# Robust Information Criterion for Model Selection in Sparse High-Dimensional Linear Regression Models

Prakash B. Gohain, *Student Member, IEEE,* Magnus Jansson, *Senior Member, IEEE.*

***Abstract*—Model selection in linear regression models is a major challenge when dealing with high-dimensional data where the number of available measurements (sample size) is much smaller than the dimension of the parameter space. Traditional methods for model selection such as Akaike information criterion, Bayesian information criterion (BIC), and minimum description length are heavily prone to overfitting in the high-dimensional setting. In this regard, extended BIC (EBIC), which is an extended version of the original BIC, and extended Fisher information criterion (EFIC), which is a combination of EBIC and Fisher information criterion, are consistent estimators of the true model as the number of measurements grows very large. However, EBIC is not consistent in high signal-to-noise-ratio (SNR) scenarios where the sample size is fixed and EFIC is not invariant to data scaling resulting in unstable behaviour. In this paper, we propose a new form of the EBIC criterion called EBIC-Robust, which is invariant to data scaling and consistent in both large sample sizes and high-SNR scenarios. Analytical proofs are presented to guarantee its consistency. Simulation results indicate that the performance of EBIC-Robust is quite superior to that of both EBIC and EFIC.**

***Index Terms*—High-dimension, linear regression, data scaling, statistical model selection, subset selection, sparse estimation, scale-invariant, variable selection.**

## I. Introduction

Selecting the true or best set of covariates from a large pool of potential covariates is a fundamental requirement in many applications of science, engineering, and biology. In this paper, our primary focus is on model selection (MS) in high-dimensional linear regression models associated with the maximum likelihood (ML) method of parameter estimation where the number of measurements, $N$, is quite small compared to the model space or parameter dimension, $p$, i.e., $N < p$. High-dimensional datasets are common phenomena in many fields of scientific studies, and as such MS is a central element of data analysis and statistical inference [1].

Consider the linear model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the measurement vector and $\mathbf{A} \in \mathbb{R}^{N\times p}$ is the known design matrix. We consider a high-dimensional setting, hence $p > N$. Also, $p$ can be linked to $N$ as $p = N^d$, where $d > 0$ is a real value. $\mathbf{e} \in \mathbb{R}^N$ is the associated noise vector whose elements are assumed to be i.i.d. following a Gaussian distribution, i.e., $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$ where $\sigma^2$ is the unknown true noise power. $\mathbf{x} \in \mathbb{R}^p$ is the unknown parameter vector. Here, $\mathbf{x}$ is assumed to be sparse, which implies that very few of the elements of $\mathbf{x}$ are non-zero. We denote $\mathcal{S}$ as the true support of $\mathbf{x}$, i.e., $\mathcal{S} = \{i : x_i \neq 0\}$ having cardinality $card(\mathcal{S}) = k_0 \ll N$ and $\mathbf{A}_\mathcal{S}$ as the set of columns of $\mathbf{A}$ corresponding to the support $\mathcal{S}$. The goal of MS is estimating $\mathcal{S}$ given $\mathbf{y}$ and $\mathbf{A}$.

A popular approach for MS is using information theoretic criteria [2], [3], [4], [5]. A typical information criterion based MS rule picks the best model that minimizes some statistical metric as shown below

$$\hat{\mathcal{S}} = \underset{\mathcal{I}\in\mathcal{J}}{\arg\min}\{f(\mathcal{M}_\mathcal{I}) + \mathcal{P}(\mathcal{I})\}, \tag{2}$$

where $\hat{\mathcal{S}}$ is the model estimate, $\mathcal{J}$ is the set of candidate models under consideration, and $\mathcal{M}_\mathcal{I}$ denotes the model with support $\mathcal{I}$. The statistical metric consists of two parts: (1) $f(\mathcal{M}_\mathcal{I})$ representing the goodness of fit of model $\mathcal{M}_\mathcal{I}$ and (2) $\mathcal{P}(\mathcal{I})$ is the penalty term that compensates for overparameterization. The literature on MS is quite extensive. Some of the popular classical MS rules include Akaike information criterion [6], Bayesian information criterion (BIC)[7], minimum description length (MDL)[8], gMDL[9], nMDL[10], penalizing adaptively the likelihood (PAL) [11], Bayesian model comparison with g-prior [12], etc. However, these classical methods in their current form fail to handle the large dimension cases and tend to overfit the final model [13], [14].

Among the classical methods of MS, BIC has been quite successful due to its simplicity and consistent performance in many fields. BIC is asymptotically consistent in selecting the true model as $N$ grows very large given that $p$ and the true noise variance $\sigma^2$ is fixed. However, its performance in high-dimensional settings when $p > N$ is not satisfactory and it has a tendency to select more co-variates than required, thus overfitting the model [13]. To handle the large-$p$ small-$N$ scenario, the authors in [13] proposed a novel extension to the original BIC called extended BIC (EBIC), that takes into account both the number of unknown parameters and the complexity of the model space. EBIC adds dynamic prior model probabilities to each of the models under consideration that is inversely proportional to the model set dimension. This eliminates the earlier assumption of assigning uniform prior to all models irrespective of their sizes, which goes against the principle of parsimony. EBIC is consistent in selecting the true model as $N$ tends to infinity [13]. However, the consistent behaviour of EBIC fails when $N$ is small and fixed and $\sigma^2$ tends to zero [14]. This new consistency requirement was

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.

The authors are with the Division of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm SE-10044, Sweden e-mail: pbg@kth.se, janssonm@kth.se.

first introduced in [15], where the authors highlighted that the original BIC is also inconsistent for fixed $N$ and decreasing noise variance scenarios where $N > p$.

To overcome the drawbacks of EBIC, the authors in [14] proposed a criterion called extended Fisher information criterion (EFIC) that is inspired by EBIC and the MS criteria with Fisher information [16]. The authors analyzed the performance of EFIC in the high-dimensional setting for two key cases: (1) when $\sigma^2$ is fixed and $N$ tends to infinity; (2) when $N$ is fixed and $\sigma^2$ tends to zero. In each case, it was shown that EFIC selects the true model with a probability approaching one. However, as indicated in our simulations, EFIC is not invariant to data scaling and it tends to suffer from overfitting issues (and sometimes underfitting) in practical sizes of $N$ when the data is scaled. This scaling problem is a result of the data-dependent penalty design that may blow the penalty to extremely small or large values depending on how the data is scaled.

Apart from the criteria mentioned above, there are other non-information theoretic methods available for MS. One such popular method is cross-validation (CV) [17], [18]. However, CV-based procedures can be computationally intensive and their performance in high-dimensional problems is not satisfactory [19], [20]. Recent additions to the list of MS methods for high-dimensional data are residual ratio thresholding (RRT)[21] and multi-beta-test (MBT) [22]. Both are non-information theoretic methods based on hypothesis testing using a test statistic. They operate along with a greedy variable selection method such as orthogonal matching pursuit (OMP) [23] and involve a tuning parameter (TP) $\in [0, 1]$, that is connected to the probability of false selection. However, there is no optimal way to set it and as such, they may tend to overfit or underfit the model depending on the chosen TP value. Moreover, in their current form, they can only be used with algorithms that generate monotonic sequences of support estimates such as OMP, which restricts their usability. Knockoff filters [24] are newly developed methods for variable selection in high dimensional inference. Initially designed for linear regression but later generalized to other regression models [25].

The contributions of the paper are as follows: (i) This paper proposes a modified criterion for MS in high-dimensional linear regression models called EBIC-Robust or $\text{EBIC}_\text{R}$ in short. $\text{EBIC}_\text{R}$ mitigates the data-scaling problem of EFIC and unlike EBIC it is consistent for both large $N$ and high-SNR scenarios. Some preliminary results have been published in [26]. (ii) To guarantee consistency of the criterion, analytical proofs are provided to show that under a suitable asymptotic identifiability condition, $\text{EBIC}_\text{R}$ selects the true model with a probability approaching one as $N \to \infty$ as well as when $\sigma^2 \to 0$. (iii) The theoretical analysis also provides a lower bound on the TP of $\text{EBIC}_\text{R}$ such that the criterion is consistent as $N \to \infty$ under the setting that $p$ grows with $N$ as $p = N^d$. (iv) In theory, $\text{EBIC}_\text{R}$ can be seen as an extension of $\text{BIC}_\text{R}$ [27], which was designed under the classical order selection setting. However, the paper highlights through extensive simulations the ineffectiveness of the classical methods (including $\text{BIC}_\text{R}$) when dealing with high-dimensional data employing greedy algorithms for predictor selection and the advantage of $\text{EBIC}_\text{R}$ over existing methods.

Notations used in the paper are as follows. Boldface letters denote matrices and vectors. The notation $(\cdot)^T$ stands for transpose. $\mathbf{A}_\mathcal{I}$ denotes a sub-matrix of the full matrix $\mathbf{A}$ formed using the columns indexed by the support set $\mathcal{I}$. $\mathbf{I}_N$ is an $N \times N$ identity matrix. $\mathbf{\Pi}_\mathcal{I} = \mathbf{A}_\mathcal{I}(\mathbf{A}_\mathcal{I}^T\mathbf{A}_\mathcal{I})^{-1}\mathbf{A}_\mathcal{I}^T$ denotes the orthogonal projection matrix on the span of $\mathbf{A}_\mathcal{I}$ and $\mathbf{\Pi}_\mathcal{I}^\perp = \mathbf{I}_N - \mathbf{\Pi}_\mathcal{I}$. The notation $\left|\mathbf{X}\right|$ denotes the determinant of the matrix $\mathbf{X}$ and $\|\cdot\|_2$ denotes the Euclidean norm. $X \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes a Gaussian distributed random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. $X \sim \chi_k^2$ is a central chi-squared distributed random variable with $k$ degrees of freedom, $X \sim \chi_k^2(\lambda)$ is a noncentral chi-squared distributed random variable with $k$ degrees of freedom and non-centrality parameter $\lambda$. Further, $\mathcal{O}(\cdot)$ denotes the standard Big-O notation and we use $\mathrm{O}(\cdot)$ to denote dominating order of growth.

## II. Background

Given the linear model (1), the entire process of MS or in other words estimating the true support set $\mathcal{S}$ involves two major steps: (i) Predictor/subset selection, which includes finding a competent set of candidate models out of all the $(2^p - 1)$ possible models. In our work, we consider the set of competing models as the collection of all plausible combinatorial models up to a maximum cardinality $K$, under the assumption that $k_0 \leq K \ll N$; (ii) estimating the true model among the candidate models using a suitable MS criterion. For a candidate model with support $\mathcal{I}$ having cardinality $card(\mathcal{I}) = k$, the linear model in (1) can be reformulated as follows

$$\mathcal{H}_\mathcal{I} : \mathbf{y} = \mathbf{A}_\mathcal{I}\mathbf{x}_\mathcal{I} + \mathbf{e}_\mathcal{I}, \tag{3}$$

where $\mathcal{H}_\mathcal{I}$ denotes the hypothesis that the data $\mathbf{y}$ is truly generated according to (3), $\mathbf{A}_\mathcal{I} \in \mathbb{R}^{N\times k}$ is the sub-design matrix consisting of columns from the known design matrix $\mathbf{A}$ with support $\mathcal{I}$, $\mathbf{x}_\mathcal{I} \in \mathbb{R}^k$ is the corresponding unknown parameter vector and $\mathbf{e}_\mathcal{I} \in \mathbb{R}^N$ is the associated noise vector following $\mathbf{e}_\mathcal{I} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathcal{I}^2\mathbf{I}_N)$ where $\sigma_\mathcal{I}^2$ is the unknown noise variance corresponding to the hypothesis $\mathcal{H}_\mathcal{I}$.

### A. Bayesian Framework for Model Selection

To motivate the proposed criterion we start by describing the Bayesian framework that leads to the maximum a-posteriori (MAP) estimator, which in turn forms the backbone for deriving BIC and its extended versions, viz., EBIC, EFIC, as well as the proposed criterion $\text{EBIC}_\text{R}$. Now, for the considered model in (3), the probability density function (pdf) of the data vector $\mathbf{y}$ is given as

$$p(\mathbf{y}|\boldsymbol{\theta}_\mathcal{I}, \mathcal{H}_\mathcal{I}) = \frac{\exp\{-\|\mathbf{y} - \mathbf{A}_\mathcal{I}\mathbf{x}_\mathcal{I}\|_2^2/2\sigma_\mathcal{I}^2\}}{(2\pi\sigma_\mathcal{I}^2)^{N/2}}, \tag{4}$$

where $\boldsymbol{\theta}_\mathcal{I} = [\mathbf{x}_\mathcal{I}^T, \sigma_\mathcal{I}^2]^T$ comprises of all the parameters of the model. Under hypothesis $\mathcal{H}_\mathcal{I}$, the maximum likelihood estimates (MLEs) of $\hat{\boldsymbol{\theta}}_\mathcal{I} = [\hat{\mathbf{x}}_\mathcal{I}^T, \hat{\sigma}_\mathcal{I}^2]^T$ are obtained as [28]

$$\hat{\mathbf{x}}_\mathcal{I} = \left(\mathbf{A}_\mathcal{I}^T\mathbf{A}_\mathcal{I}\right)^{-1}\mathbf{A}_\mathcal{I}^T\mathbf{y} \quad \& \quad \hat{\sigma}_\mathcal{I}^2 = \frac{\mathbf{y}^T\mathbf{\Pi}_\mathcal{I}^\perp\mathbf{y}}{N}. \tag{5}$$

EBIC$_{\mathrm{R}}$ is derived using the Bayesian framework of MS. The maximum a-posteriori (MAP) criterion is first derived and after some suitable modifications and reasonable assumptions we arrive at the EBIC$_{\mathrm{R}}$. We follow similar steps as in [27], [26] but include them here for completeness. Let us denote the prior pdf of the parameter vector $\boldsymbol{\theta}_{\mathcal{I}}$ as $p(\boldsymbol{\theta}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$, the marginal of $\mathbf{y}$ as $p(\mathbf{y}|\mathcal{H}_{\mathcal{I}})$ and the prior probability of the model with support $\mathcal{I}$ as $\Pr(\mathcal{H}_{\mathcal{I}})$. Then the MAP estimate of $\mathcal{S}$ is equivalently given by [29], [27]

$$\hat{\mathcal{S}}_{\text{MAP}} = \arg\max_{\mathcal{I}} \Big\{ \ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}}) + \ln \Pr(\mathcal{H}_{\mathcal{I}}) \Big\}. \tag{6}$$

The classical derivation employs a second-order Taylor series expansion around the MLE to obtain an approximation of $\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}})$ under the premise that $N$ is large or/and SNR is high (see [29], [30])

$$\begin{aligned}\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}}) &\approx \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}}) + \ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) \\ &+ \left(\frac{k+1}{2}\right)\ln(2\pi) - \frac{1}{2}\ln\big|\hat{\mathbf{F}}_{\mathcal{I}}\big|,\end{aligned} \tag{7}$$

where $k = card(\mathcal{I})$ and $\hat{\mathbf{F}}_{\mathcal{I}}$ is the sample Fisher information matrix (FIM) under $\mathcal{H}_{\mathcal{I}}$ given as [28]

$$\hat{\mathbf{F}}_{\mathcal{I}} = -\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}})}{\partial \boldsymbol{\theta}_{\mathcal{I}} \partial \boldsymbol{\theta}_{\mathcal{I}}^T}\bigg|_{\boldsymbol{\theta}_{\mathcal{I}} = \hat{\boldsymbol{\theta}}_{\mathcal{I}}}. \tag{8}$$

