# A Model Selection Criterion For High-Dimensional Linear Regression 

Arash Owrang, Student Member, IEEE, and Magnus Jansson, Senior Member, IEEE


#### Abstract

Statistical model selection is a great challenge when the number of accessible measurements is much smaller than the dimension of the parameter space. We study the problem of model selection in the context of subset selection for highdimensional linear regressions. Accordingly, we propose a new model selection criterion with the Fisher information that leads to the selection of a parsimonious model from all the combinatorial models up to some maximum level of sparsity. We analyze the performance of our criterion as the number of measurements grows to infinity, as well as when the noise variance tends to zero. In each case, we prove that our proposed criterion gives the true model with a probability approaching one. Additionally, we devise a computationally affordable algorithm to conduct model selection with the proposed criterion in practice. Interestingly, as a side product, our algorithm can provide the ideal regularization parameter for the Lasso estimator such that Lasso selects the true variables. Finally, numerical simulations are included to support our theoretical findings.


Index Terms-Model selection, high-dimensional inference, subset selection, Bayesian information criterion, Lasso, sparse estimation, regularization

## I. Introduction

STATISTICAL model selection is the task of selecting a parsimonious model from a set of available competing models in the $N$-dimensional parameter space that best fits the measurements. This task is of such fundamental importance that many problems in statistical inference can be considered to be problems linked to model selection [1]. For this reason, many model selection methods have been developed in the classical setting in which the number of measurements, $m$, can grow to infinity, but $N$ and the number of available competing models is fixed [2]-[8]. The Bayesian information criterion (BIC) is one of these classical methods that has been very successful due to its simplicity and effectiveness [3]. In fact, when $N$ is fixed, it is shown that BIC is asymptotically consistent in selecting the true model as $m$ grows to infinity [5], [9]. However, in high-dimensional settings, in which the number of accessible measurements is much smaller than the dimension of the parameter space $(N \gg m)$, BIC and other classical model selection methods are ineffective and prone to overfitting [10], [11].

To overcome this issue, a novel extension to BIC (EBIC) is proposed in [11]. EBIC is a generic model selection method

[^0]in the sense that it generalizes the set of competing models to the collection of all plausible combinatorial models up to cardinality $K$, with $K \ll m$. It is shown that under a suitable asymptotic identifiability condition, EBIC can consistently select the true model as $m$ grows to infinity [11]. However, as indicated in our simulations the empirical performance of EBIC can sometimes be unsatisfactory for practical sizes of $m$. Moreover, in scenarios when $m$ is fixed but the noise variance, $\sigma^{2}$, tends to zero, our results show that EBIC is inconsistent.

In this paper, we consider the problem of generic model selection for high-dimensional data. Accordingly, inspired by EBIC and the model selection criteria with Fisher information [6], [12], we propose a model selection criterion named as extended Fisher information criteria (EFIC). Some preliminary results of this work have been previously published in [13]. Here, we analyze the performance of EFIC for the linear regression problem as $m \rightarrow \infty$, as well as when the noise variance tends to zero. For both cases, we prove that the EFIC's selected model coincides with the true subset with a probability approaching one. Our theoretical findings are also confirmed by numerical simulations. More specifically, the numerical simulation illustrates the superiority of EFIC to EBIC for practical sizes of $m$ or when $\sigma^{2} \rightarrow 0$. Additionally, to practically perform model selection with EFIC, we devise a computationally affordable algorithm that is assisted by the modified-LARS algorithm [14] providing the Lasso path. As a side effect of our algorithm, one can find the ideal regularization parameter in the Lasso estimator, in the sense that Lasso provides the variables of interest.

To assist the reader, we list some of the most frequently addressed notations in the following. Given sequences $f(m)$ and $h(m)$, as $m \rightarrow \infty$, the notation $f(m)=\mathrm{o}(h(m))$ means that $|f(m) / h(m)| \rightarrow 0, f(m)=\mathrm{O}(h(m))$ means that there exists a constant $C_{1}$ such that $|f(m)| \leq C_{1}|h(m)|$ and $f(m)=\Omega(h(m))$ means that there exists a constant $C_{2}>0$ such that $|f(m)| \geq C_{2}|h(m)|[15]$.

| $\mathcal{I}$ | $\left\{i_{1}, i_{2}, \ldots, i_{k}\right\}$ |
| :---: | :---: |
| $\mathcal{S}$ | the set corresponding to the true model |
| $\mathbf{x}_{\mathcal{I}}$ | $\left[x_{i_{1}}, x_{i_{2}}, \ldots, x_{i_{k}}\right]^{T}$ |
| $\mathcal{J}$ | $\bigcup_{k=0}^{K}\{\mathcal{I}\| \| \mathcal{I} \mid=k\}$ |
| $\dot{\mathcal{I}}^{k}$ | $\{\mathcal{I}\|\|\mathcal{I}\|=k, \mathcal{S} \subset \mathcal{I}\}$ |
| $\check{\mathcal{I}}^{k}$ | $\{\mathcal{I}\|\|\mathcal{I}\|=k, \mathcal{S} \not \subset \mathcal{I}\}$ |

## II. Generalized Model Selection

Let $\mathbf{y} \in \mathbb{R}^{m}$ be the measurement vector and suppose that under hypothesis $\mathscr{H}_{\mathcal{I}}$ we have

$$
\mathscr{H}_{\mathcal{I}}: \mathbf{y}=s\left(\mathbf{x}_{\mathcal{I}}\right)+\sigma \boldsymbol{\epsilon}, \quad \mathcal{I} \in \mathcal{J}
$$

