NEW IMPROVED CRITERION FOR MODEL SELECTION IN SPARSE HIGH-DIMENSIONAL LINEAR REGRESSION MODELS

Prakash B. Gohain, Magnus Jansson *

Division of Information Science and Engineering, KTH Royal Institute of Technology, Sweden
Email ids: pbg@kth.se, janssonm@kth.se

ABSTRACT

Extended Bayesian information criterion (EBIC) and extended Fisher information criterion (EFIC) are two popular criteria for model selection in sparse high-dimensional linear regression models. However, EBIC is inconsistent in scenarios when the signal-to-noise-ratio (SNR) is high but the sample size is small, and EFIC is not invariant to data scaling, which affects its performance under different signal and noise statistics. In this paper, we present a refined criterion called EBIC_R where the 'R' stands for robust. EBIC_R is an improved version of EBIC and EFIC. It is scale-invariant and a consistent estimator of the true model as the sample size grows large and/or when the SNR tends to infinity. The performance of EBIC_R is compared to existing methods such as EBIC, EFIC and multi-beta-test (MBT). Simulation results indicate that the performance of EBIC_R in identifying the true model is either at par or superior to that of the other considered methods.

Index Terms— High-dimensional inference, model selection, Lasso, OMP, sparse estimation, subset selection.

1. INTRODUCTION

In this paper, we study the model selection problem also known as best subset selection in high-dimensional linear regression models of the form $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where the number of samples, N, is quite small compared to the parameter dimension, p ($N \ll p$), $\mathbf{y} \in \mathbb{R}^N$ is the measurement vector, $\mathbf{A} \in \mathbb{R}^{N\times p}$ is the known design matrix, whose columns are referred to as predictors. $\mathbf{x} \in \mathbb{R}^p$ is the unknown regression vector. We denote the true support of $\mathbf{x}$ as $\mathcal{S} = \{i : x_i \neq 0\}$ and cardinality $card(\mathcal{S}) = k_0$. $\mathbf{x}$ is assumed to be sparse such that $k_0 \ll N$. $\mathbf{e} \in \mathbb{R}^N$ is the noise vector following $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$ where σ^2 is the true noise variance and $\mathbf{I}_N$ is an $N \times N$ identity matrix. The goal of model selection is estimating the true subset $\mathcal{S}$ given $\mathbf{y}$ and $\mathbf{A}$.

Model selection in general is a well studied subject and there exists a vast literature on this topic [1, 2, 3, 4]. However, in the last two decades, the model selection problem in high-dimensional cases has gained immense attention due to the complexities and challenges involved in it. In fact, for $p \gg N$ problems, most of the popular and traditional information theoretic methods such as Akaike information criterion (AIC) [5], Bayesian information criterion (BIC) [6] and minimum description length (MDL) [7] perform poorly and often lead to overfitted models with unwanted predictors [8, 9, 10].

In order to tackle this large-p small-N problem, the authors in [8] proposed an extended version of BIC called extended BIC (EBIC). Compared to BIC, EBIC assigns prior probability to a candidate model that is inversely proportional to the size of its model space. Thus, a model with larger dimension is assigned smaller prior probability as compared to a model with smaller dimension, which is in tune with the law of parsimony. EBIC is a consistent estimator of the true model as $N \to \infty$. However, as indicated in [10], the empirical performance of EBIC can sometimes be unsatisfactory for practical sizes of N. Moreover, in scenarios when N is fixed but $\sigma^2 \to 0$, it is shown that EBIC is inconsistent.

To overcome the consistency issue in EBIC for high-SNR scenarios, the authors in [10], proposed a novel criterion called extended Fisher information criterion (EFIC). It is inspired by EBIC and the model selection criteria with Fisher information [11]. EFIC is a consistent criterion, as $N \to \infty$ as well as when $\sigma^2 \to 0$. However, as shown in this paper, EFIC is not invariant to data scaling and this causes the behaviour of EFIC to become unstable when the data is scaled or equivalently under changing signal and noise statistics. This scaling problem is a result of the data dependent penalty design in EFIC that may blow the penalty to extremely small or large values depending on how the signal is scaled leading to higher overfitting or underfitting losses.