Evaluating (8) using (4) and (5) we get [29]

$$\hat{\mathbf{F}}_{\mathcal{I}} = \begin{bmatrix} \frac{1}{\hat{\sigma}_{\mathcal{I}}^2}\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \frac{N}{2\hat{\sigma}_{\mathcal{I}}^4} \end{bmatrix}. \tag{9}$$

Now, for the considered linear model we have

$$-2\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}}) = N\ln\hat{\sigma}_{\mathcal{I}}^2 + \text{const.} \tag{10}$$

Therefore, using (10), we can rewrite (7) as

$$\begin{aligned}-2\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}}) \approx N\ln\hat{\sigma}_{\mathcal{I}}^2 + \ln\big|\hat{\mathbf{F}}_{\mathcal{I}}\big| - 2\ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) \\ -k\ln 2\pi + \text{const.}\end{aligned} \tag{11}$$

Traditionally, the prior term in (7), i.e., $\ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ is assumed to be flat and uninformative, and hence omitted from the analysis. Thus, dropping the constants and the terms independent of the model dimension $k$, we can equivalently reformulate the MAP based model estimate as

$$\hat{\mathcal{S}}_{\text{MAP}} = \arg\min_{\mathcal{I}}\Big\{N\ln\hat{\sigma}_{\mathcal{I}}^2 + \ln\big|\hat{\mathbf{F}}_{\mathcal{I}}\big| - k\ln 2\pi - 2\ln\Pr(\mathcal{H}_{\mathcal{I}})\Big\}. \tag{12}$$

## B. BIC

The classical BIC can be derived from the MAP estimator in (12). The term $-k\ln 2\pi$ is ignored as it weakly depends on the model dimension $k$ and hence is typically much smaller than the dominating terms. Moreover, the prior probability of each candidate model is assumed to be equiprobable. Hence, the $-2\ln\Pr(\mathcal{H}_{\mathcal{I}})$ term is dropped as well. Now, expanding the $|\hat{\mathbf{F}}_{\mathcal{I}}|$ term of (12) using (9) we have

$$\ln\big|\hat{\mathbf{F}}_{\mathcal{I}}\big| = \ln(N/2) - (k+2)\ln\hat{\sigma}_{\mathcal{I}}^2 + \ln\big|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\big|. \tag{13}$$

*Assumption 1:* The further analysis of BIC considers the following property of the design matrix $\mathbf{A}$ [29], [31]

$$\lim_{N\to\infty}\left\{N^{-1}(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}})\right\} = \mathbf{M}_{\mathcal{I}}, \tag{14}$$

where $\mathbf{M}_{\mathcal{I}}$ is a $k \times k$ positive definite matrix and bounded as $N \to \infty$.

Assumption 1 is true in many applications but not all (see [32] for more details). Using (14), it is possible to show that for large $N$

$$\ln\big|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\big| = \ln\Big|N \cdot N^{-1}(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}})\Big| = k\ln N + \mathcal{O}(1). \tag{15}$$

Furthermore, $\hat{\sigma}_{\mathcal{I}}^2$ is considered to be of $\mathcal{O}(1)$ since it does not grow with $N$. Dropping the $\mathcal{O}(1)$ term, $(k+2)\ln\hat{\sigma}_{\mathcal{I}}^2$ and $\ln(N/2)$ (a constant) from (13) leads to the BIC

$$\text{BIC}(\mathcal{I}) = N\ln\hat{\sigma}_{\mathcal{I}}^2 + k\ln N. \tag{16}$$

BIC is consistent when $p$ is fixed and $N \to \infty$. However, it is inconsistent when $N$ is fixed and $\sigma^2 \to 0$ [33], [27] as well as when $p > N$ and $p$ grows with $N$ [13].

## C. EBIC

The authors in [13] proposed an extended version of the BIC, i.e., EBIC, to mitigate the drawbacks of BIC for large-$p$ small-$N$ scenarios. EBIC can be derived from the MAP estimator in (12), using the same assumptions as in BIC, except for the prior probability term $\Pr(\mathcal{H}_{\mathcal{I}})$. In EBIC, the idea of equiprobable models is discredited and instead a prior probability is assigned that is inversely proportional to the size of the model space. Thus, a model with dimension $k$ is assigned prior probability of $\Pr(\mathcal{H}_{\mathcal{I}}) \propto \binom{p}{k}^{-\gamma}$, where $0 \le \gamma \le 1$ is a TP. Thus, the EBIC is

$$\text{EBIC}(\mathcal{I}) = N\ln\hat{\sigma}_{\mathcal{I}}^2 + k\ln N + 2\gamma\ln\binom{p}{k}. \tag{17}$$

When $\gamma = 0$, EBIC boils down to BIC (16). Moreover, unlike BIC, EBIC is consistent in selecting the true model for $p \gg N$ cases where $p$ grows with $N$. However, it has been observed in [14] that EBIC is inconsistent when $N$ is fixed and $\sigma^2 \to 0$.

## D. EFIC

To circumvent the shortcomings of EBIC in high-SNR cases, the authors in [14] proposed EFIC. In EFIC, the assumptions imposed on the sample FIM (13) are removed and the entire structure is included as it is in the criterion except for the constant term $\ln(N/2)$. Some further simplifications are involved:

$$N\ln\hat{\sigma}_{\mathcal{I}}^2 = N\ln\big\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\big\|_2^2 - N\ln N \tag{18}$$

$$(k+2)\ln\hat{\sigma}_{\mathcal{I}}^2 = (k+2)\left[\ln\big\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\big\|_2^2 - \ln N\right]. \tag{19}$$

The $-N\ln N$ and $-2\ln N$ term of (18) and (19) respectively are independent of the model dimension $k$ and hence ignored. Similar to EBIC the prior probability term is assumed to be proportional to the model space, hence $\Pr(\mathcal{H}_{\mathcal{I}}) \propto \binom{p}{k}^{-c}$, where $c > 0$ is a TP. Furthermore, under the large-$p$ approximation and since $k \leq K \ll p$, the $\ln\binom{p}{k}$ term is approximated as

$$\ln\binom{p}{k} = \sum_{i=0}^{k-1} \ln(p-i) - \ln(k!) \approx k \ln p. \tag{20}$$

Hence, for large-$p$ case, we can set $-2\ln p(\mathcal{H}_\mathcal{I}) \approx 2ck\ln p$. Thus, the EFIC is given as

$$\begin{aligned} \text{EFIC}(\mathcal{I}) = N \ln \left\|\mathbf{\Pi}_\mathcal{I}^\perp \mathbf{y}\right\|_2^2 + k \ln N + \ln\left|\mathbf{A}_\mathcal{I}^T\mathbf{A}_\mathcal{I}\right| \\ -(k+2)\ln\left\|\mathbf{\Pi}_\mathcal{I}^\perp \mathbf{y}\right\|_2^2 + 2ck\ln p. \end{aligned} \tag{21}$$

EFIC is consistent in both large-$N$ and high-SNR scenarios [14]. However, EFIC suffers from a data scaling problem due to the inclusion of the data-dependent penalty term and as such the performance of EFIC is not invariant to data scaling. See further in Section III-A.

## III. Proposed criterion: EBIC-Robust (EBIC$_\text{R}$)

In this section, we present the necessary steps for deriving EBIC$_\text{R}$. EBIC$_\text{R}$ can be seen as a natural extension of BIC$_\text{R}$ [27] for performing MS in large-$p$ small-$N$ scenarios. Below, we provide a detailed derivation and establish the connection to BIC$_\text{R}$. First we factorize the $\ln\left|\hat{\mathbf{F}}_\mathcal{I}\right|$ term in (12) in the following manner [29], [27]

$$\begin{aligned} \ln\left|\hat{\mathbf{F}}_\mathcal{I}\right| &= \ln\left[\left|\mathbf{L}\right| \left|\mathbf{L}^{-1/2}\hat{\mathbf{F}}_\mathcal{I}\mathbf{L}^{-1/2}\right|\right] \\ &= \ln\left|\mathbf{L}\right| + \ln\underbrace{\left|\mathbf{L}^{-1/2}\hat{\mathbf{F}}_\mathcal{I}\mathbf{L}^{-1/2}\right|}_{\text{T}}. \end{aligned} \tag{22}$$

The goal here is to choose a suitable $\mathbf{L}$ matrix that normalizes the sample FIM $\hat{\mathbf{F}}_\mathcal{I}$ such that the T term in (22) is $\mathcal{O}(1)$, i.e., in this case T should be bounded as $N \to \infty$ and/or $\sigma^2 \to 0$. To accomplish this objective, we choose the following $\mathbf{L}^{-1/2}$ matrix

$$\mathbf{L}^{-1/2} = \begin{bmatrix} \sqrt{\frac{1}{N}}\sqrt{\frac{\hat{\sigma}_\mathcal{I}^2}{\hat{\sigma}_0^2}}\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \sqrt{\frac{1}{N}\frac{\hat{\sigma}_\mathcal{I}^2}{\hat{\sigma}_0^2}} \end{bmatrix}, \tag{23}$$

where $\hat{\sigma}_0^2 = \|\mathbf{y}\|_2^2/N$. The factor, $\hat{\sigma}_0^2$, is used in $\mathbf{L}^{-1/2}$ in order to neutralize the data scaling problem and is motivated by the fact that given (14), when the SNR is a constant, we have

$$\mathbb{E}[\hat{\sigma}_0^2] \to \text{const.} \quad \& \quad \text{Var}[\hat{\sigma}_0^2] \to 0 \tag{24}$$

as $N \to \infty$. Also, from the considered generating model in (1), when $N$ is fixed, (24) is also satisfied as $\sigma^2 \to 0$ (see Appendix C for details on $\hat{\sigma}_0^2$). Now using (9), (23), the assumption 1 in (14), and (24) it is possible to show that

$$\left|\mathbf{L}^{-1/2}\hat{\mathbf{F}}_\mathcal{I}\mathbf{L}^{-1/2}\right| = \begin{vmatrix} \frac{1}{\hat{\sigma}_0^2}\frac{\mathbf{A}_\mathcal{I}^T\mathbf{A}_\mathcal{I}}{N} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\hat{\sigma}_0^4} \end{vmatrix} = \mathcal{O}(1), \tag{25}$$

and therefore may be discarded without much effect on the criterion. Further, the $\ln\left|\mathbf{L}\right|$ term can be expanded as

$$\begin{aligned} \ln\left|\mathbf{L}\right| &= \ln\begin{vmatrix} N\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_\mathcal{I}^2}\right)\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & N\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_\mathcal{I}^2}\right)^2 \end{vmatrix} \\ &= (k+1)\ln N + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_\mathcal{I}^2}\right). \end{aligned} \tag{26}$$

Therefore, using (25) and (26) we can rewrite (22) as

$$\ln\left|\hat{\mathbf{F}}_\mathcal{I}\right| = k\ln N + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_\mathcal{I}^2}\right) + \mathcal{O}(1) + \ln N. \tag{27}$$

Next, for the model prior probability term $-2\ln\Pr(\mathcal{H}_\mathcal{I})$ in (12), a similar proposition is taken as in EBIC such that $\Pr(\mathcal{H}_\mathcal{I}) \propto \binom{p}{k}^{-\zeta}$, where $\zeta \geq 0$ is a TP. For large-$p$, we follow a similar approach as in EFIC by employing the following approximation $\ln\binom{p}{k} \approx k\ln p$ . This gives

$$-2\ln\Pr(\mathcal{H}_\mathcal{I}) = 2\zeta k\ln p + \text{const.} \tag{28}$$

Now, substituting (27), (28) in (12) and dropping the $\mathcal{O}(1)$, the $\ln N$ term (independent of $k$), the constant and the $p(\hat{\boldsymbol{\theta}}_\mathcal{I}|\mathcal{H}_\mathcal{I})$ term we arrive at the EBIC$_\text{R}$:

$$\begin{aligned} \text{EBIC}_\text{R}(\mathcal{I}) = N\ln\hat{\sigma}_\mathcal{I}^2 + k\ln\left(\frac{N}{2\pi}\right) + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_\mathcal{I}^2}\right) \\ + 2k\zeta\ln p. \end{aligned} \tag{29}$$

The true model is estimated as

$$\hat{\mathcal{S}}_{\text{EBIC}_\text{R}} = \arg\min_\mathcal{I}\{\text{EBIC}_\text{R}(\mathcal{I})\}, \tag{30}$$

It can be observed from (29) that the penalty of EBIC$_\text{R}$ is a function of $N$, the ratio $(\hat{\sigma}_0^2/\hat{\sigma}_\mathcal{I}^2)$ and $p$. Furthermore, when $\mathcal{S} \not\subset \mathcal{I}$, the ratio $(\hat{\sigma}_0^2/\hat{\sigma}_\mathcal{I}^2) = \mathcal{O}(1)$ and for $\mathcal{S} \subset \mathcal{I}$ we have $(\hat{\sigma}_0^2/\hat{\sigma}_\mathcal{I}^2) = \mathcal{O}(\text{SNR}+1)$. Hence, the behaviour of the penalty can be summarized as follows: (i) For fixed $p$ and SNR, as $N \to \infty$ the penalty grows as $\text{O}(\ln N)$; (ii) If $N$ and $p$ are constant, as SNR $\to \infty$, the penalty grows as $\text{O}(\ln(\text{SNR}+1))$ for all $\mathcal{I} \supset \mathcal{S}$; (iii) when SNR is a constant and given that $p$ grows with $N$, then as $N \to \infty$ the penalty grows as $\text{O}(\ln N)$ + $\text{O}(\ln p)$.

