*)T}\hat{\boldsymbol{\theta}}_{ML}^{(m*)}(\mathbf{x}). \tag{21}$$

By using (15)-(18), we obtain:

$$\begin{aligned} B^{(m)} &= \left(\tilde{\mathbf{h}}^{(P)T}\boldsymbol{\theta}_t - \tilde{\mathbf{h}}^{(m)T}\boldsymbol{\theta}_0^{(m)}\right)^2 \\ &\quad + \sigma^2\tilde{\mathbf{h}}^{(m)T}\left(\mathbf{H}^{(m)T}\mathbf{H}^{(m)}\right)^{-1}\tilde{\mathbf{h}}^{(m)}. \end{aligned} \tag{22}$$

Since the bound in (22) depends on the unknown parameters $\boldsymbol{\theta}_t$, and P, an approximation is necessary. Unbiased estimators for $\tilde{\mathbf{h}}^{(P)T}\boldsymbol{\theta}_t, \boldsymbol{\theta}_0^{(m)}$ are given by x_{ext}, $\hat{\boldsymbol{\theta}}_{ML}^{(m)}(\mathbf{x})$, respectively, and an unbiased estimator of the bound is given by:

$$\hat{B}^{(m)} = \left(x_{ext} - \tilde{\mathbf{h}}^{(m)T}\hat{\boldsymbol{\theta}}_{ML}^{(m)}(\mathbf{x})\right)^2. \tag{23}$$

Thus, the criterion in (10) can be approximated as:

$$m^* = \arg\min_{m=0,\ldots,M}\left\{\hat{B}^{(m)}\right\}. \tag{24}$$

The proposed criterion will be compared to MDL [3–9], AIC [1], and AICc [8]:

$$MDL(m) = -2\hat{L}^{(m)} + (m+1)\log(J+1), \tag{25}$$

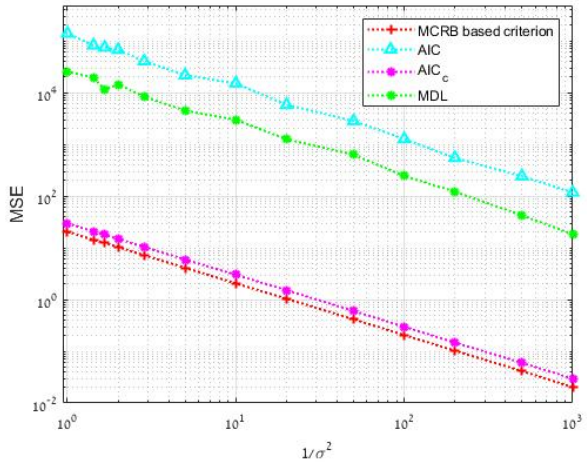
$$AIC(m) = -2\hat{L}^{(m)} + 2(m+1), \tag{26}$$

$$AIC_c(m) = -2\hat{L}^{(m)} + 2\frac{J+1}{(J-(m+1))}(m+1), \tag{27}$$

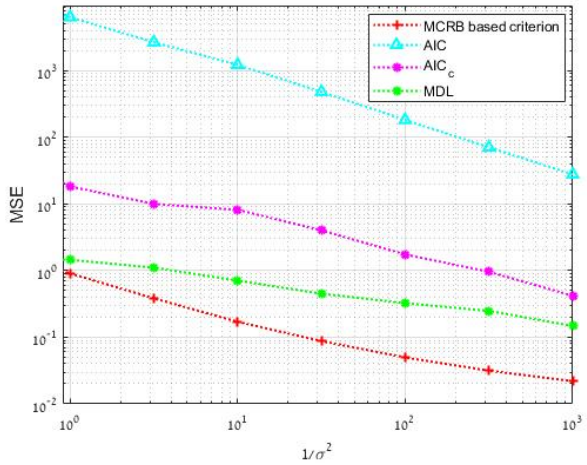
where $m+1$ is the number of unknown parameters, and $\hat{L}^{(m)}$ is the log-likelihood function for estimating $\boldsymbol{\varphi}^{(m)}(\boldsymbol{\theta}^{(m)})$ under model order m after maximization w.r.t. $\boldsymbol{\theta}^{(m)}$:

$$\hat{L}^{(m)} = -\frac{(J+1)}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\|\mathbf{P}^{(m)\perp}\tilde{\mathbf{x}}\right\|^2, \tag{28}$$

where $\mathbf{P}^{(m)\perp}$ denotes the projection matrix into the null space spanned by the columns of $\left[\mathbf{H}^{(m)}, \tilde{\mathbf{h}}^{(m)}\right]$. In order to conduct a fair comparison between the proposed criterion versus the criteria in (25)-(27), $\tilde{\mathbf{x}} = [\mathbf{x}^T, x_{ext}]^T$, and $\mathbf{P}^{(m)\perp}$ include the extrapolation sample for the considered criteria. The performances of existing model order selection methods are usually measured in terms of probability of correct decision. MDL is a consistent criterion, that is, its probability of error in model selection approaches zero as the number of samples goes to infinity. This criterion results in large MSE when the model is underestimated. Overestimation may also lead to large MSE. In contrast, our criterion will not necessarily detect the true model order, but it minimizes the evaluated bound on the MSE. In fact, when we try to approximate an arbitrary model with a set of nested candidate models, asymptotically the best model would be the one with the highest possible order. The time instances $[\tau_1, \ldots, \tau_J]$ were randomly generated from white Gaussian distribution with unit variance, and τ_{ext} was set to be $\min_j(\tau_j) - 0.1$. The performances of the considered algorithms were evaluated in terms of MSE using $K = 10^4$ realizations as a function of the noise variance inverse, $1/\sigma^2$, with the following parameters: case a: $\boldsymbol{\theta}_{t,a} = [0.8, -1.44, -0.74, -0.78, -1.8, 0.51, 1.21, 0.84]$, $P = 7, M = 12, J = 15$, case b: $\boldsymbol{\theta}_{t,b} = [0.12, 0.16, 0.18, -0.05, -0.08, -0.19, -0.08, -0.2, -0.07, -0.06, 0.1]$, $P = 10, M = 15, J = 50$. Fig. 1 presents the evaluated MSE using the considered criteria for the two cases as a function of the noise variance inverse. Fig. 2 presents the evaluated MSE for the considered criteria with case (a), $\sigma^2 = 0.1$, as a function of J. It can be seen that the proposed criterion outperforms the other tested criteria in terms of



(a)



(b)

Fig. 1. MSE for estimating φ using the considered criteria for the following parameter sets: (a) $P = 7, M = 12, J = 15, \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t,a}$, (b) $P = 10, M = 15, J = 50, \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t,b}$.

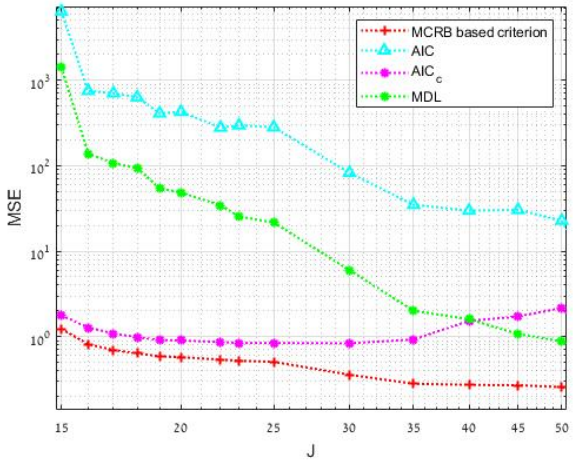


Fig. 2. MSE as a function on the number of samples for estimating φ using the considered criteria for case (a) and $\sigma^2 = 0.1$.

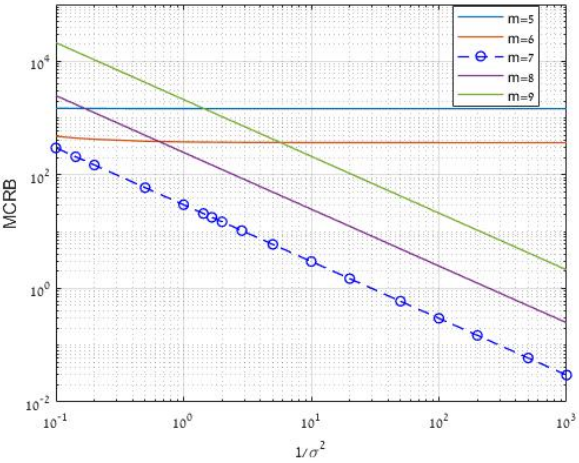


Fig. 3. MCRB for various model orders, m, with $P = 7, J = 15$.