This paper presents an improved version of EBIC and EFIC that is invariant to data scaling and consistent as $N \to \infty$ and/or as $\sigma^2 \to 0$ (or equivalently SNR $\to \infty$). We call it EBIC_R, where the 'R' stands for robust. EBIC_R can be combined with different predictor selection algorithms such as greedy methods like orthogonal matching pursuit (OMP)[12] or shrinkage methods, e.g., least absolute shrinkage and selection operator (LASSO) [13] to perform model selection.

The rest of the paper is organized as follows: In Section

*This research was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 742648.

2 we provide the necessary background to motivate the new criterion. Section 3 presents the proposed criterion in detail. Section 4 shows the simulations results and Section 5 concludes the paper.

2. BACKGROUND

Consider the linear model under the following hypothesis

$$\mathcal{H}_{\mathcal{I}} : \mathbf{y} = \mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}} + \mathbf{e}_{\mathcal{I}}, \tag{1}$$

where $\mathcal{H}_{\mathcal{I}}$ denotes the hypothesis that the data $\mathbf{y}$ is truly generated according to (1), $\mathbf{A}_{\mathcal{I}} \in \mathbb{R}^{N\times k}$ is the sub-design matrix consisting of columns from the original design matrix $\mathbf{A}$ with support $\mathcal{I}$ and cardinality $card(\mathcal{I}) = k$. $\mathbf{x}_{\mathcal{I}} \in \mathbb{R}^k$ is the corresponding unknown regression vector and $\mathbf{e}_{\mathcal{I}} \in \mathbb{R}^N$ is the associated noise vector following $\mathbf{e}_{\mathcal{I}} \sim \mathcal{N}(0, \sigma_{\mathcal{I}}^2\mathbf{I}_N)$ where $\sigma_{\mathcal{I}}^2$ is the unknown noise variance corresponding to the hypothesis $\mathcal{H}_{\mathcal{I}}$. The maximum likelihood estimates (MLEs) of $\mathbf{x}_{\mathcal{I}}$ and $\sigma_{\mathcal{I}}^2$ under $\mathcal{H}_{\mathcal{I}}$ are obtained as [14]

$$\hat{\mathbf{x}}_{\mathcal{I}} = \left(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right)^{-1}\mathbf{A}_{\mathcal{I}}^T\mathbf{y} \quad \& \quad \hat{\sigma}_{\mathcal{I}}^2 = \mathbf{y}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}/N \tag{2}$$

where $\mathbf{\Pi}_{\mathcal{I}}^{\perp} = \mathbf{I}_N - \mathbf{A}_{\mathcal{I}}(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}})^{-1}\mathbf{A}_{\mathcal{I}}^T$ denotes the orthogonal projection matrix on the right null space of $\mathbf{A}_{\mathcal{I}}^T$.

To motivate the proposed criterion we start with the source i.e., the maximum a-posteriori (MAP) estimator. For the considered model selection problem, the MAP estimate of $\mathcal{S}$ is equivalently given as (see [15, 16] for details):

$$\begin{aligned}\hat{\mathcal{S}}_{\text{MAP}} = \arg\min_{\mathcal{I}} \big\{ & N\ln\hat{\sigma}_{\mathcal{I}}^2 + \ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| - k\ln 2\pi \\ & -2\ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) - 2\ln p(\mathcal{H}_{\mathcal{I}})\big\}.\end{aligned} \tag{3}$$

where $p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ is the prior probability for the model parameter vector $\hat{\boldsymbol{\theta}}_{\mathcal{I}} = [\hat{\mathbf{x}}_{\mathcal{I}}^T, \hat{\sigma}_{\mathcal{I}}^2]^T$, $p(\mathcal{H}_{\mathcal{I}})$ is the prior probability of the model with support $\mathcal{I}$, $|\cdot|$ denotes determinant, and $\hat{\mathbf{F}}_{\mathcal{I}}$ is the sample Fisher information matrix under $\mathcal{H}_{\mathcal{I}}$ given as [10]

$$\hat{\mathbf{F}}_{\mathcal{I}} = \begin{bmatrix} \frac{1}{\hat{\sigma}_{\mathcal{I}}^2}\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \frac{N}{2\hat{\sigma}_{\mathcal{I}}^4} \end{bmatrix}. \tag{4}$$