### A. Scaling Robustness as Compared to EFIC

In this section, we elaborately discuss the data scaling problem. Ideally, any MS criterion should be invariant to data scaling, which means that if $\mathbf{y}$ is scaled by any arbitrary constant $C > 0$, the equivalent penalty for each of the models $\mathcal{I}$ should not change. This property is necessary because otherwise the behaviour of the MS criterion will be unreliable and may suffer from overfitting or underfitting issues when the data is scaled. As mentioned before, the penalty of EFIC is not invariant to data scaling. This can be observed from the following analysis. Let $\Delta = card(\mathcal{I}) - card(\mathcal{S})$. Now, consider the difference assuming $\mathcal{I} \neq \mathcal{S}$

$$\begin{aligned} &\text{EFIC}(\mathcal{I}) - \text{EFIC}(\mathcal{S}) \\ &= (N-2)\ln\frac{\left\|\mathbf{\Pi}_\mathcal{I}^\perp\mathbf{y}\right\|_2^2}{\left\|\mathbf{\Pi}_\mathcal{S}^\perp\mathbf{y}\right\|_2^2} + \ln\frac{\left|\mathbf{A}_\mathcal{I}^T\mathbf{A}_\mathcal{I}\right|}{\left|\mathbf{A}_\mathcal{S}^T\mathbf{A}_\mathcal{S}\right|} - k\ln\left\|\mathbf{\Pi}_\mathcal{I}^\perp\mathbf{y}\right\|_2^2 \\ &\quad + k_0\ln\left\|\mathbf{\Pi}_\mathcal{S}^\perp\mathbf{y}\right\|_2^2 + \Delta\left(\ln N + 2c\ln p\right) = D_\text{EFIC} \text{ (say)}. \end{aligned} \tag{31}$$

Ideally, for correct MS, $D_\text{EFIC} > 0$ for all $\mathcal{I} \neq \mathcal{S}$. Now, if we scale the data $\mathbf{y}$ by a constant $C > 0$, the data dependent term becomes $\ln\|\mathbf{\Pi}_\mathcal{I}^\perp C\mathbf{y}\|_2^2 = \ln C^2 + \ln\|\mathbf{\Pi}_\mathcal{I}^\perp\mathbf{y}\|_2^2$ and the difference becomes

$$\text{EFIC}(\mathcal{I}) - \text{EFIC}(\mathcal{S}) = D_\text{EFIC} - \Delta\ln C^2. \tag{32}$$

It is evident that (31) and (32) are unequal and the difference after scaling contains an additional term $-\Delta \ln C^2$. This implies that scaling the data 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 $C \geq 1$) and $\Delta > 0$ or $\Delta < 0$, the difference in (32) may become negative leading to a false MS. Thus, EFIC is not invariant to data scaling. On the contrary, consider the difference for $\text{EBIC}_\text{R}$,

$$\begin{aligned}
&\text{EBIC}_\text{R}(\mathcal{I}) - \text{EBIC}_\text{R}(\mathcal{S}) \\
&= (N-2)\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}}}{\hat{\sigma}^2_{\mathcal{S}}}\right) - k\ln\hat{\sigma}^2_{\mathcal{I}} + k_0\ln\hat{\sigma}^2_{\mathcal{S}} + \Delta\ln\hat{\sigma}^2_0 \\
&\quad + \Delta\left(\ln(N/2\pi) + 2\zeta\ln p\right) = D_{\text{EBIC}_\text{R}}(\text{say}) \qquad (33)
\end{aligned}$$

Now, scaling $\mathbf{y}$ by $C$, scales the noise variance estimates $\hat{\sigma}^2_{\mathcal{I}}$, $\hat{\sigma}^2_{\mathcal{S}}$ and $\hat{\sigma}^2_0$ by $C^2$, however, the difference remains the same, i.e., $D_{\text{EBIC}_\text{R}}$. This is because in this case the $-\Delta\ln C^2$ term is cancelled by $+\Delta\ln C^2$ generated by $\Delta\ln\hat{\sigma}^2_0$. Hence, $\text{EBIC}_\text{R}$ is invariant to data scaling, which is a desired property of any MS criterion.

## IV. Consistency of $\text{EBIC}_\text{R}$

This section discusses the consistency of the proposed $\text{EBIC}_\text{R}$. Generally speaking, a MS criterion with $\hat{\mathcal{S}}$ as its estimate of the true model $\mathcal{S}$ is consistent if it satisfies the following conditions [14]

$$\begin{aligned}
&\lim_{\sigma^2\to 0} \Pr\{\hat{\mathcal{S}} = \mathcal{S}\} = 1 \text{ when } N \text{ is fixed,} \\
&\lim_{N\to\infty} \Pr\{\hat{\mathcal{S}} = \mathcal{S}\} = 1 \text{ when } \sigma^2 \text{ is fixed, and } p = N^d.
\end{aligned} \qquad (34)$$

Here $p$ is allowed to grow with $N$. This is a common setting in the MS literature (see, e.g., [13], [14], [34]). Now, let us define the set of all overfitted models of dimension $k$ as $\mathcal{I}^k_o = \{\mathcal{I} : card(\mathcal{I}) = k, \mathcal{S} \subset \mathcal{I}\}$ and the set of all misfitted models of dimension $k$ as $\mathcal{I}^k_m = \{\mathcal{I} : card(\mathcal{I}) = k, \mathcal{S} \not\subset \mathcal{I}\}$. Furthermore, let $\mathbb{O}$ denote the set of all $\mathcal{I}^k_o$ for $k = k_0 + 1, \ldots, K$, and let $\mathbb{M}$ denote the set of all $\mathcal{I}^k_m$ for $k = 1, \ldots, K$, i.e.,

$$\mathbb{O} = \bigcup_{k=k_0+1}^{K} \mathcal{I}^k_o \quad \text{and} \quad \mathbb{M} = \bigcup_{k=1}^{K} \mathcal{I}^k_m\,, \qquad (35)$$

where $K$ is some upper bound for $k_0$ and $k_0 \leq K \ll N$. In practice, $\text{EBIC}_\text{R}$ picks the true model $\mathcal{S}$, if the following conditions are satisfied:

$$\mathcal{C}_1 : \text{EBIC}_\text{R}(\mathcal{S}) < \text{EBIC}_\text{R}(\mathcal{I}) \quad \forall\, \mathcal{I} \in \mathbb{O} \qquad (36)$$

$$\mathcal{C}_2 : \text{EBIC}_\text{R}(\mathcal{S}) < \text{EBIC}_\text{R}(\mathcal{I}) \quad \forall\, \mathcal{I} \in \mathbb{M}. \qquad (37)$$

### A. Asymptotic Identifiability of the Model

In general, the model is identifiable if no model of comparable size other than the true submodel can predict the noise free response almost equally well [13]. In the context of linear regression, this is equivalent to say $\mathbf{y} = \mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} \neq \mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}}$ for $\{\mathcal{I} : card(\mathcal{I}) \leq card(\mathcal{S}), \mathcal{I} \neq \mathcal{S}\}$. The identifiability of the true model in the high-dimensional linear regression setup is uniformly maintained if the minimal eigenvalue of all restricted sub-matrices, $\mathbf{A}^T_{\mathcal{I}}\mathbf{A}_{\mathcal{I}}$ for $\{\mathcal{I} : card(\mathcal{I}) \leq 2K\}$, is bounded away from zero [14].

*Assumption 2:* A sufficient assumption on the design matrix $\mathbf{A}$ to prove the consistency of $\text{EBIC}_\text{R}$ is the sparse Riesz condition [35]:

$$\lim_{N\to\infty} \left\{N^{-1}\left(\mathbf{A}^T_{\mathcal{I}}\mathbf{A}_{\mathcal{I}}\right)\right\} = \mathbf{M}_{\mathcal{I}}, \quad \forall\, card(\mathcal{I}) \leq 2K, \qquad (38)$$

where $\mathbf{M}_{\mathcal{I}}$ denotes a bounded positive definite matrix.

While a weaker assumption may be possible (cf. [13]), we believe this is a natural condition. It connects well to Assumption 1 in (14) and should hold in well-designed experiments where, e.g., near identical regressors have been pruned away.

Now, we present the consistency theorems of $\text{EBIC}_\text{R}$ for large-$N$ and high-SNR. The consistency of $\text{EBIC}_\text{R}$ as $\sigma^2 \to 0$ or SNR $\to \infty$ for fixed $N$ is summarized as a theorem stated below.

*Theorem 1:* Assume that $N$ and $p$ are fixed and the matrix $\mathbf{A}$ satisfies the condition given by (38). If $K \geq k_0$, then $\Pr\{\text{EBIC}_\text{R}(\mathcal{S}) < \text{EBIC}_\text{R}(\mathcal{I})\} \to 1$ as $\sigma^2 \to 0$ for all $\mathcal{I} \neq \mathcal{S}$ and $card(\mathcal{I}) = 1, \ldots, K$.

The proof of Theorem 1 is given in Appendix A.

Next, we consider the consistency of $\text{EBIC}_\text{R}$ as the sample size $N \to \infty$ given that $\sigma^2$ is fixed and under the setting $p = N^d$ for some $d > 0$. This leads to the following theorem.

*Theorem 2:* Assume that $p = N^d$ for some constant $d > 0$, the SNR is fixed and the matrix $\mathbf{A}$ satisfies (38). If $K \geq k_0$, then $\Pr\{\text{EBIC}_\text{R}(\mathcal{S}) < \text{EBIC}_\text{R}(\mathcal{I})\} \to 1$ as $N \to \infty$ for all $\mathcal{I} \neq \mathcal{S}$ and $card(\mathcal{I}) = 1, \ldots, K$ under the condition $\zeta > 1 - 1/2d$.

The proof of Theorem 2 is given in Appendix B.

## V. Predictor selection algorithms

In the high-dimensional scenario, when $p$ is large, it is infeasible to perform MS in the conventional manner. For a design matrix with parameter dimension $p$, the number of possible candidate models is $2^p - 1$. Hence, the candidate model space grows with $p$ and we cannot afford to calculate the model score for all possible models. Therefore, to perform MS, we combine a MS criterion with a predictor selection (support recovery) algorithm such as OMP or LASSO (least absolute shrinkage and selection operator) [36]. The goal of predictor selection is to pick a subset of important predictors from the entire set of $p$ predictors. In this context, the most important predictors refer to the positions of the nonzero elements of the input signal $\mathbf{x}$. Using a predictor selection algorithm we reduce the cardinality of the candidate model space to some upper bound $K$ such that $k_0 \leq K \ll N$ under the assumption of a sparse parameter vector. This enables us to apply the MS criterion to the smaller set of candidate models to pick the best model. The OMP algorithm is shown in Algorithm 1. To perform MS, we combine OMP with $\text{EBIC}_\text{R}$ as shown in Algorithm 2.

LASSO is a shrinkage method for variable selection/estimation in linear regression models developed by Tibshirani [36]. Given the linear model in (1), the LASSO solution for $\mathbf{x}$ for a particular choice of the regularization parameter $\lambda \geq 0$ is obtained as

$$\hat{\mathbf{x}}_{\text{lasso}}(\lambda) = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{\frac{1}{2N}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2_2 + \lambda\|\mathbf{x}\|_1\right\}, \qquad (39)$$

**Algorithm 1** OMP with $K$ iterations

```
Inputs: Design matrix $\mathbf{A}$, measurement vector $\mathbf{y}$.
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: $d^i = \arg\max_j \left|\mathbf{a}_j^T \mathbf{r}^{i-1}\right|$
    Add current index: $\mathcal{S}_{\text{OMP}}^i = \mathcal{S}_{\text{OMP}}^{i-1} \cup \{d^i\}$
    Update residual: $\mathbf{r}^i = \left(\mathbf{I}_N - \mathbf{\Pi}_{\mathcal{S}_{\text{OMP}}^i}\right)\mathbf{y}$
end for
Output: OMP generated index sequence $\mathcal{S}_{\text{OMP}}^K$
```

where $\|\cdot\|_1$ denotes the $l_1$ norm. The parameter $\lambda$ determines the level of sparsity. When $\lambda \to \infty$ the objective function in (39) attains the minimum with $\hat{\mathbf{x}}_{\text{lasso}}(\lambda)$ being a zero vector. As we gradually lower the $\lambda$ value, the number of non-zero components in $\hat{\mathbf{x}}_{\text{lasso}}(\lambda)$ starts increasing. MS combining LASSO and EBIC$_\text{R}$ can be performed as shown in Algorithm 3. Gradually decrease $\lambda$ from a high value so that the number of non-zero components in $\hat{\mathbf{x}}_{\text{lasso}}(\lambda)$ gradually increases. Therefore, for each decreasing unique value of $\lambda$ say $\lambda_i$, we acquire a different solution $\hat{\mathbf{x}}_{\text{lasso}}(\lambda_i)$, with increasing support and thus obtaining a sequence of candidate models with maximum cardinality $K$. The value of EBIC$_\text{R}$ is computed for each of the candidate models and the model corresponding to the smallest EBIC$_\text{R}$ score is selected as the final model. A most useful method for solving LASSO in our context is the (modified) least angle regression (LARS) algorithm [37], since it also provides the required sequence of regularization parameters for which the support changes.

## VI. Simulation Results

In this section, we provide numerical simulation results to illustrate the empirical performance of EBIC$_\text{R}$. The performance of EBIC$_\text{R}$ is compared with the 'oracle', EBIC, EFIC, RRT, and knockoff filter. The 'oracle' criterion assumes *a priori* knowledge of the true cardinality $k_0$. Thus, the oracle provides the upper bound on the MS performance that can be achieved using a particular predictor selection algorithm and for a given set of data settings. Additionally, we also provide simulation results to highlight the drawbacks of classical methods for model selection in high-dimensional linear regression models with a sparse parameter vector.

*General Simulation Setup:* In the simulations, we consider the 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 $\mathcal{N}(0, 1)$. Furthermore, without loss of generality, we assume that the true support is $\mathcal{S} = [1, 2, \ldots, k_0]$, therefore, $\mathbf{x}_\mathcal{S} = [x_1, x_2, \ldots, x_{k_0}]^T$ and $\mathbf{A}_\mathcal{S} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k_0}]$. This implies that the elements of $\mathbf{x}$ follows $x_k \neq 0$ for $k = 1, \ldots, k_0$ and $x_k = 0$ for $k > k_0$. The SNR in dB is SNR (dB) = $10\log_{10}(\sigma_s^2/\sigma^2)$, where $\sigma_s^2$ and $\sigma^2$ denote signal and true noise power, respectively. The signal power is computed as $\sigma_s^2 = ||\mathbf{A}_\mathcal{S}\mathbf{x}_\mathcal{S}||_2^2/N$. Based on $\sigma_s^2$ and the chosen SNR (dB), the noise power is set as $\sigma^2 = \sigma_s^2/10^{\text{SNR (dB)}/10}$. Using this $\sigma^2$, the noise vector $\mathbf{e}$ is generated following $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. The probability of correct model selection (PCMS), i.e., $\Pr\{\hat{\mathcal{S}} = \mathcal{S}\}$, is estimated over 300 Monte Carlo trials. To maintain randomness in the data, a new design matrix $\mathbf{A}$ is generated at each Monte Carlo trial. OMP is used for predictor selection for its simplicity and wider range of applicability.