MSE for various noise levels and various number of samples. Note that the evaluated MSE in case (b) is lower due to the dependency on the polynomial coefficients and the larger number of samples J. Fig. 3 presents the MSE bounds in (22) for $m = 5, \ldots, 9$ and $P = 7$ (case a). The norm squared of the bias (first term of (22)) has large values for $m < P$ (underfitting), and thus those bounds are large (independent of σ^2). For $m \geq P$ the bias term equals zero and the bound is minimal for $m = P$. The second term increases (like penalty of overfitting) as σ^2 increases. Dimensions of the matrix $\left(\mathbf{H}^{(m)T}\mathbf{H}^{(m)}\right)^{-1}$ are defined by m, and thus the bound increases with m.

6. CONCLUSION

In this paper, we proposed a new model selection approach for parameter estimation in learning systems. In this approach, the MCRB for parameter estimation based on a family of models is computed and minimized. The MCRB takes into account the effect of both overfitting and model misspecification. This approach was applied to model order selection in linear regression problems. In a polynomial fitting problem it was shown that this criterion outperforms the MDL and the AIC for model order selection.

7. REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, pp. 716–723, Dec. 1974.

[2] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, Mar. 1978.

[3] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, Sep. 1978.

[4] J. Rissanen, "Estimation of structure by minimum description

length," *Circuits Syst. Signal Process.*, vol. 1, pp. 395–406, Sep. 1982.

[5] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.

[6] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Amer. Statist. Assoc.*, vol. 96, pp. 746–774, June 2001.

[7] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques—an overview," *IEEE Signal Process. Mag.*, vol. 35, no 6, pp. 16–34, 2018.

[8] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no 4, pp. 36–47, 2004.

[9] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 279–290, 2004.

[10] H. Cramér, *Mathematical Methods of Statistics.*, Princeton, NJ, USA, Princeton Univ. Press, 1946.

[11] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, vol. 37, pp. 81–91, 1945.

[12] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, pp. 1–25, 1982.

[13] H. White, *Estimation, Inference and Specification Analysis.*, Cambridge University Press, 1996.

[14] Y. Noam and J. Tabrikian, "Marginal likelihood for estimation and detection theory," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 3963–3974, 2007.

[15] C. D. Richmond and L. L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification," *IEEE Trans. Signal Process.*, vol. 63, no 9, pp. 2263–2278, 2015.

[16] C. D. Richmond and L. L. Horowitz, "Parameter bounds under misspecified models," *Proc. Conf. Signals, Systems and Computers (Asilomar)*, pp. 176–180, 2013.

[17] S. Fortunati, F. Gini, and M.S. Greco, "The misspecified Cramér-Rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions," *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2387–2399, 2016.

[18] S. Fortunati, F. Gini, M.S. Greco, and C.D. Richmond, "Performance bounds for parameter estimation under misspecified models fundamental findings and applications," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 142–157, 2017.

[19] P. Wang, T. Koike-Akino, M. Pajovic, P. V. Orlik, W. Tsujita, and F. Gini, "Misspecified CRB on parameter estimation for a coupled mixture of polynomial phase and sinusoidal FM signals," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5302–5306.

[20] S. Fortunati, F. Gini, and M.S. Greco, "The constrained misspecified Cramér-Rao bound," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 718–721, 2016.

[21] S. Fortunati, "Misspecified Cramér–Rao bounds for complex unconstrained and constrained parameters," *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2017.

[22] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," *Proc. 5th Berkley Symp. Math. Statist. Probab*, p. 221–233, 1967.

[23] Q. H. Vuong, "Cramér–Rao bounds for misspecified models," *Div. of the Humanities and Social Sci., California Inst. of Technol., Pasadena, CA, USA*, 1986.