Both EBIC and EFIC can be derived from (3) under different assumptions. First, the $-k\ln 2\pi$ term is ignored as it weakly depends on k. Also, the $p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ term is dropped as it is considered to be uniform and uninformative. In EBIC, it is assumed that $\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| \approx k\ln N + \mathcal{O}(1)$ and $p(\mathcal{H}_{\mathcal{I}}) \propto \binom{p}{k}^{-\gamma}$ [8]. Thus, the EBIC score for a model with support $\mathcal{I}$ is

$$\text{EBIC}(\mathcal{I}) = N\ln\hat{\sigma}_{\mathcal{I}}^2 + k\ln N + 2\gamma\ln\binom{p}{k} \tag{5}$$

where $0 \le \gamma \le 1$ is a tuning parameter. In EFIC, the $\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right|$ term is fully retained. Further approximation under large-p include $\ln\binom{p}{k} \approx k\ln p$. The final form of the EFIC is [10]

$$\begin{aligned}\text{EFIC}(\mathcal{I}) = & N\ln\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 + k\ln N + \ln\left|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right| \\ & - (k+2)\ln\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 + 2ck\ln p\end{aligned} \tag{6}$$

where $c > 0$ is a tuning parameter.

3. PROPOSED METHOD EBIC$_{\text{R}}$

In this section, we present the necessary steps for deriving EBIC$_{\text{R}}$. An important assumption in this regard is the following property of the design matrix $\mathbf{A}$ [15, 17]

$$\lim_{N\to\infty}\left\{N^{-1}\left(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right)\right\} = \mathbf{M}_{\mathcal{I}} = \mathcal{O}(1), \tag{7}$$

where $\mathbf{M}_{\mathcal{I}}$ is a positive definite matrix that is bounded as $N \to \infty$. The above assumption is true in many applications but not all [15, 18]. Next, a similar approach as in [15] is considered, but here we perform normalization of $\hat{\mathbf{F}}_{\mathcal{I}}$ under both large-N and high-SNR assumption. Consider the following matrix

$$\mathbf{L}^{-1/2} = \begin{bmatrix} \sqrt{\frac{1}{N}}\sqrt{\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_0^2}}\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \sqrt{\frac{1}{N}\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_0^2}} \end{bmatrix}, \tag{8}$$

where $\hat{\sigma}_0^2 = \|\mathbf{y}\|_2^2/N$. The factor, $\hat{\sigma}_0^2$, is used in $\mathbf{L}^{-1/2}$ to neutralize the data scaling problem and is motivated by the fact that given (7), when the SNR is a constant, we have

$$\mathbb{E}[\hat{\sigma}_0^2] \to \text{const.} \quad \& \quad \text{Var}[\hat{\sigma}_0^2] \to 0 \tag{9}$$

as $N \to \infty$. Furthermore, from the considered generating model in (1), when N is fixed, (9) is also satisfied as $\sigma^2 \to 0$. Now using (4), (7) and (8) it is possible to show that

$$\left|\mathbf{L}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{L}^{-1/2}\right| = \begin{vmatrix} \frac{1}{\hat{\sigma}_0^2}\frac{\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}}{N} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\hat{\sigma}_0^4} \end{vmatrix} = \mathcal{O}(1). \tag{10}$$

Using (10), we can express the $\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right|$ in (3) as

$$\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| = \ln\left[\left|\mathbf{L}\right|\left|\mathbf{L}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{L}^{-1/2}\right|\right] = \ln|\mathbf{L}| + \mathcal{O}(1) \tag{11}$$

where $\mathcal{O}(1)$ is a term that is bounded as $N \to \infty$ and/or $\sigma_{\mathcal{I}}^2 \to 0$ and therefore may be discarded without much effect on the criterion. Expanding the $\ln|\mathbf{L}|$ term we have

$$\begin{aligned}\ln|\mathbf{L}| &= \ln\begin{vmatrix} N\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & N\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right)^2 \end{vmatrix} \\ &= (k+1)\ln N + (k+2)\ln\left(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2\right).\end{aligned} \tag{12}$$