**Algorithm 2** Model selection combining EBIC$_\text{R}$ with OMP

```
Run OMP for $K$ iterations to obtain $\mathcal{S}_{\text{OMP}}^K$
for $k = 1$ to $K$ do
    $\mathcal{I} = \mathcal{S}_{\text{OMP}}^k$
    Compute EBIC$_\text{R}(\mathcal{I})$
end for
Estimated true support: $\hat{\mathcal{S}}_{\text{EBIC}_\text{R}} = \arg\min_{\mathcal{I}}\{\text{EBIC}_\text{R}(\mathcal{I})\}$
```

**Algorithm 3** Model selection combining EBIC$_\text{R}$ with LASSO

```
Compute LASSO estimates $\{\hat{\mathbf{x}}_{\text{lasso}}(\lambda_1), \ldots, \hat{\mathbf{x}}_{\text{lasso}}(\lambda_{K_{\max}})\}$
where $card(supp\,(\hat{\mathbf{x}}_{\text{lasso}}(\lambda_{K_{\max}}))) = K$
for $i = 1$ to $K_{\max}$ do
    $\mathcal{I} = supp\,(\hat{\mathbf{x}}_{\text{lasso}}(\lambda_i))$
    Compute EBIC$_\text{R}(\mathcal{I})$
end for
Estimated true support: $\hat{\mathcal{S}}_{\text{EBIC}_\text{R}} = \arg\min_{\mathcal{I}}\ \{\text{EBIC}_\text{R}(\mathcal{I})\}$
```

*Tuning Parameter Selection:* An important step in MS is the choice of the TP. In the case of EBIC$_\text{R}$, if $\zeta$ is too high, the overall penalty may become too large as $k$ increases. In such a situation, it is more likely that the minimum EBIC$_\text{R}$ score occurs at some $k < k_0$ such that some of the weaker signals are left out leading to an underfitted model. On the contrary, if $\zeta$ is too low, the penalty may not be sufficiently large to compensate for the overparameterization due to large parameter dimension. As such, the minimum EBIC$_\text{R}$ score may occur at some $k > k_0$ consequently leading to an overfitted model. However, in practical scenarios, it is hard to decide if the chosen $\zeta$ is high or low. Theorem 2 provides a lower bound on $\zeta$ such that consistency is guaranteed as $N$ grows large. Recall that $\zeta$ was introduced in the derivation of EBIC$_\text{R}$ as a parameter of the prior probability (see (28)). The most natural choice from that Bayesian perspective is to set $\zeta = 1$. Furthermore, it can be shown that the probability in (97) will correspond to $1 - 1/\sqrt{N}$ by choosing $\zeta = 1$ (and $\Delta = 1$). Again, this appears as a natural rate as the standard deviations of the parameter estimation errors decay as $1/\sqrt{N}$. Hence, both these arguments suggest that $\zeta = 1$ is a natural choice. This is also illustrated in the following simulation. Fig. 1 shows a performance comparison of EBIC$_\text{R}$ for four different values of $\zeta$ (0.4, 0.6, 1, and 2). Here, we set $p = N^d$ where $d = 1.1$. Hence, from Theorem 2 we require $\zeta > 1 - 1/2d = 0.55$ to achieve consistency. From the figure, we see that for $\zeta = 0.4$, the performance of EBIC$_\text{R}$ degrades after a certain point with increasing $N$, which justifies the theory. For all other $\zeta > 0.55$, the performances improve with increasing $N$. For $\zeta = 0.6$, which is very close to the lower bound, the convergence to PCMS = 1 is slow and will require a very large sample size. For, $\zeta = 2$, the performance suffers (due to underfitting) in the small-$N$ regime, but does achieve

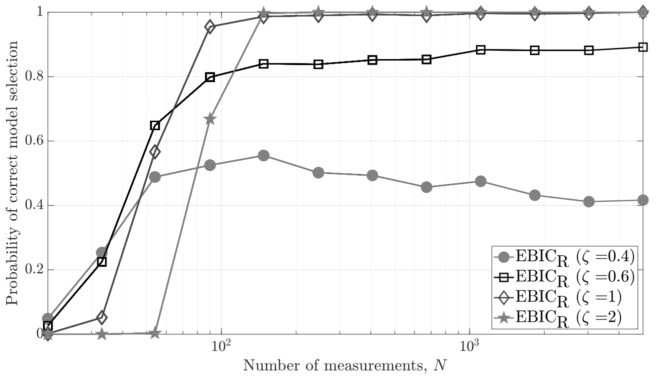


Fig. 1: PCMS vs $N$ with $k_0 = 5$, $\mathbf{x}_{\mathcal{S}} = [1, 1, 1, 1, 1]$, SNR = 5 dB, $p = N^d$ and $d = 1.1$.

perfect selection as $N$ increases. In this case, $\zeta = 1$ provides a much better overall performance for a broader range of $N$. A similar trend as in EBIC$_\mathrm{R}$ is observed even in EBIC and EFIC for different choices of $\gamma$ and $c$. Hence, to maintain fairness, the following TP settings are considered for further analysis: $\zeta = 1$ (EBIC$_\mathrm{R}$), $c = 1$ (EFIC), and $\gamma = 1$ (EBIC). For RRT, $\lim_{N\to\infty}$ PCMS $\to 1$ as $\alpha \to 0$. Hence, we choose $\alpha = 0.01$ [21]. For the knockoff filter, the false discovery rate (FDR) is set to FDR = 0.1 for all cases (which is the default setting in the R package and it gave the best results among other choices of FDR such as 0.05 and 0.2).

### A. Performance comparison with classical methods

In this section, we present simulation results for MS using classical methods in high-dimensional linear regression models and compare their performances with EBIC$_\mathrm{R}$. The purpose of these results is to highlight the limitations of the classical methods in dealing with large-$p$ small-$N$ scenarios. The classical methods used here are BIC [7], $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$[29], BIC$_\mathrm{R}$[27], gMDL [9], and PAL [11]. Here, we choose $k_0 = 5$ and the true parameter vector as $\mathbf{x}_{\mathcal{S}} = [5, 4, 3, 2, 1]^T$. Fig. 2 presents the plot for PCMS versus $N$ for SNR = 30 dB with $p = N^d$ where $d = 1.1$. The figure shows that EBIC$_\mathrm{R}$ ($\zeta = 1$) clearly surpasses the classical methods with huge differences in performance. In general, when $p$ is fixed and $N \to \infty$, the classical methods are consistent [27]. However, when $p$ is varying and grows with $N$, the consistency attribute does not hold any longer, hence, we see the decreasing performance trend in Fig. 2.

Fig. 3 illustrates the PCMS versus SNR in dB for fixed $N = 100$ and $p = 500$. This gives $d = \log(p)/\log(N) \approx 1.35$, and hence from Theorem 2, $\zeta > 1 - 1/2d \approx 0.63$. The first major observation from the figure is that EBIC$_\mathrm{R}$ ($\zeta = 1$) clearly outperforms all the classical methods by a huge margin. Secondly, for the considered setting, the performances of BIC$_\mathrm{R}$ and gMDL are quite similar followed by $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$. The criteria BIC$_\mathrm{R}$, gMDL and $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$ do achieve convergence to detection probability one but at the expense of very high values of SNR. The performances of PAL and BIC are extremely poor in this case, even in the high-SNR regions.

### B. Performance comparison with the latest methods

In Section VI-A, we highlighted the drawbacks of classical methods in MS under the high-dimensional setting. We observed that the performance of the classical methods collapses when $p$ grows with $N$ and the consistency property breaks down. In this section, we present simulation results for MS comparing EBIC$_\mathrm{R}$ to existing state-of-the-art methods, designed to deal with the large-$p$ small-$N$ scenarios.

To raise the simulation complexity, here we consider correlated predictors. The rows of $\mathbf{A}$ are generated as i.i.d. following $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where the covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is chosen as

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & & \vdots & \vdots \\ \rho & \rho & \dots & \rho & 1 \end{bmatrix}_{p\times p} . \tag{40}$$

For the above choice of $\mathbf{\Sigma}$, the $\rho$ denotes the level of correlation between the predictor $\mathbf{a}_i$'s. This structure of $\mathbf{\Sigma}$ further indicates that the $\mathbf{a}_i$'s are statistically equiangular. As a side remark, please note that for this model of $\mathbf{A}$, the matrix $\mathbf{M}_{\mathcal{I}}$ in the sparse Riesz assumption (38) is $\mathbf{M}_{\mathcal{I}} = \mathbf{\Sigma}_{\mathcal{I}}$, where $\mathbf{\Sigma}_{\mathcal{I}} \in \mathbb{R}^{k\times k}$ is a positive definite sub-covariance matrix of $\mathbf{\Sigma}$. To highlight the scale-invariant and consistent behaviour of EBIC$_\mathrm{R}$, we consider

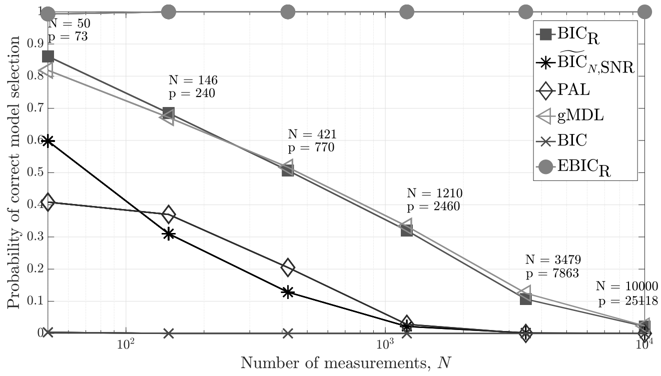


Fig. 2: The PCMS versus $N$ for SNR = 30 dB with $\mathbf{x}_{\mathcal{S}} = [5, 4, 3, 2, 1]$ and $p = N^d$ where $d = 1.1$.

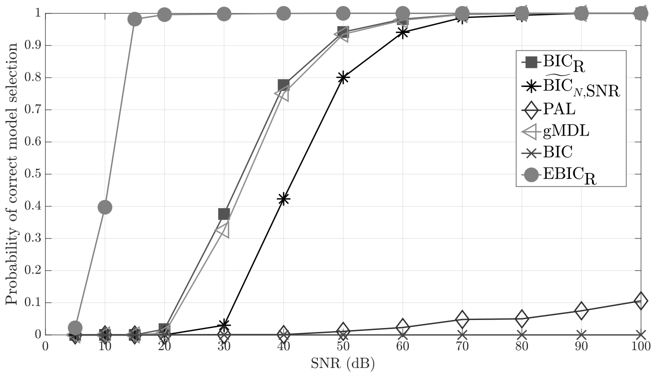
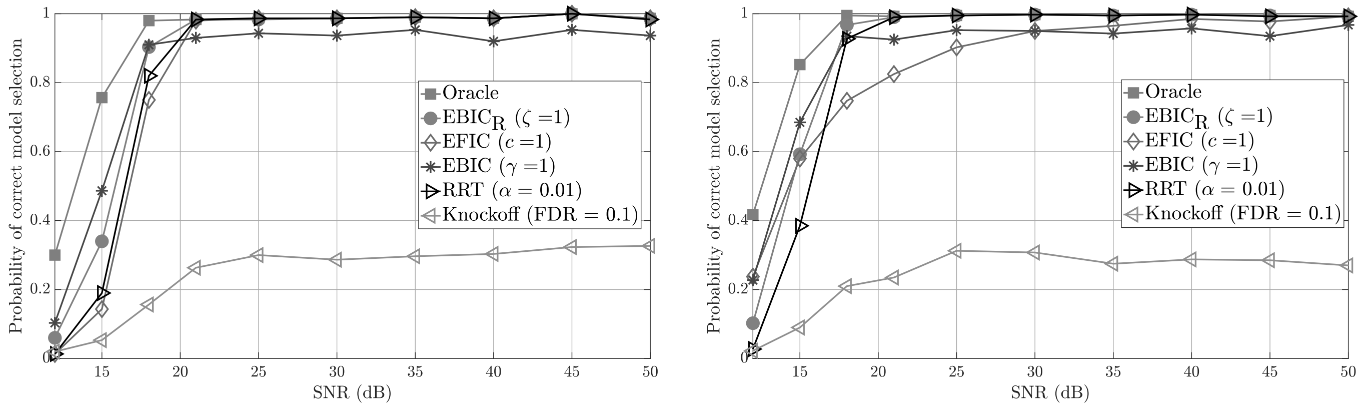


Fig. 3: The PCMS versus SNR (dB) for $N = 100$, $p = 500$ and $\mathbf{x}_{\mathcal{S}} = [5, 4, 3, 2, 1]$.



(a) Case I: $\mathbf{x}_{\mathcal{S}} = [0.055, -0.05, \ldots, 0.015, 0.01]^T$

(b) Case II: $\mathbf{x}_{\mathcal{S}} = [55, -50, \ldots, 15, 10]^T$

Fig. 4: The PCMS versus SNR (dB) for $N = 100$, $p = 500$, $k_0 = 10$, $\rho = 0.2$, $K = 30$.

two scenarios. In the first scenario, we choose $\mathbf{x}_{\mathcal{S}} = [0.055, -0.05, 0.045, -0.04, 0.035, -0.03, 0.025, -0.02, 0.015, 0.01]^T$, which we denote as Case I. In the second scenario, we choose $\mathbf{x}_{\mathcal{S}} = [55, -50, 45, -40, 35, -30, 25, -20, 15, 10]^T$ denoted as Case II. Observe that Case II is a scaled version of Case I, where the scaling factor is 1000.

*1) Model Selection versus SNR:* To simulate the PCMS versus SNR in a high-dimensional setting we fixed $N = 100$ and $p = 500$. This gives $d = \log(p)/\log(N) \approx 1.35$, hence, $\zeta > 1 - 1/2d \approx 0.63$. The correlation factor $\rho$ in (40) is chosen as 0.2 for both cases. Fig. 4 shows the empirical PCMS versus SNR (dB). Fig. 4a and Fig. 4b correspond to Case I and Case II, respectively. Both the figures depict a fixed-$N$ increasing-SNR scenario. Comparing the figures, the first clear observation is that, 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 SNR range. This illustrates the scaling problem present in EFIC that leads to either high underfitting or overfitting issues. This 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 depends on the value $\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2$, which in turn is influenced by the signal and noise powers $\sigma_s^2$ and $\sigma^2$, 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 Case I (Fig. 4a). On the contrary if $\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 \gg 1$, then DDPT $\ll 0$, thus lowering the overall penalty leading to overfitting issues (Case II, Fig. 4b). The second major observation is that EBIC is inconsistent when SNR is high but $N$ is small and fixed. This behaviour of EBIC is already reported in [14]. The performance of knockoff is also quite unsatisfactory compared to all methods even though it appears to be scale-invariant. In general, EFIC, RRT (for $\alpha \to 0$), and EBIC$_\mathrm{R}$ are consistent for increasing SNR scenarios given that $N$ is fixed, but while EBIC$_\mathrm{R}$ and RRT are invariant to data scaling EFIC is not.

*2) Model Selection versus N:* Fig. 5 illustrates the empirical PCMS versus $N$ for SNR = 8 dB, $\rho = 0.2$ and a fixed $p = 500$. Further we consider Case I for the choice of $\mathbf{x}_{\mathcal{S}}$, i.e., $[55, -50, \ldots, 15, 10]^T$. It depicts a low-SNR increasing-$N$ scenario. It is clearly seen that compared to the other criteria, EFIC suffers from the scaling issue and requires a large sample size to achieve detection probability one. It is interesting to note that the behavior of knockoff is quite subpar even in the large-$N$ scenario. Among all the criteria, the performance of EBIC and EBIC$_\mathrm{R}$ are closest to the oracle. Furthermore, observe that the performance of EBIC$_\mathrm{R}$ and EBIC are more or

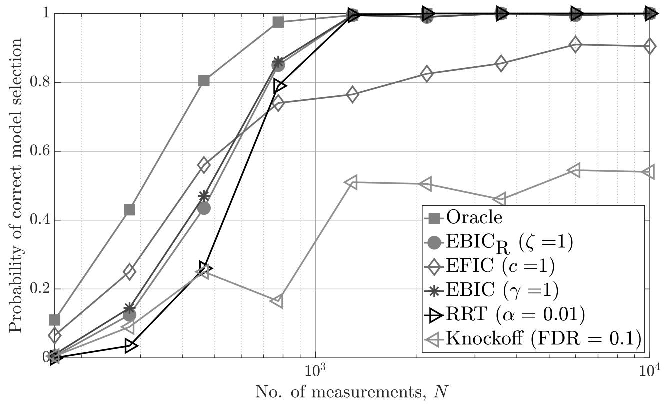


Fig. 5: The PCMS vs $N$ for SNR = 8 dB, $p = 500$.

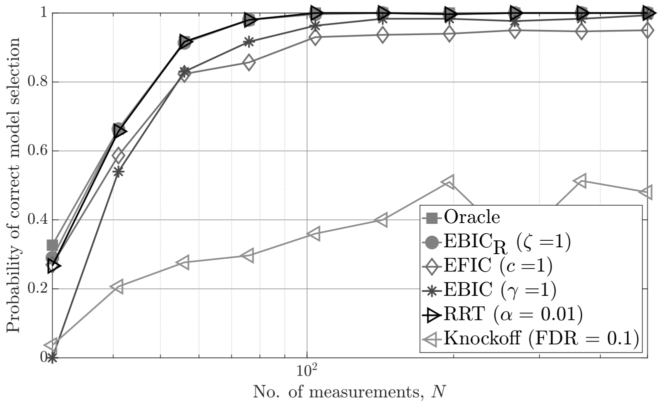
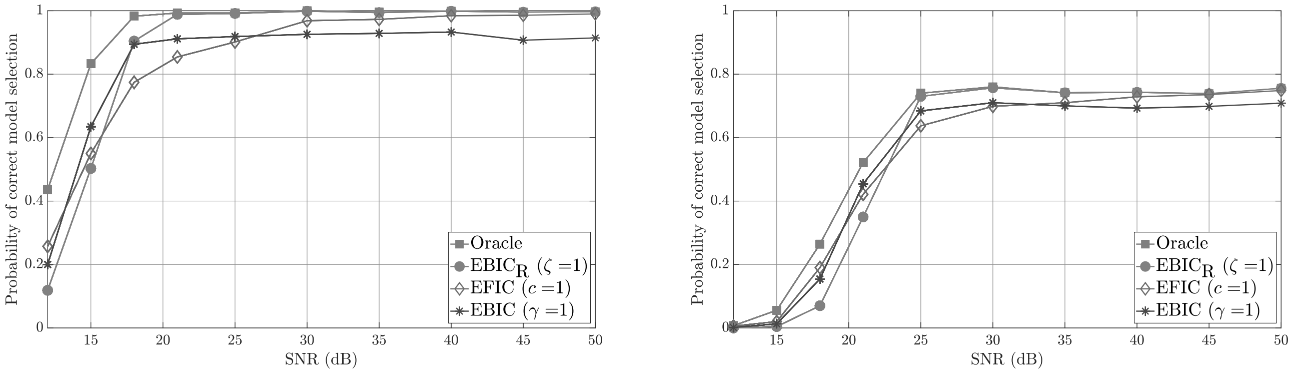


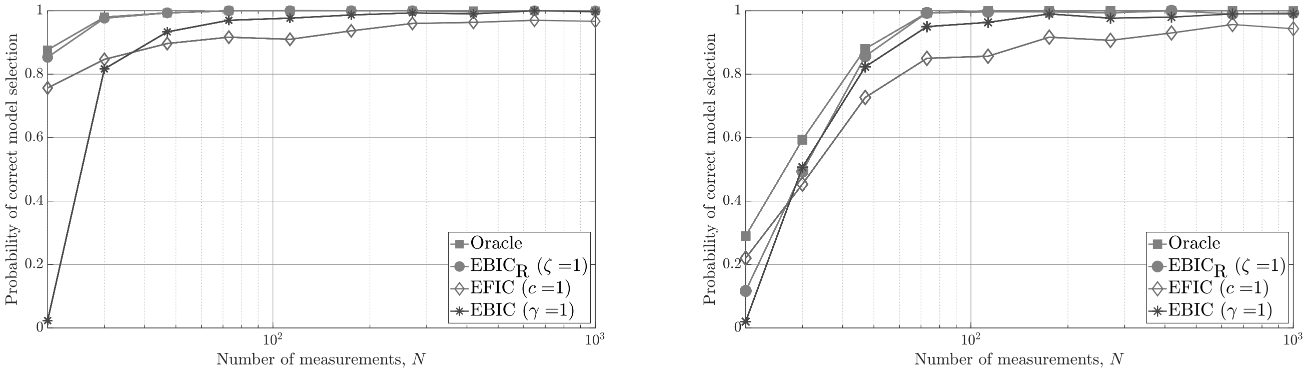
Fig. 6: The PCMS vs $N$ for SNR = 25 dB, $p = N^d$, $d = 1.2$.



(a) $\rho = 0.05$

(b) $\rho = 0.5$

Fig. 7: The PCMS versus SNR (dB) for $N = 55$ and $p = 1000$ and $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$.



(a) $\rho = 0.05$

(b) $\rho = 0.5$

Fig. 8: The PCMS versus $N$ for SNR = 25 dB and $p = N^d$, $d = 1.2$ and $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$.

less alike for the current setting. This is primarily because the SNR is low (8 dB) hence the $(k+2)\ln(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2)$ term of EBIC$_\text{R}$ behaves very close to a $\mathcal{O}(1)$ quantity for $k \geq k_0$. Thus, for low-SNR scenarios, the penalties of EBIC and EBIC$_\text{R}$ are similar, and as such the behaviour of these two criteria overlaps. However, this is not true in the high-SNR cases, which will be evident from the discussion following Fig. 6.

The plots shown in Fig. 4 and Fig. 5 represent fixed-$N$ increasing-SNR and low-SNR increasing-$N$ scenarios, respectively. In Fig. 6, we present a high-SNR increasing-$N$ case. Here, we consider a varying $p$ such that $p = N^d$ and $d = 1.2$. It is clearly observed that for high-SNR scenarios, EBIC$_\text{R}$ and RRT provide much faster convergence to oracle behaviour as compared to EBIC which requires a higher sample size to achieve detection probability one. Furthermore, we also notice that EFIC suffers from a higher false selection error and performs worse than EBIC in a certain region of the sample size. This clearly shows the effects of scaling on the behaviour of EFIC.

The simulation results so far covered two important aspects: (1) failure of classical methods in high-dimensional MS and (2) performance comparison of the improved criterion EBIC$_\text{R}$ with existing methods with a special highlight on the data scaling problem and high-SNR consistency. However, since a constant value of the correlation coefficient, $\rho$ is chosen, it is hard to visualize its effect for different values on the overall performance. In the final part, we emphasize the effect of $\rho$ on the MS performance. For this we choose two different values of $\rho$, one small and the other slightly larger i.e., $\rho = 0.05$ and $\rho = 0.5$. Fig. 7 presents PCMS vs SNR (dB) with $N = 55$ and $p = 1000$, where Fig. 7a corresponds to $\rho = 0.05$ and Fig.7b to $\rho = 0.5$. A straightforward comparison clearly reveals that as the $\rho$ increases, the performance curves of all methods shifts towards the right, i.e., a bigger SNR is required when $\rho = 0.5$ to reach the same PCMS when $\rho = 0.05$. In fact, the interesting observation is the oracle in Fig. 7b does not reach PCMS = 1 as SNR increases and gets saturated at some PCMS $< 0.8$. However, the performance of the other methods relative to the oracle is similar in both plots. Since the oracle performance is the upper bound, the MS methods are bounded by the oracle behaviour in this case.

Fig. 8 presents PCMS vs $N$, with a fixed SNR of 25 dB, where Fig. 8a corresponds to $\rho = 0.05$ and Fig. 8b to $\rho = 0.5$. A similar shift of performance curves to the right is observed here as well when $\rho$ is increased from 0.05 to 0.5. However, in this case, the PCMS of the methods in both plots tends to one as $N$ grows large given that the SNR is fixed. Also, the performance of the methods relative to the oracle remains identical where EBIC$_\text{R}$ provides higher PCMS compared to the other methods.

### C. Real data analysis

In this section, we apply different MS methods to real-world data and analyze the results. Two different datasets are considered, viz. (1) Air pollution and mortality in US metropolitan areas [38] and (2) Breast cancer gene expression data from The Cancer Genome Atlas [39]. For each dataset, we compute the model order (cardinality of the estimated model excluding the intercept term) selected by different methods for different values of their respective TP. However, the data analysis steps are different for both the datasets which are further discussed below.

*1) Air pollution and mortality data analysis:* In this dataset, $N = 60$ and $p = 15$, hence $N > p$. The response vector $\mathbf{y}$ denotes the total age-adjusted mortality from all causes (annual deaths per 100,000 people). The rows of the design matrix $\mathbf{A}$ are labeled with the name of the metropolitan area. Each column denotes a specific predictor variable, for example, SO2 (pollution potential of sulfur dioxide), House (population per household), etc. OMP is employed for predictor selection with $K = 15$. Note that a pre-processing step before using OMP is centering $\mathbf{y}$ (i.e., subtracting the mean) and standardizing the $\mathbf{A}$ matrix (i.e., columns have zero mean and variance one). However, while computing the information criterion scores we use the $\mathbf{y}$ and $\mathbf{A}$ without centering or standardizing, but append an all one column at the beginning of $\mathbf{A}$ to take care of the intercept term. Three different TP values are chosen for EBIC$_\mathrm{R}$, EFIC, and EBIC, viz., 0, 0.5, and 1. Note that, even though this is a low-dimensional scenario (i.e., $N > p$), but since $p > \sqrt{N} \approx 8$, thus as per [13], the original BIC is inconsistent. Here, $d = \log(p)/\log(N) \approx 0.661$. Therefore, for EBIC$_\mathrm{R}$, $\zeta > 1 - 1/2d \approx 0.25$ as recommended from theory. For RRT three values of $\alpha$ are chosen as well, viz. 0.5, 0.1, and 0.01. Further, MS results for CV-Lasso (using glmnet package in R) are also provided. In CV-Lasso, the best value of $\lambda$ is chosen using CV. Here, $\lambda_{\text{min}}$ is the value of $\lambda$ that gives minimum mean cross-validated error, while $\lambda_{\text{1se}}$ is the value of $\lambda$ that gives the most regularized model such that the cross-validated error is within one standard error of the minimum. Table I presents the MS results for the air pollution and mortality data. The $R^2$ (R-squared) value is computed for each selected model, which gives us an idea of the goodness of fit. The $R^2$ value is computed as $1 - \frac{\text{SSE}}{\text{SST}}$ where SST is the total sum of squares and SSE is the residual sum of squares, i.e., SST $= \sum_{i=1}^{N}(y_i - \bar{y})^2$ ($\bar{y}$ is the mean) and SSE $= \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$. Observe that $\zeta = \gamma = 0$ implies EBIC$_\mathrm{R}$ $\equiv$ BIC$_\mathrm{R}$ and EBIC $\equiv$ BIC. As we increase the TP from 0 to 1, the model order decreases. For RRT the estimated model order is the same (i.e., 2) for $\alpha = 0.1, 0.01$ and it is identical to that of EBIC$_\mathrm{R}$ for $\zeta = 0.5, 1$ ($> 0.25$). However, for $\alpha = 0.5$, the model order picked by RRT is 4 which is in line with EBIC$_\mathrm{R}$ with $\zeta = 0$. Also, for model orders, $k \geq 5$, the $R^2$ values are $> 0.7$ and close to that of the least-squares ($k = p = 15$). This gives an indication that $k \geq 5$ is most likely the overfitting region (under the assumption that $k_0 < p$).

| Method | Tuning Para. | Model Order | $R^2$ |
|---|---|---|---|
| EBIC$_\mathrm{R}$ | 0 | 4 | 0.683 |
| | 0.5 | 2 | 0.562 |
| | 1 | 2 | 0.562 |
| EFIC | 0 | 11 | 0.755 |
| | 0.5 | 6 | 0.735 |
| | 1 | 4 | 0.683 |
| EBIC | 0 | 5 | 0.717 |
| | 0.5 | 4 | 0.683 |
| | 1 | 4 | 0.683 |
| RRT | 0.5 | 4 | 0.683 |
| | 0.1 | 2 | 0.562 |
| | 0.01 | 2 | 0.562 |
| CV-Lasso | $\lambda_{\text{min}} = 2.02$ | 9 | 0.745 |
| | $\lambda_{\text{1se}} = 7.4$ | 5 | 0.717 |
| Least-squares | $\lambda = 0$ | $k = p = 15$ | 0.765 |

TABLE I: Results for air pollution and mortality data.

*2) Breast cancer gene expression data analysis:* This dataset comes from breast cancer tissue samples deposited to The Cancer Genome Atlas (TCGA) project. It compiles results obtained using Agilent mRNA expression microarrays. BRCA1 is the first gene identified that increases the risk of early onset breast cancer. Here we have expression measurements of 17,322 genes from 536 patients, hence $N = 536$ and $p = 17322$ ($p \gg N$). The response $\mathbf{y}$ denotes the gene expression measurement for BRCA1 gene and the design matrix $\mathbf{A}$ represents gene expression measurements for the remaining genes. In this case, we take a different approach to demonstrate the broader applicability of the MS criteria. The data is split into 75% training ($\mathcal{I}_{train}$) and 25% test dataset ($\mathcal{I}_{test}$) with $N_{train}$ and $N_{test}$ number of samples, respectively. This gives $N_{train} = 536 \times 0.75 = 402$ samples for estimating the model and $N_{test} = 536 \times 0.25 = 134$ for evaluating the mean squared prediction error (MSPE) = $1/N_{test} \sum_{i \in \mathcal{I}_{test}}(y_i - \hat{y}_i)^2$ of the chosen model, where $y_i$'s are responses in the test set $\mathcal{I}_{test}$ and $\hat{y}_i$ the estimated response. Please note that the MSPE measure was not used in the air pollution example because it contained very few data points and splitting the dataset into training and test sets led to unreliable results. Lasso is employed to generate the set of candidate models that corresponds to a particular value of $\lambda$ in the lasso path. Model scores are evaluated for each candidate model, and the final model is picked as the one with the minimum score. RRT is excluded in this case since in its current form RRT cannot be used to estimate the true model. Table II presents the results for the breast cancer data. We run the experiment 100 times with a random selection of training

| Method | Tuning Parameter | Mean Order | Std. Dev. Order | Average MSPE |
|---|---|---|---|---|
| EBIC$_\mathrm{R}$ | 0.3 | 6.45 | 1.51 | 0.240 |
| | 0.7 | 5.27 | 1.37 | 0.248 |
| | 1 | 4.95 | 1.13 | 0.255 |
| EFIC | 0.3 | 6.55 | 1.47 | 0.240 |
| | 0.7 | 5.28 | 1.34 | 0.248 |
| | 1 | 5.03 | 1.21 | 0.252 |
| EBIC | 0.3 | 6.82 | 1.74 | 0.240 |
| | 0.7 | 5.33 | 1.39 | 0.248 |
| | 1 | 5.04 | 1.21 | 0.252 |
| CV-Lasso | $\lambda_{\text{min}}$ | 139.10 | 62.39 | 0.256 |
| | $\lambda_{\text{1se}}$ | 44.77 | 19.12 | 0.230 |

TABLE II: Results for breast cancer gene expression data.

and test sets to evaluate the average MSPE, the empirical mean order selected by each method, and the standard deviation of the model order. Here, $d = \log(p)/\log(N_{train}) = 1.62$. Hence, as per Theorem 2, for EBIC$_{\text{R}}$, $\zeta > 1 - 1/2d \approx 0.69$ will guarantee consistency if the sample size is assumed to be large. However, it does not imply optimal $\zeta$ since it also depends on the true SNR, which is unknown. CV-lasso with $\lambda = \lambda_{\text{1se}}$ as the regularization parameter generates the lowest average MSPE (0.23). For the remaining MS criteria, the TP choice of $0.3$ generates the lowest average MSPE ($\approx 0.24$). However, note that the mean order for CV-lasso ($\lambda_{\text{1se}}$) is $\approx 45$ whereas, for the remaining MS criteria with TP = 0.3, the mean model order is $\approx 7$. If we compute the difference between the average MSPE we have $0.24 - 0.23 = 0.01$, which is practically negligible. This clearly highlights that CV overfits the final model and picks more parameters than needed. Furthermore, the standard deviation of the model order shows that there is a high variability in the model order for CV as compared to the rest. This indicates that for different training and test sets of the same dataset, CV will pick models with a wide range of model orders. This makes it an unreliable method for MS, especially in the high-dimensional scenario.