Therefore, using (12) we can rewrite (11) as

$$\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| = k\ln N + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right) + \ln N + \mathcal{O}(1). \tag{13}$$

Using (3), (13), $-\ln p(\mathcal{H}_{\mathcal{I}}) \approx \zeta k\ln p$, and ignoring the $\mathcal{O}(1)$, $\ln N$ (independent of k) and the $p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ term we arrive at the modified criterion EBIC$_{\text{R}}$:

$$\begin{aligned}\text{EBIC}_{\text{R}}(\mathcal{I}) = N\ln\hat{\sigma}_{\mathcal{I}}^2 + k\ln\left(\frac{N}{2\pi}\right) + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right) \\ + 2\zeta k\ln p\end{aligned} \tag{14}$$

where $\zeta > 0$ is a tuning parameter. The term $\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2 \ge 1\ \forall\,\mathcal{I}$, therefore, $\ln(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2) \ge 0,\ \forall\,\mathcal{I}$. Hence, the penalty of EBIC$_{\text{R}}$ is a monotonic function of N, $\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2$ ($\approx$ SNR $+ 1$) and p.

3.1. Scaling robustness as compared to EFIC

Here, we discuss the scaling problem present in EFIC. Let $\Delta = k - k_0$. Now consider the difference assuming $\mathcal{I} \neq \mathcal{S}$

$$\text{EFIC}(\mathcal{I}) - \text{EFIC}(\mathcal{S}) = (N-2)\ln\left(\frac{\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2}{\left\|\mathbf{\Pi}_{\mathcal{S}}^{\perp}\mathbf{y}\right\|_2^2}\right) + \\ + \ln\left|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right| - \ln\left|\mathbf{A}_{\mathcal{S}}^T\mathbf{A}_{\mathcal{S}}\right| - k\ln\left(\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2\right) \\ + k_0\ln\left(\left\|\mathbf{\Pi}_{\mathcal{S}}^{\perp}\mathbf{y}\right\|_2^2\right) + \Delta\left(\ln N + 2c\ln p\right) = D_{\text{EFIC}}. \quad (15)$$

Ideally, for correct model selection, $D_{\text{EFIC}} > 0$ for all $\mathcal{I} \neq \mathcal{S}$. Now, if we scale the data $\mathbf{y}$ by a constant $C > 0$, the difference becomes

$$\text{EFIC}(\mathcal{I}) - \text{EFIC}(\mathcal{S}) = D_{\text{EFIC}} - \Delta\ln C^2. \quad (16)$$

It is clearly observed that (15) and (16) are unequal and the difference after scaling contains an additional term $-\Delta\ln C^2$. This implies that scaling changes the EFIC score difference between any arbitrary model $\mathcal{I}$ and the true model $\mathcal{S}$. Hence depending on the C value ($C < 1$ or $C > 1$) and $\Delta > 0$ or $\Delta < 0$, the difference in (16) may become negative leading to a false model selection. Thus, EFIC is not immune to scaling issues. On the contrary, consider the difference for EBIC_{R},

$$\text{EBIC}_{\text{R}}(\mathcal{I}) - \text{EBIC}_{\text{R}}(\mathcal{S}) = (N-2)\ln\left(\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_{\mathcal{S}}^2}\right) - k\ln\hat{\sigma}_{\mathcal{I}}^2 + \\ k_0\ln\hat{\sigma}_{\mathcal{S}}^2 + \Delta\ln\hat{\sigma}_0^2 + \Delta\left(\ln(N/2\pi) + 2\zeta\ln p\right) = D_{\text{EBIC}_{\text{R}}}.$$

Now, scaling $\mathbf{y}$ by C, scales the estimates $\hat{\sigma}_{\mathcal{I}}^2$, $\hat{\sigma}_{\mathcal{S}}^2$ and $\hat{\sigma}_0^2$ by C^2, however, the difference remains the same

$$\text{EBIC}_{\text{R}}(\mathcal{I}) - \text{EBIC}_{\text{R}}(\mathcal{S}) = D_{\text{EBIC}_{\text{R}}} \quad (17)$$

because in this case the $-\Delta\ln C^2$ term is cancelled by $+\Delta\ln C^2$ generated by $\Delta\ln\hat{\sigma}_0^2$. Hence, EBIC_{R} is scale-invariant, which is a desired property in any model selection method.

4. SIMULATION RESULTS

In the simulations, we consider the linear regression model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where the design matrix $\mathbf{A} \in \mathbb{R}^{N\times p}$ is generated with independent entries following normal distribution $\mathcal{N}(0,1)$. The cardinality of the true support $\mathcal{S}$ is chosen to be $card(\mathcal{S}) = k_0 = 5$. Furthermore, without loss of generality, we assume $\mathcal{S} = [1,2,3,4,5]$. This also implies that the elements of $\mathbf{x} \in \mathbb{R}^p$ follows $x_i \neq 0$ for $i = 1,\ldots,k_0$ and $x_i = 0$ for $i > k_0$. We represent the true regression vector as $\mathbf{x}_{\mathcal{S}} = [x_1, x_2, x_3, x_4, x_5]^T$. The SNR in dB is SNR (dB) = $10\log_{10}(\sigma_s^2/\sigma^2)$, where σ_s^2 is the signal power computed as $\sigma_s^2 = ||\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}||_2^2/N$. Based on σ_s^2 and the chosen SNR (dB), the noise power is set as $\sigma^2 = \sigma_s^2/10^{\text{SNR (dB)}/10}$. Using this σ^2, the noise vector $\mathbf{e}$ is generated following $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$.

Algorithm 1 OMP with K iterations

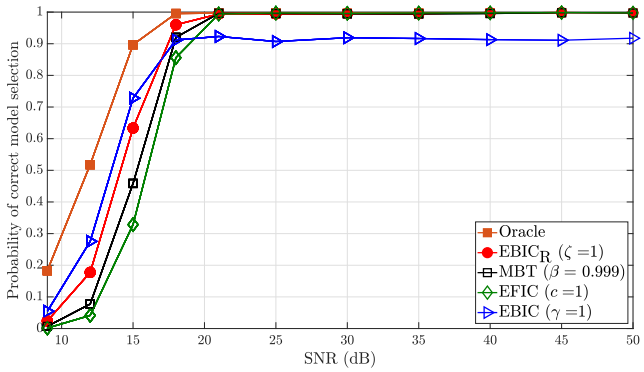
Inputs: Design matrix $\mathbf{A}$, measurement vector $\mathbf{y}$, K.
Initialization: $\|\mathbf{a}_j\|_2 = 1\ \forall j$, $\mathbf{r}^0 = \mathbf{y}$, $\mathcal{S}_{\text{OMP}}^0 = \emptyset$
for $i = 1$ to K **do**
  Find next column index $t^i = \underset{j\in[1,\ldots,p]}{\arg\max}\left|\mathbf{a}_j^T\mathbf{r}^{i-1}\right|$
  Add current index: $\mathcal{S}_{\text{OMP}}^i = \mathcal{S}_{\text{OMP}}^{i-1} \cup \{t^i\}$
  Update residual: $\mathbf{r}^i = \mathbf{\Pi}_{\mathcal{S}_{\text{OMP}}^i}^{\perp}\mathbf{y}$
end for
Output: OMP generated index sequence $\mathcal{S}_{\text{OMP}}^K$

The probability of correct model selection is estimated over 1000 Monte Carlo trials. To maintain randomness in the data, a new design matrix $\mathbf{A}$ is generated at each Monte Carlo trial. OMP (Algorithm 1) is used for predictor selection with $K = 20$. Algorithm 2 illustrates executing model selection combining OMP and EBIC_{R}. The performance of EBIC_{R} is compared to the 'oracle' (OMP with *a priori* knowledge of k_0), EBIC, EFIC and MBT [19], which is a non-information theoretic method based on hypothesis testing using a test statistic. The hyperparameter settings for the criteria are as follows: $\zeta = 1$ (EBIC_{R}), $\beta = 0.999$ (MBT), $c = 1$ (EFIC) and $\gamma = 1$ (EBIC). Note that LASSO can also be employed for predictor selection instead of OMP, but we prefer OMP for now, primarily because (i) MBT in its current form cannot be combined with LASSO for model selection because the support index sequence generated by LASSO is not monotonic in nature and (ii) OMP is less computationally intensive.