*3) Remarks from real data analysis:* In this section, we provide some further discussion on the results obtained in sections VI-C1 and VI-C2. For the air pollution data, EBIC$_{\text{R}}$ and RRT appear to have a consensus on the model order. EFIC seems to have high variability in the model orders for different values of the TP. This might be a result of the scaling issue with EFIC. EBIC gives decent and stable model orders for all the choices of parameters. Furthermore, CV-Lasso ($\lambda_{\min}$) is clearly an overfitting case while CV-Lasso ($\lambda_{\text{1se}}$) delivers parsimonious model order.

In the breast cancer data, increasing the TP from 0.3 to 1 led to an increase in the average MSPE. However, the differences are not that significant for this case. Model orders in the range of $5 - 7 \ll 17322$ provide a decent average MSPE. We see that choosing the TP above the theoretical consistency condition $> 0.69$ provides stable model order estimates around 5 for all information criteria. From the real data examples, we do not really see any strong arguments speaking against the previously discussed suggestion of using the TP $\zeta = 1$ in EBIC$_{\text{R}}$.

## VII. Conclusion

In this paper, we provided a new criterion, which is an extension of BIC$_{\text{R}}$, to handle model selection in sparse high-dimensional linear regression models employing sparse methods for predictor selection. The extended version is named EBIC$_{\text{R}}$, where the subscript 'R' stands for robust and it is a scale-invariant and consistent model selection criterion. Additionally, we analytically examined the behaviour of EBIC$_{\text{R}}$ as $\sigma^2 \to 0$ and as $N \to \infty$. In both cases, it is shown that the probability of detecting the true model approaches one. The paper further highlighted the data scaling issue present in EFIC, which is a consistent criterion for both large sample size and high-SNR scenarios. Extensive simulation results show that the performance of EBIC$_{\text{R}}$ is either similar or superior to that of EBIC, EFIC, RRT, MBT, and knockoff filters.

## Appendix A
## Proof of Theorem 1

*Proof.* The proof consists of two parts. In part (a) we show that the probability of overfitting ($\mathcal{S} \subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}}$) tends to 0 as $\sigma^2 \to 0$, which in this case is equivalent to showing $\lim_{\sigma^2 \to 0} \Pr(\mathcal{C}_1) = 1$, cf. (36). In part (b) we show that the probability of misfitting ($\mathcal{S} \not\subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}}$) also tends to 0 as $\sigma^2 \to 0$, which is equivalent to $\lim_{\sigma^2 \to 0} \Pr(\mathcal{C}_2) = 1$, cf. (37).

($a$) *Over-fitting case* ($\mathcal{S} \subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}}$): Consider the set of overfitted subsets having cardinality $k$, which we have denoted as $\mathcal{I}_o^k$. Let $\mathcal{I}_j$ denote the $j$th subset in the set $\mathcal{I}_o^k$. The total number of subsets in $\mathcal{I}_o^k$ is $\binom{p-k_0}{\Delta}$ where $\Delta = k - k_0$ . For any overfitted subset $\mathcal{I}_j \in \mathcal{I}_o^k$, consider the following inequality

$$\text{EBIC}_\text{R}(\mathcal{S}) < \text{EBIC}_\text{R}(\mathcal{I}_j), \quad \mathcal{I}_j \in \mathcal{I}_o^k, \tag{41}$$

where $j = 1, \ldots, \binom{p-k_0}{\Delta}$. Using the relation $p = N^d$ and after some straightforward rearrangement of (41) we get

$$\begin{aligned}(N - k_0 - 2)\ln \hat{\sigma}^2_{\mathcal{S}} - (N - k - 2)\ln \hat{\sigma}^2_{\mathcal{I}_j} \\ -\Delta(1 + 2\zeta d)\ln N - \Delta \ln \hat{\sigma}_0^2 + \Delta \ln 2\pi < 0.\end{aligned} \tag{42}$$

Let us define a random variable $X_{\mathcal{I}_j} = \hat{\sigma}^2_{\mathcal{I}_j}/\sigma^2$, then

$$N \cdot X_{\mathcal{I}_j} \sim \chi^2_{N-k}, \quad \forall\, \mathcal{I}_j \in \mathcal{I}_o^k. \tag{43}$$

This implies that the variables $X_{\mathcal{I}_j}$ are independent of $\sigma^2$. Now, we can express

$$(N - k - 2)\ln \hat{\sigma}^2_{\mathcal{I}_j} = \ln X_{\mathcal{I}_j}^{N-k-2} + (N - k - 2)\ln \sigma^2, \tag{44}$$

and similarly by defining $X_{\mathcal{S}} = \hat{\sigma}^2_{\mathcal{S}}/\sigma^2$ we get

$$(N - k_0 - 2)\ln \hat{\sigma}^2_{\mathcal{S}} = \ln X_{\mathcal{S}}^{N-k_0-2} + (N - k_0 - 2)\ln \sigma^2. \tag{45}$$

Using (44) and (45) in (42) and after exponentiation we get

$$\left(\frac{X_{\mathcal{S}}^{N-k_0-2}}{X_{\mathcal{I}_j}^{N-k-2}}\right)\left(\frac{1}{N}\right)^{\Delta(1+2\zeta d)}\left(\frac{2\pi}{\hat{\sigma}_0^2}\right)^{\Delta} < \left(\frac{1}{\sigma^2}\right)^{\Delta}. \tag{46}$$

Let $E^k_{\mathcal{I}_j}$ denote the entire left hand-side and let $\eta_k$ denote the right-hand side of the inequality in (46). Let $\mathcal{I}^* \in \mathcal{I}_o^k$ denote the subset that produces the maximum value of $E^k_{\mathcal{I}_j}$ among all such subsets $\mathcal{I}_j \in \mathcal{I}_o^k$ . Then, let us denote

$$E^k_{\mathcal{I}^*} = \max_{\mathcal{I}_j \in \mathcal{I}_o^k}\left\{E^k_{\mathcal{I}_j}\right\}, \quad j = 1, 2, \ldots, \binom{p-k_0}{\Delta}. \tag{47}$$

The condition $\mathcal{C}_1$ in (36) is satisfied as $\sigma^2 \to 0$ under the event $E^k_{\mathcal{I}^*} < \eta_k$, for all $k = k_0 + 1, \ldots, K$. Now, we can express the probability that $E^k_{\mathcal{I}^*} < \eta_k$ as follows

$$\begin{aligned}\Pr\left(E^k_{\mathcal{I}^*} < \eta_k\right) &= \Pr\left\{\bigcap_{j=1}^{\binom{p-k_0}{\Delta}}\left(E^k_{\mathcal{I}_j} < \eta_k\right)\right\} \\ &= 1 - \Pr\left\{\bigcup_{j=1}^{\binom{p-k_0}{\Delta}}\left(E^k_{\mathcal{I}_j} > \eta_k\right)\right\} \\ &\geq 1 - \binom{p-k_0}{\Delta}\Pr\left(E^k_{\mathcal{I}_j} > \eta_k\right) \\ \implies \Pr\left(E^k_{\mathcal{I}^*} > \eta_k\right) &\leq \binom{p-k_0}{\Delta}\Pr\left(E^k_{\mathcal{I}_j} > \eta_k\right),\end{aligned} \tag{48}$$

where the inequality follows from the union bound. Now consider the following probability $\Pr\left\{E^k_{\mathcal{I}_j} > \eta_k\right\}$ for any arbitrary subset $\mathcal{I}_j \in \mathcal{I}^k_o$, which can be expressed as

$$\Pr\left\{\left(\frac{X_{\mathcal{S}}^{N-k_0-2}}{X_{\mathcal{I}_j}^{N-k-2}}\right)\left(\frac{1}{N}\right)^{\Delta(1+2\zeta d)}\left(\frac{2\pi}{\hat{\sigma}_0^2}\right)^{\Delta} > \left(\frac{1}{\sigma^2}\right)^{\Delta}\right\}. \tag{49}$$

Let $W = X_{\mathcal{S}}^{N-k_0-2}/X_{\mathcal{I}_j}^{N-k-2}$. Notice that the random variable $W$ is independent of the noise variance $\sigma^2$ and since $N$ is fixed $W$ is bounded as $\sigma^2 \to 0$. Furthermore, $\lim\limits_{\sigma^2\to 0} \hat{\sigma}_0^2 = c$ (see Appendix C) and the right-hand side of the inequality in (49) grows unbounded as $\sigma^2 \to 0$. Thus, we have

$$\lim_{\sigma^2\to 0} \Pr\left\{E^k_{\mathcal{I}_j} > \eta_k\right\} = 0. \tag{50}$$

Therefore, using (48) and the result in (50), we have

$$\lim_{\sigma^2\to 0} \Pr\left(E^k_{\mathcal{I}^*} > \eta_k\right) = 0, \qquad \forall\ k = k_0+1,\ldots,K. \tag{51}$$

Finally, using the union bound, and the result in (51), we get

$$\begin{aligned}\Pr\left\{\mathcal{C}_1\right\} =& \Pr\left\{\bigcap_{k=k_0+1}^{K} E^k_{\mathcal{I}^*} < \eta_k\right\}\\ \geq& 1 - \sum_{k=k_0+1}^{K} \Pr\left\{E^k_{\mathcal{I}^*} > \eta_k\right\} \to 1, \end{aligned} \tag{52}$$

as $\sigma^2 \to 0$.

$(b)$ *Misfitting case* ($\mathcal{S} \not\subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}}$): Let $\mathcal{I}_j$ be any arbitrary $j$th subset belonging to the set of misfitted subsets of dimension $k$, i.e., $\mathcal{I}^k_m$. We consider the following inequality

$$\text{EBIC}_\text{R}(\mathcal{S}) < \text{EBIC}_\text{R}(\mathcal{I}_j), \quad \mathcal{I}_j \in \mathcal{I}^k_m, \tag{53}$$

where $j = 1,\ldots,t$. Here, $t$ denotes the total number of subsets in the set $\mathcal{I}^k_m$ and $t = \binom{p}{k}$ if $k < k_0$, otherwise $t = \binom{p}{k} - \binom{p-k_0}{\Delta}$ if $k \geq k_0$, where $\Delta = k - k_0$. Denoting $X_{\mathcal{S}} = \hat{\sigma}^2_{\mathcal{S}}/\sigma^2$, rearranging and applying exponentiation we can express (53) as

$$\left(\frac{X_{\mathcal{S}}^{N-k_0-2}}{(\hat{\sigma}^2_{\mathcal{I}_j})^{N-k-2}}\right)\left(\frac{1}{N}\right)^{\Delta(1+2\zeta d)}\left(\frac{2\pi}{\hat{\sigma}_0^2}\right)^{\Delta} < \left(\frac{1}{\sigma^2}\right)^{N-k_0-2}. \tag{54}$$

Similar to the overfitting case, let $E^k_{\mathcal{I}_j}$ denote the entire left-hand side and $\eta$ the right-hand side of (54). Also, let $E^k_{\mathcal{I}^*} = \max\limits_{\mathcal{I}_j\in\mathcal{I}^k_m}\left\{E^k_{\mathcal{I}_j}\right\}$ for $j = 1,\ldots,t$, where $\mathcal{I}^*$ is the subset that leads to the maximum value of $E^k_{\mathcal{I}_j}$ among all such subsets of dimension $k$. The condition $\mathcal{C}_2$ in (37) is satisfied as $\sigma^2 \to 0$ under the event $E^k_{\mathcal{I}^*} < \eta$, for all $k = 1,\ldots,K$. Now, we can express the probability that $E^k_{\mathcal{I}^*} < \eta$ as

$$\begin{aligned}\Pr\left(E^k_{\mathcal{I}^*} < \eta\right) =& \Pr\left\{\bigcap_{j=1}^{t}\left(E^k_{\mathcal{I}_j} < \eta\right)\right\}\\ \implies \Pr\left(E^k_{\mathcal{I}^*} > \eta\right) \leq& t\Pr\left(E^k_{\mathcal{I}_j} > \eta\right), \end{aligned} \tag{55}$$

where the inequality follows from the union bound. Now consider the following probability for any arbitrary subset $\mathcal{I}_j \in \mathcal{I}^k_m$

$$\begin{aligned}\Pr\left(E^k_{\mathcal{I}_j} > \eta\right) = \Pr\Bigg\{&\left(\frac{X_{\mathcal{S}}^{N-k_0-2}}{(\hat{\sigma}^2_{\mathcal{I}_j})^{N-k-2}}\right)\left(\frac{1}{N}\right)^{\Delta(1+2\zeta d)}\\ &\times\left(\frac{2\pi}{\hat{\sigma}_0^2}\right)^{\Delta} > \left(\frac{1}{\sigma^2}\right)^{N-k_0-2}\Bigg\}. \end{aligned} \tag{56}$$

Here, $X_{\mathcal{S}}^{N-k_0-2}$ is independent of $\sigma^2$ and $N$ is fixed, therefore $X_{\mathcal{S}}^{N-k_0-2}$ is bounded as $\sigma^2 \to 0$. Also $\hat{\sigma}^2_{\mathcal{I}_j} \to \|\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_2^2/N$ in probability as $\sigma^2 \to 0$ and since we are in the misfitting scenario, from Lemma 4 in Appendix E we have $\|\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_2^2/N > 0$. Furthermore, $\lim\limits_{\sigma^2\to 0}\hat{\sigma}_0^2 = \text{const.}$ (see Appendix C) and the right-hand side of the inequality in (56) grows unbounded as $\sigma^2 \to 0$. Hence,

$$\lim_{\sigma^2\to 0}\Pr\left\{E^k_{\mathcal{I}_j} > \eta\right\} = 0. \tag{57}$$

Using (55) and the result in (57) we get

$$\lim_{\sigma^2\to 0}\Pr\left\{E^k_{\mathcal{I}^*} > \eta\right\} = 0, \quad \forall\ k = 1,\ldots,K. \tag{58}$$

Finally, using the union bound and the result in (58), we get

$$\Pr\left\{\mathcal{C}_2\right\} \geq 1 - \sum_{k=1}^{K}\Pr\left\{E^k_{\mathcal{I}^*} > \eta\right\} \to 1 \quad \text{as} \quad \sigma^2 \to 0. \tag{59}$$

From (52) and (59) we can conclude that $\text{EBIC}_\text{R}$ is consistent as $\sigma^2 \to 0$, which proves Theorem 1.