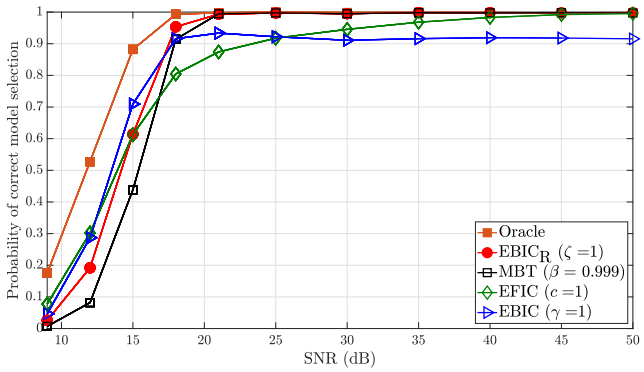
Fig. 1 shows the empirical probability of correct model selection versus SNR (dB) for $N = 54$ and $p = 1000$. To highlight the scaling problem, we consider two different design of $\mathbf{x}_{\mathcal{S}}$. Fig. 1a and Fig. 1b correspond to $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]$ and $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]$, respectively. Comparing the figures, the first clear observation is that EBIC is inconsistent when SNR is high but N is small and fixed as reported in [10]. Secondly, for the considered hyperparameters, EFIC, MBT and EBIC_{R} are empirically consistent for increasing SNR scenarios given that N is fixed. Furthermore, unlike the other criteria, the behaviour of EFIC is not identical for the two different $\mathbf{x}_{\mathcal{S}}$ given that the other parameters viz, N, p and k_0 are constant and the performance is evaluated for the same

Algorithm 2 Model selection combining EBIC_{R} with OMP

Run OMP for K iterations to obtain $\mathcal{S}_{\text{OMP}}^K$
for $i = 1$ to K **do**
  $\mathcal{I} = \mathcal{S}_{\text{OMP}}^i$
  Compute $\text{EBIC}_{\text{R}}(\mathcal{I})$ using (14)
end for
Estimated true support: $\hat{\mathcal{S}} = \underset{\mathcal{I}}{\arg\min}\ \{\text{EBIC}_{\text{R}}(\mathcal{I})\}$.

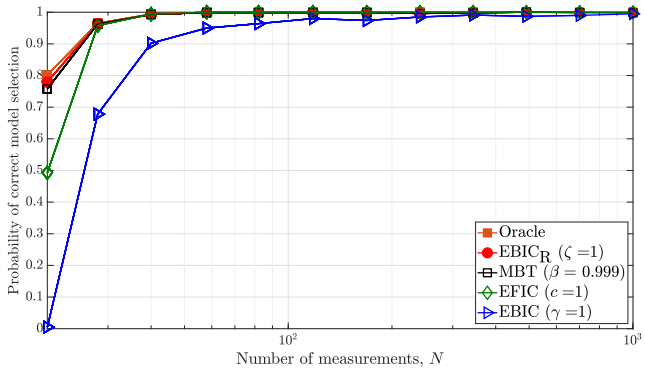


(a) $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$

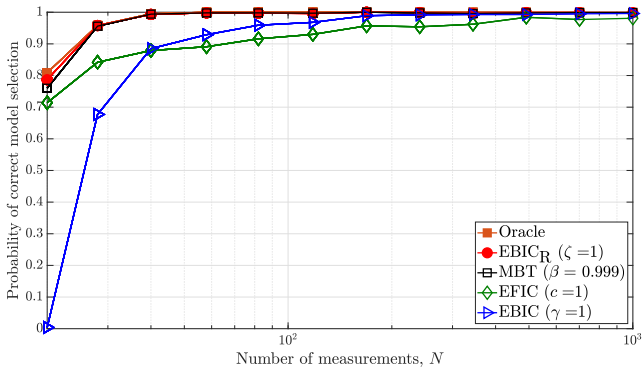


(b) $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$

Fig. 1: The probability of correct model selection versus SNR (dB) for $N = 54$, $p = 1000$ and $k_0 = 5$.



(a) $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$



(b) $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$

Fig. 2: The probability of correct model selection versus N (20 to 10^3) for SNR = 25 dB, $p = N^d$ where $d = 1.3$.

SNR range. This highlights the scaling problem present in EFIC that produces irregular penalties leading to either high underfitting or overfitting issues. For the considered setting, the above behavior of EFIC can be explained as follows: The data dependent penalty term (DDPT) of EFIC is DDPT $= -(k+2)\ln\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2$, whose overall value highly depends on $\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2$, which in turn is influenced by the signal and noise powers σ_s^2 and σ^2 respectively. If $\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 \ll 1$, then DDPT $\gg 0$, which may blow the overall penalty to a large value leading to underfitting issues. This is most likely the case when $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$ (Fig. 1a). On the contrary if $\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 \gg 1$, then DDPT $\ll 0$, thus lowering the overall penatly leading to overfitting issues (when $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$, Fig. 1b).

Fig. 2, illustrates the empirical probability of correct model selection versus N for SNR = 25 dB. A varying parameter space is considered where p grows with N as follows $p = N^d$, where $d = 1.3$. Fig. 2a and Fig. 2b correspond to $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$ and $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$ respectively. Comparing Fig. 2a and 2b it is clearly observed that for high-SNR scenarios, EBIC$_R$ and MBT provides much faster convergence to oracle behaviour as compared to EBIC that requires higher sample size to achieve convergence. We also notice that EFIC suffers from a higher overfitting error for $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$ and performs lower than EBIC in a certain region of the sample size. This clearly underlines the effect of scaling on the behaviour of EFIC.

5. CONCLUSION

In this paper, we presented an improved model selection criterion for sparse high-dimensional linear regression models called EBIC$_R$ where the subscript 'R' stands for robust. EBIC$_R$ solves the inconsistency issue of EBIC for high-SNR and the scaling problem of EFIC. Simulation results indicated that EBIC$_R$ is a consistent criterion and its performance is quite appreciable in many regions of the settings compared to EBIC, EFIC and MBT. Furthermore, compared to MBT, EBIC$_R$ has more flexibility on the choice of predictor selection algorithms. In a future extension of this paper, we will provide detailed analytical consistency guarantees of EBIC$_R$ and extensive simulation results.

6. REFERENCES

[1] CR Rao, Y Wu, Sadanori Konishi, and Rahul Mukerjee, "On model selection," *Lecture Notes-Monograph Series*, pp. 1–64, 2001.

[2] Petre Stoica and Yngve Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[3] Jie Ding, Vahid Tarokh, and Yuhong Yang, "Model selection techniques: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 6, pp. 16–34, 2018.

[4] D Anderson and K Burnham, "Model selection and multi-model inference," *Second Edition. NY: Springer-Verlag*, vol. 63, no. 2020, pp. 10, 2004.

[5] Hirotugu Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[6] Gideon Schwarz et al., "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[7] Mark H Hansen and Bin Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[8] Jiahua Chen and Zehua Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[9] Arash Owrang and Magnus Jansson, "Model selection for high-dimensional data," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 606–609.

[10] Arash Owrang and Magnus Jansson, "A model selection criterion for high-dimensional linear regression," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3436–3446, 2018.

[11] Hamparsum Bozdogan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[12] T Tony Cai and Lie Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[13] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[14] S. Kay, *Fundamental of Statistical Signal Processing: Volume I Estimation Theory*, Prentice Hall, 1998.

[15] Petre Stoica and Prabhu Babu, "On the proper forms of BIC for model order selection," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4956–4961, 2012.

[16] Andrew A Neath and Joseph E Cavanaugh, "The Bayesian information criterion: background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 199–203, 2012.

[17] Daniel F Schmidt and Enes Makalic, "The consistency of MDL for linear regression models with increasing signal-to-noise ratio," *IEEE transactions on signal processing*, vol. 60, no. 3, pp. 1508–1510, 2011.

[18] Petar M Djuric, "Asymptotic MAP criteria for model selection," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2726–2735, 1998.

[19] Prakash B Gohain and Magnus Jansson, "Relative cost based model selection for sparse high-dimensional linear regression models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5515–5519.