## Appendix B
## Proof of Theorem 2

*Proof.* As in the previous section, we have two parts of the proof. Part $(a)$ is the overfitting case where we show that $\Pr(\mathcal{C}_1) \to 1$ (cf. (36)) as $N \to \infty$ and part $(b)$ is the misfitting case where we show that $\Pr(\mathcal{C}_2) \to 1$ (cf. (37)) as $N \to \infty$.

$(a)$ *Overfitting case* ($\mathcal{S} \subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}}$): Let $\mathcal{I}_j \in \mathcal{I}^k_o$ be any overfitted subset of dimension $k$. Consider the following inequality

$$\text{EBIC}_\text{R}(\mathcal{I}_j) > \text{EBIC}_\text{R}(\mathcal{S}), \quad \mathcal{I}_j \in \mathcal{I}^k_o. \tag{60}$$

Denoting $\Delta = k - k_0$ and rearranging (60) we get

$$\begin{aligned}(N-k-2)\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right) &+ \Delta(1+2\zeta d)\ln N\\ &+\Delta\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{S}}}\right) - \Delta\ln 2\pi > 0. \end{aligned} \tag{61}$$

Let $E^k_{\mathcal{I}_j}$ denote the entire left side of the inequality (61) and $\mathcal{I}^*$ denote the subset that leads to the minimum value of $E^k_{\mathcal{I}_j}$ among all such subsets of dimension $k$. Hence,

$$E^k_{\mathcal{I}^*} = \min_{\mathcal{I}_j\in\mathcal{I}^k_o}\left\{E^k_{\mathcal{I}_j}\right\}, \quad j = 1,2,\ldots,\binom{p-k_0}{\Delta}. \tag{62}$$

The condition $\mathcal{C}_1$ in (36) is satisfied as $N \to \infty$ under the event $E^k_{\mathcal{I}^*} > 0$, for all $k = k_0 + 1, \ldots, K$. Expanding the ratio we have

$$\begin{aligned}
\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right) &= \ln\left(\frac{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}}{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}\right) \\
&= \ln\left[\frac{\mathbf{e}^T\left(\mathbf{I} - \boldsymbol{\Pi}_{\mathcal{I}_j} + \boldsymbol{\Pi}_{\mathcal{S}} - \boldsymbol{\Pi}_{\mathcal{S}}\right)\mathbf{e}}{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}\right] \\
&= \ln\left(\frac{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e} - \mathbf{e}^T\boldsymbol{\Pi}_{\mathcal{I}_j\backslash\mathcal{S}}\mathbf{e}}{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}\right) \\
&= \ln\left(1 - \frac{\mathbf{e}^T\boldsymbol{\Pi}_{\mathcal{I}_j\backslash\mathcal{S}}}{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}\right), \qquad (63)
\end{aligned}$$

where $\boldsymbol{\Pi}_{\mathcal{I}_j\backslash\mathcal{S}} = \boldsymbol{\Pi}_{\mathcal{I}_j} - \boldsymbol{\Pi}_{\mathcal{S}}$. Now we can write

$$\begin{aligned}
&\min_{1\le j\le T}\left\{(N-k-2)\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right)\right\} = \\
&(N-k-2)\ln\left[1 - \frac{\max\limits_{1\le j\le T}\left\{\left(\mathbf{e}^T\boldsymbol{\Pi}_{\mathcal{I}_j\backslash\mathcal{S}}\mathbf{e}\right)/\sigma^2\right\}}{\left(\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}\right)/\sigma^2}\right], \qquad (64)
\end{aligned}$$

where $T = \binom{p-k_0}{\Delta}$. Now the term, $(\mathbf{e}^T\boldsymbol{\Pi}_{\mathcal{I}_j\backslash\mathcal{S}}\mathbf{e})/\sigma^2 \sim \chi^2_{\Delta}$ (see Appendix D). Then from Lemma 2 in Appendix E we have the following upper bound

$$\max_{1\le j\le T}\left\{(\mathbf{e}^T\boldsymbol{\Pi}_{\mathcal{I}_j\backslash\mathcal{S}}\mathbf{e})/\sigma^2\right\} \le \Delta + 2\sqrt{\Delta\psi\ln T} + 2\psi\ln T, \qquad (65)$$

with probability approaching one as $N \to \infty$ if $\psi > 1$. Now, for sufficiently large $p = N^d$ we can write $\ln T = \ln\binom{p-k_0}{\Delta} \approx \Delta d\ln N$. This gives

$$\begin{aligned}
\max_{1\le j\le T}\left\{(\mathbf{e}^T\boldsymbol{\Pi}_{\mathcal{I}_j\backslash\mathcal{S}}\mathbf{e})/\sigma^2\right\} &\le \Delta + 2\Delta\sqrt{\psi d\ln N} + 2\psi\Delta d\ln N \\
&= 2\psi\Delta d\ln N\left(1 + \frac{1}{\sqrt{\psi d\ln N}} + \frac{1}{2\psi d\ln N}\right) \\
&\approx 2\psi\Delta d\ln N, \qquad (66)
\end{aligned}$$

as $N$ grows large. Furthermore, the term in the denominator in (64), $(\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e})/\sigma^2 \sim \chi^2_{N-k_0}$ and based on the law of large numbers tends to $N - k_0 \approx N$. Therefore, using (66) in (64) and $(N-k-2) \approx N$ under the large-$N$ approximation we get

$$\begin{aligned}
\min_{1\le j\le T}\left\{N\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right)\right\} &\ge N\ln\left(1 - \frac{2\Delta\psi d\ln N}{N}\right) \\
&\approx -2\Delta\psi d\ln N, \qquad (67)
\end{aligned}$$

where the last approximation follows by linearization of the logarithm for small $2\Delta\psi d\ln N/N$ value. Thus, we can write

$$\begin{aligned}
E^k_{\mathcal{I}^*} &\ge -2\Delta\psi d\ln N + \Delta(1+2\zeta d)\ln N + \Delta\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{S}}}\right) \\
&\qquad - \Delta\ln 2\pi \\
&= \Delta\left(1 + 2\zeta d - 2\psi d\right)\ln N + \Delta\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{S}}}\right) - \Delta\ln 2\pi. \qquad (68)
\end{aligned}$$

Since $\lim_{N\to\infty}\hat{\sigma}^2_0 = \text{const.} > 0$ (see Appendix C) and $\lim_{N\to\infty}\hat{\sigma}^2_{\mathcal{S}} = \sigma^2$ (see Appendix D), $E^k_{\mathcal{I}^*} \to \infty$ as $N \to \infty$ for all $k = k_0+1,\ldots,K$ under the condition $1 + 2\zeta d - 2\psi d > 0$ for any $\psi > 1$. Hence, the lower bound on $\zeta$ becomes

$$\boxed{\zeta > 1 - \frac{1}{2d}}. \qquad (69)$$

From the above analysis, we can say that

$$\lim_{N\to\infty}\Pr\left\{E^k_{\mathcal{I}^*} < 0\right\} = 0, \quad \forall\; k = k_0+1,\ldots,K. \qquad (70)$$

Finally, using the union bound and the result in (70) we can express the probability of $\mathcal{C}_1$ (36) happening as

$$\begin{aligned}
\Pr\{\mathcal{C}_1\} &= \Pr\left\{\bigcap_{k=k_0+1}^{K} E^k_{\mathcal{I}^*} > 0\right\} \\
&\ge 1 - \sum_{k=k_0+1}^{K}\Pr\left\{E^k_{\mathcal{I}^*} < 0\right\} \to 1 \qquad (71)
\end{aligned}$$

as $N \to \infty$.

($b$) *Misfitting case* ($\mathcal{S} \not\subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}}$): Let $\mathcal{I}_j \in \mathcal{I}^k_m$ be any misfitted subset of dimension $k$. Consider the following inequality

$$\text{EBIC}_\text{R}(\mathcal{I}_j) > \text{EBIC}_\text{R}(\mathcal{S}), \quad \mathcal{I}_j \in \mathcal{I}^k_m. \qquad (72)$$

Denoting $\Delta = k - k_0$ and rearranging (72) we get

$$\begin{aligned}
(N-k-2)\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right) + (1+2\zeta d)\Delta\ln N \\
+\Delta\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{S}}}\right) + \Delta\ln\left(\frac{1}{2\pi}\right) > 0.
\end{aligned} \qquad (73)$$

Let $E^k_{\mathcal{I}_j}$ denote the entire left hand side of the inequality in (73) and $\mathcal{I}^*$ denote the subset that generates the minimum value of $E^k_{\mathcal{I}_j}$ among all such subsets of dimension $k$. Then we have

$$E^k_{\mathcal{I}^*} = \min_{\mathcal{I}_j\in\mathcal{I}^k_m}\left\{E^k_{\mathcal{I}_j}\right\}, \qquad j = 1,2,\ldots,T, \qquad (74)$$

where $T = \binom{p}{k}$ if $k < k_0$ otherwise $T = \binom{p}{k} - \binom{p-k_0}{\Delta}$ if $k \ge k_0$. The condition $\mathcal{C}_2$ in (37) is satisfied as $N \to \infty$ under the event $E^k_{\mathcal{I}^*} > 0$, for all $k = 1,\ldots,K$. Now, let $\mathbf{u} = \mathbb{E}[\mathbf{y}] = \mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}$. Using this, the ratio $\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}$ can be expanded as

$$\begin{aligned}
\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}} &= \frac{\mathbf{y}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{y}}{\mathbf{y}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{y}} = \frac{(\mathbf{u}+\mathbf{e})^T\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_j}(\mathbf{u}+\mathbf{e})}{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}} \\
&= \frac{\mathbf{u}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u} + 2\sigma\sqrt{\mathbf{u}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}}\cdot Z_j + \mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}}{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}, \qquad (75)
\end{aligned}$$

where

$$Z_j = \frac{\mathbf{u}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}}{\sigma\sqrt{\mathbf{u}^T\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}}} \sim \mathcal{N}(0,1). \qquad (76)$$

Now

$$\begin{aligned}
&\min_{1\leq j\leq T}\{\hat{\sigma}^2_{\mathcal{I}_j}/\hat{\sigma}^2_{\mathcal{S}}\} = \\
&\min_{1\leq j\leq T}\left\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u} + 2\sigma\sqrt{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}}\cdot Z_j + \mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}\right\}\Big/\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e} \\
&\geq \Bigg[\min_{1\leq j\leq T}\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}\} + \sigma^2\min_{1\leq j\leq T}\{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}/\sigma^2\} \\
&\quad - 2\sigma\sqrt{\max_{1\leq j\leq T}\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}\}}\cdot\max_{1\leq j\leq T}\{Z_j\}\Bigg]\Big/\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}. \quad (77)
\end{aligned}$$

In the misfitting scenario we have two cases: (i) $k < k_0$ (ii) $k \geq k_0$. We consider case (i) in our further analysis, which also encapsulates case (ii). For $k < k_0$ we have $\ln T = \ln\binom{p}{k} \approx kd\ln N$. Therefore, using the result in Lemma 2 we have the following lower bound under large-$N$ approximation

$$\begin{aligned}
\min_{1\leq j\leq T}\left\{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}/\sigma^2\right\} &= \mathbf{e}^T\mathbf{e}/\sigma^2 - \max_{1\leq j\leq T}\left\{\mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}_j}\mathbf{e}/\sigma^2\right\} \\
&\geq\ N - 2\psi' kd\ln N, \quad (78)
\end{aligned}$$

where $\psi' > 1$ and $\mathbf{e}^T\mathbf{e}/\sigma^2 \approx N$ for large-$N$. Furthermore, from the result in Lemma 3 we have the following upper bound

$$\max_{1\leq j\leq T}\{Z_j\} \leq \sqrt{2\psi' kd\ln N}, \quad (79)$$

where $\psi' > 1$. Now, let $C_{\text{min}} = \min_{1\leq j\leq T}\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}\}$ and $C_{\text{max}} = \max_{1\leq j\leq T}\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}\}$. Also as $N \to \infty$ we can approximate $(N - k - 2) \approx N$ and $\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e} \approx \sigma^2 N$. Using this, and the results in (78) and (79) we get

$$\begin{aligned}
&\min_{1\leq j\leq T}\left\{N\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right)\right\} = N\ln\left[\min_{1\leq j\leq T}\left\{\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right\}\right] \\
&\geq N\ln\Bigg[\Bigg\{C_{\text{min}} - 2\sigma\sqrt{C_{\text{max}}}\cdot\sqrt{2\psi' kd\ln N} \\
&\qquad + \sigma^2\left(N - 2\psi' kd\ln N\right)\Bigg\}\Big/\sigma^2 N\Bigg]. \quad (80)
\end{aligned}$$

Now, observe that $C_{\min} = \mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}^*}\mathbf{u} = \mathbf{x}^T_{\mathcal{S}}\mathbf{A}^T_{\mathcal{S}}\mathbf{\Pi}^{\perp}_{\mathcal{I}^*}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}$. Since, we are in the misfitting scenario, from Lemma 4, in Appendix E, we can express $C_{\min} = Nb_{\min}$ where $b_{\min} = \text{const.} > 0$. Similarly, $C_{\max} = Nb_{\max}$ where $b_{\max} = \text{const.} > 0$ and $0 < b_{\min} \leq b_{\max}$. Hence, we can rewrite (80) as

$$\begin{aligned}
&\min_{1\leq j\leq T}\left\{N\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right)\right\} \geq \\
&N\ln\left(1 + \frac{b_{\text{min}}}{\sigma^2} - \frac{2\sqrt{b_{\text{max}}}}{\sigma}\sqrt{\frac{2\psi' kd\ln N}{N}} - \frac{2\psi' kd\ln N}{N}\right) \\
&\approx N\ln\left(1 + \frac{b_{\text{min}}}{\sigma^2}\right) \quad (81)
\end{aligned}$$

as $N$ grows large. For $k < k_0$, we get $\Delta < 0$, therefore, in this case we have

$$\begin{aligned}
E^k_{\mathcal{I}^*} \geq N\ln\left(1 + \frac{b_{\text{min}}}{\sigma^2}\right) &- |\Delta|(1 + 2\zeta d)\ln N \\
&-|\Delta|\ln\left(\frac{\hat{\sigma}^2_0}{2\pi\hat{\sigma}^2_{\mathcal{S}}}\right) \to \infty
\end{aligned} \quad (82)$$

as $N \to \infty$ for all $k = 1, \ldots, K$, since $N\ln(1 + b_{\text{min}}/\sigma^2)$ is the dominating term as it tends to infinity much faster than the $\ln N$ term and $\lim_{N\to\infty}\hat{\sigma}^2_0 = \text{const.} > 0$ (see Appendix C) and $\lim_{N\to\infty}\hat{\sigma}^2_{\mathcal{S}} = \sigma^2$ (see Appendix D). From the above analysis, we can say that

$$\lim_{N\to\infty}\Pr\left\{E^k_{\mathcal{I}^*} < 0\right\} = 0, \quad \forall\ k = 1, \ldots, K. \quad (83)$$

Finally, using the union bound and the result in (83) we can express the probability of $\mathcal{C}_2$ (37) happening as

$$\begin{aligned}
\Pr\{\mathcal{C}_2\} &= \Pr\left\{\bigcap_{k=1}^{K} E^k_{\mathcal{I}^*} > 0\right\} \\
&\geq 1 - \sum_{k=1}^{K}\Pr\left\{E^k_{\mathcal{I}^*} < 0\right\} \to 1 \quad \text{as} \quad N \to \infty. \quad (84)
\end{aligned}$$

From (71) and (84) we can conclude that EBIC$_{\text{R}}$ is consistent as $N \to \infty$, which proves Theorem 2.

## Appendix C
### Statistical analysis of the factor $\hat{\sigma}^2_0$

From the generating model (1), the true data vector follows $\mathbf{y} \sim \mathcal{N}\left(\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}, \sigma^2\mathbf{I}_N\right)$. Consider the factor $\hat{\sigma}^2_0$, which is defined as

$$\hat{\sigma}^2_0 = \frac{\|\mathbf{y}\|^2_2}{N} = \left(\frac{\sigma^2}{N}\right)\frac{\mathbf{y}^T\mathbf{I}_N\mathbf{y}}{\sigma^2}. \quad (85)$$

From Lemma 1 in Appendix E we have

$$\frac{\mathbf{y}^T\mathbf{I}_N\mathbf{y}}{\sigma^2} \sim \chi^2_N(\lambda) \text{ where } \lambda = \frac{\|\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|^2_2}{\sigma^2}. \quad (86)$$

This implies that $\left(\frac{N}{\sigma^2}\right)\hat{\sigma}^2_0 \sim \chi^2_N(\lambda)$. Therefore, the mean and variance of $\hat{\sigma}^2_0$ are:

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2_0] &= \frac{\sigma^2}{N}(N + \lambda) = \sigma^2 + \frac{\|\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|^2_2}{N} \\
\text{Var}[\hat{\sigma}^2_0] &= 2\frac{\sigma^4}{N^2}(N + 2\lambda) = 2\frac{\sigma^4}{N} + 4\frac{\sigma^2}{N^2}\|\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|^2_2.
\end{aligned} \quad (87)$$

Hence, for a fixed $N$,

$$\lim_{\sigma^2\to 0}\mathbb{E}[\hat{\sigma}^2_0] = \frac{\|\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|^2_2}{N} \quad \& \quad \lim_{\sigma^2\to 0}\text{Var}[\hat{\sigma}^2_0] = 0. \quad (88)$$

Further, when SNR or $\sigma^2$ is fixed, using the assumption $\lim_{N\to\infty}\left\{\frac{\mathbf{A}^T_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}}{N}\right\} = \mathbf{M}_{\mathcal{S}}$ we get

$$\lim_{N\to\infty}\mathbb{E}[\hat{\sigma}^2_0] = \sigma^2 + \mathbf{x}^T_{\mathcal{S}}\mathbf{M}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} \quad \& \quad \lim_{N\to\infty}\text{Var}[\hat{\sigma}^2_0] = 0, \quad (89)$$

where $\mathbf{M}_{\mathcal{S}}$ is a bounded positive definite matrix and as such $\mathbf{x}^T_{\mathcal{S}}\mathbf{M}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} = \mathcal{O}(1)$ as $N$ grows large.

## Appendix D
### Statistical analysis of $\hat{\sigma}^2_{\mathcal{I}}$ when $\mathcal{S} \subseteq \mathcal{I}$

The noise variance estimate under hypothesis $\mathcal{H}_{\mathcal{I}}$ can be rewritten as

$$\hat{\sigma}^2_{\mathcal{I}} = \left(\frac{\sigma^2}{N}\right)\frac{\mathbf{y}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}}\mathbf{y}}{\sigma^2}. \quad (90)$$

The true model $\mathbf{u} = \mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}$ lies in a linear subspace spanned by the columns of $\mathbf{A}_{\mathcal{S}}$. Consequently, for $\mathcal{I} \supseteq \mathcal{S}$ we have

$\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{u} = \mathbf{0}$. This implies that $\mathbf{y}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y} = \mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{e}$. Thus we have,

$$\frac{\mathbf{y}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}}{\sigma^2} = \frac{\mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{e}}{\sigma^2} \sim \chi^2_{N-k} \text{ (Using Lemma 1)}, \tag{91}$$

where $k = card(\mathcal{I}) \geq k_0$. This implies that $\left(\frac{N}{\sigma^2}\right)\hat{\sigma}^2_{\mathcal{I}} \sim \chi^2_{N-k}$. Therefore, the mean and variance of $\hat{\sigma}^2_{\mathcal{I}}$ for $\mathcal{I} \supseteq \mathcal{S}$ are:

$$\mathbb{E}[\hat{\sigma}^2_{\mathcal{I}}] = \frac{\sigma^2}{N}(N-k) \quad \& \quad \text{Var}[\hat{\sigma}^2_{\mathcal{I}}] = 2\frac{\sigma^4}{N^2}(N-k). \tag{92}$$

Hence, when $\sigma^2$ is a constant,

$$\lim_{N\to\infty} \mathbb{E}[\hat{\sigma}^2_{\mathcal{I}}] = \sigma^2 \quad \& \quad \lim_{N\to\infty} \text{Var}[\hat{\sigma}^2_{\mathcal{I}}] = 0. \tag{93}$$

## Appendix E

*Lemma 1:* Let $\mathbf{y}$ be a $N \times 1$ dimensional vector following $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_N)$ and $\mathbf{\Pi}$ be a $N \times N$ symmetric, idempotent matrix with rank$(\mathbf{\Pi}) = r$. Then the ratio $\mathbf{y}^T\mathbf{\Pi}\mathbf{y}/\sigma^2$ has a non-central chi-square distribution $\chi^2_r(\lambda)$ with $r$ degrees of freedom and non-centrality parameter $\lambda = \boldsymbol{\mu}^T\mathbf{\Pi}\boldsymbol{\mu}/\sigma^2$ (see, e.g., Chapter 5 of [40]).

*Lemma 2:* Let $Z_{\text{max}} = \max\limits_{i}\{Z_i\}_{i=1}^m$ where $Z_1, Z_2, \ldots, Z_m$ is a sequence of identically distributed random variables (not necessarily independent) having a Chi-square distribution with $k$ degrees of freedom where $k < m$. Then $Z_{\text{max}} \leq k + 2\sqrt{k\psi \ln m} + 2\psi \ln m$ for some constant $\psi > 1$ with probability approaching one as $m \to \infty$.

*Proof:* From the union bound we have

$$\Pr(Z_{\text{max}} \leq \eta) \geq 1 - m\Pr(Z_i \geq \eta). \tag{94}$$

Since $Z_i \sim \chi^2_k$, then from the Chi-square tail bound (Lemma 1 of [41]) we have the following result

$$\Pr\left(Z_i \geq k + 2\sqrt{kt} + 2t\right) \leq e^{-t}. \tag{95}$$

Setting $t = \psi \ln m$ in (95) where $\psi > 1$ we get

$$\Pr\left(Z_i \geq k + 2\sqrt{k\psi\ln m} + 2\psi\ln m\right) \leq e^{-\psi\ln m} = m^{-\psi}. \tag{96}$$

Using (96) in (94) we get

$$\Pr\left(Z_{\text{max}} \leq k + 2\sqrt{k\psi\ln m} + 2\psi\ln m\right) \geq 1 - \frac{1}{m^{\psi-1}}. \tag{97}$$

Therefore, $Z_{\text{max}} \leq k + 2\sqrt{k\psi\ln m} + 2\psi\ln m$ with probability approaching one as $m \to \infty$ if $\psi > 1$.

*Lemma 3:* Let $X_{\text{max}} = \max\limits_{i}\{X_i\}_{i=1}^m$ where $X_1, X_2, \ldots, X_m$ is a sequence of identically distributed random variables (not necessarily independent) having a Gaussian distribution with zero mean and variance one. Then $X_{\text{max}} \leq \sqrt{2\ln m}$ with probability approaching one as $m \to \infty$.

*Proof:* From the union bound we have

$$\Pr(X_{\text{max}} \leq \eta) \geq 1 - m\Pr(X_i \geq \eta). \tag{98}$$

Since $X_i \sim \mathcal{N}(0,1)$, from the Gaussian tail bound we have

$$\Pr(X_i \geq \eta) \leq \frac{1}{\eta}\frac{e^{-\eta^2/2}}{\sqrt{2\pi}}, \tag{99}$$

for all $\eta > 0$. Setting $\eta = \sqrt{2\ln m}$ in (99) we get

$$\Pr\left(X_i \geq \sqrt{2\ln m}\right) \leq \frac{m^{-1}}{2\sqrt{\pi\ln m}}. \tag{100}$$

Using (100) in (98) we get

$$\Pr\left(X_{\text{max}} \leq \sqrt{2\ln m}\right) \geq 1 - \frac{1}{2\sqrt{\pi\ln m}}. \tag{101}$$

Therefore, $X_{\text{max}} \leq \sqrt{2\ln m}$ with probability approaching one as $m \to \infty$.

*Lemma 4:* For any arbitrary support $\mathcal{I} \in \mathcal{I}_m^k \in \mathbb{M}$, under the asymptotic identifiability condition in (38) the following inequality holds

$$\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\right\|_2^2 > 0.$$

*Proof:* Let $\mathcal{S}' = \{\mathcal{S} \setminus \mathcal{I}\}$. The true support $\mathcal{S}$ can be split into two disjoint subsets as $\mathcal{S} = \{\mathcal{S} \cap \mathcal{I}\} \cup \{\mathcal{S} \setminus \mathcal{I}\}$. Since span$(\mathbf{A}_{\mathcal{S}\cap\mathcal{I}}) \subset$ span$(\mathbf{A}_{\mathcal{I}})$ we have

$$\begin{aligned}\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_2^2 =& \|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}'}\mathbf{x}_{\mathcal{S}'}\|_2^2 \\ =& N\mathbf{x}_{\mathcal{S}'}^T\left(N^{-1}\mathbf{A}_{\mathcal{S}'}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}'}\right)\mathbf{x}_{\mathcal{S}'}.\end{aligned}$$

Now, consider the matrix $\mathbf{M} = \begin{bmatrix}\mathbf{A}_{\mathcal{S}'} & \mathbf{A}_{\mathcal{I}}\end{bmatrix}$ where $card(\mathcal{S}') \leq K$ and $card(\mathcal{I}) \leq K$, such that $card(\mathcal{S}' \cup \mathcal{I}) \leq 2K$. Under the assumption (38)

$$N^{-1}\mathbf{M}^T\mathbf{M} = N^{-1}\begin{bmatrix}\mathbf{A}_{\mathcal{S}'}^T\mathbf{A}_{\mathcal{S}'} & \mathbf{A}_{\mathcal{S}'}^T\mathbf{A}_{\mathcal{I}} \\ \mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{S}'} & \mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\end{bmatrix} \tag{102}$$

is a bounded positive definite matrix. Then the Schur complement of the block matrix $\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}$ is

$$\begin{aligned}&N^{-1}\left[\mathbf{A}_{\mathcal{S}'}^T\mathbf{A}_{\mathcal{S}'} - \mathbf{A}_{\mathcal{S}'}^T\mathbf{A}_{\mathcal{I}}(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}})^{-1}\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{S}'}\right] \\ =&N^{-1}\mathbf{A}_{\mathcal{S}'}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}'}\end{aligned}$$

is also positive definite and bounded as $N \to \infty$. Let $\widetilde{\mathbf{M}} = N^{-1}\mathbf{A}_{\mathcal{S}'}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}'}$, then, $\mathbf{x}_{\mathcal{S}'}^T\widetilde{\mathbf{M}}\mathbf{x}_{\mathcal{S}'} = b$ (say) $=$ const. $> 0$. Hence, $\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_2^2 = Nb > 0$ for all $\mathcal{I} \in \mathcal{I}_m^k \in \mathbb{M}$.

## References

[1] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 6, pp. 16–34, 2018.

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

[3] C. Rao, Y. Wu, S. Konishi, and R. Mukerjee, "On model selection," *Lecture Notes-Monograph Series*, pp. 1–64, 2001.

[4] D. Anderson and K. Burnham, "Model selection and multi-model inference," *Second. NY: Springer-Verlag*, vol. 63, p. 10, 2004.

[5] A. Chakrabarti and J. K. Ghosh, "AIC, BIC and recent advances in model selection," *Philosophy of statistics*, pp. 583–605, 2011.

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

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

[8] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

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

[10] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[11] P. Stoica and P. Babu, "Model order estimation via penalizing adaptively the likelihood (PAL)," *Signal Processing*, vol. 93, no. 11, pp. 2865–2871, 2013.

[12] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, and S. H. Jensen, "Bayesian model comparison with the g-prior," *IEEE transactions on signal processing*, vol. 62, no. 1, pp. 225–238, 2013.

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

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

[15] S. Kay, "Exponentially embedded families-new approaches to model order estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 1, pp. 333–345, 2005.

[16] H. 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.

[17] R. R. Picard and R. D. Cook, "Cross-validation of regression models," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 575–583, 1984.

[18] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[19] M. Chichignoud, J. Lederer, and M. J. Wainwright, "A practical scheme and fast algorithm to tune the lasso with optimality guarantees," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8162–8181, 2016.

[20] L. de Torrenté and T. Hastie, "Does cross-validation work when $p \gg n$?" 2012.

[21] S. Kallummil and S. Kalyani, "Signal and noise statistics oblivious orthogonal matching pursuit," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2429–2438.

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

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

[24] R. F. Barber and E. J. Candès, "A knockoff filter for high-dimensional selective inference," *The Annals of Statistics*, vol. 47, no. 5, pp. 2504–2537, 2019.

[25] E. Candes, Y. Fan, L. Janson, and J. Lv, "Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 3, pp. 551–577, 2018.

[26] P. B. Gohain and M. Jansson, "New improved criterion for model selection in sparse high-dimensional linear regression models," in *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5692–5696.

[27] ——, "Scale-invariant and consistent Bayesian information criterion for order selection in linear regression models," *Signal Processing*, p. 108499, 2022.

[28] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice Hall PTR, 1993.

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

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

[31] D. F. Schmidt and E. 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.

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

[33] Q. Ding and S. Kay, "Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 1959–1969, 2011.

[34] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[35] C.-H. Zhang and J. Huang, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *The Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[36] R. 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.

[37] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[38] G. C. McDonald and R. C. Schwing, "Instabilities of regression estimates relating air pollution to mortality," *Technometrics*, vol. 15, no. 3, pp. 463–481, 1973. [Online]. Available: https://myweb.uiowa.edu/pbreheny/data/pollution.html

[39] N. C. Institute. The cancer genome atlas program. [Online]. Available: https://myweb.uiowa.edu/pbreheny/data/bcTCGA.html

[40] A. M. Mathai and S. B. Provost, *Quadratic forms in random variables: theory and applications*. Dekker, 1992.

[41] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000.