Here, the set $\mathcal{I}=\left\{i_{1}, i_{2}, \ldots, i_{k}\right\}$,

$$
1 \leq i_{1}<i_{2}<\cdots<i_{k} \leq N
$$

represents the data model with the convention of $\mathcal{I}=\emptyset$ for $k=0$. The unknown vector $\mathbf{x}_{\mathcal{I}} \in \mathbb{R}^{k}$ is the collection of $x_{i}$ 's with the support $\mathcal{I}$ from the general parameter vector $\mathbf{x} \in \mathbb{R}^{N}, \mathbf{x}_{\mathcal{I}}=\left[\begin{array}{llll}x_{i_{1}}, & x_{i_{2}}, & \ldots, & x_{i_{k}}\end{array}\right]^{T}$. The set $\mathcal{J}=\bigcup_{k=0}^{K}\{\mathcal{I}| | \mathcal{I} \mid=k\}$ is the collection of all the combinatorial competing models up to cardinality $K \ll m$ and $|\mathcal{I}|$ denotes the cardinality of the set $\mathcal{I}$. It is assumed that the continuous function $s(\cdot)$ is known and maps $\mathbb{R}^{N}$ to $\mathbb{R}^{m}$, and $N$ is linked to $m$ by $N=m^{d}$ where $d>1$ is a constant. Moreover, the elements of the noise vector $\epsilon$ are independent and identically Gaussian distributed with $\epsilon_{i} \sim \mathcal{N}(0,1)$, and the unknown standard deviation $\sigma \geq 0$ is considered as a nuisance parameter. Subsequently, under each equi-probable $\mathscr{H}_{\mathcal{I}}$, the probability density function $p\left(\mathbf{y} ; \underline{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)$ is known apart from $\underline{\boldsymbol{\theta}}_{\mathcal{I}} \triangleq\left[\sigma^{2}, \mathbf{x}_{\mathcal{I}}\right]^{T}$. Our main interest is to propose a criterion to identify the true unknown subset $\mathcal{S}$ from the collection of competing subsets $\mathcal{J}$, given that the measurement $\mathbf{y} \in \mathbb{R}^{m}$ is generated by hypothesis $\mathscr{H}_{\mathcal{S}}$. The estimate of this criterion, $\hat{\mathcal{I}}$, should fulfill

$$
\begin{array}{ll}
\operatorname{Pr}\{\hat{\mathcal{I}}=\mathcal{S}\} \rightarrow 1 & \text { as } \sigma \rightarrow 0 \\
\operatorname{Pr}\{\hat{\mathcal{I}}=\mathcal{S}\} \rightarrow 1 & \text { as } m \rightarrow \infty
\end{array}
$$

A standard approach to find $\hat{\mathcal{I}}$ is to minimize a penalized log-likelihood [2]-[7]. To adapt this approach to the needs of generalized model selection, [11] suggests to add a binomial coefficient penalty to the BIC's objective function. As a result, EBIC selects $\hat{\mathcal{I}}$ by

$$
\begin{equation*}
\min _{\mathcal{I} \in \mathcal{J}}\left\{-2 \ln p\left(\mathbf{y} ; \hat{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)+(|\mathcal{I}|+1) \ln m+2 \dot{c} \ln \binom{N}{|\mathcal{I}|}\right\}, \tag{1}
\end{equation*}
$$

where $\hat{\boldsymbol{\theta}}_{\mathcal{I}}$ is the maximum likelihood (ML) estimate of $\underline{\boldsymbol{\theta}}_{\mathcal{I}}$, $\binom{N}{|\mathcal{I}|}$ is a binomial coefficient, and $\dot{c}>1-1 /(2 d)$ is a constant that controls the penalty level. The advent of the binomial coefficient penalty term is the consequence of considering the binomial model spaces. By introducing this penalty, EBIC penalizes the log-likelihood rigorously to compensate for the effect of a large collection of binomial spaces. It is proven that the EBIC's estimate satisfies $\operatorname{Pr}\{\hat{\mathcal{I}}=\mathcal{S}\} \rightarrow 1$ as $m \rightarrow \infty$ [11]. However, its empirical performance for problems of practical size is not satisfactory. See e.g. Fig. 3 in Sec. V. Moreover, EBIC exploits BIC's criterion, in which some approximations have been considered. Namely, approximating the determinant of the Fisher information matrix by $m^{|\mathcal{I}|+1}$. This is acceptable for many cases but not always appropriate, e.g. for polynomial regressions and in high signal-to-noise ratio scenarios [16]. Due to such an approximation, EBIC is inconsistent in selecting the true model as $\sigma \rightarrow 0$. See Fig. 4 in Sec. V.

To fulfill our goal in estimating the model, we propose a new model selection criterion for high-dimensional data, EFIC, as

$$
\begin{equation*}
\min _{\mathcal{I} \in \mathcal{J}}\left\{-2 \ln p\left(\mathbf{y} ; \hat{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)+\ln \operatorname{det} \mathbf{F}\left(\underline{\hat{\boldsymbol{\theta}}}_{\mathcal{I}}\right)+2 c d|\mathcal{I}| \ln m\right\} . \tag{2}
\end{equation*}
$$

Here, $\mathbf{F}(\cdot)$ is the Fisher information matrix defined by

$$
\mathbf{F}\left(\underline{\boldsymbol{\theta}}_{\mathcal{I}}\right) \triangleq-\mathbb{E}\left(\begin{array}{ll}
\frac{\partial^{2} \ln p\left(\mathbf{y} ; \boldsymbol{\theta}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)}{\partial \mathbf{x}_{\mathcal{I}} \partial \mathbf{x}_{\mathcal{I}}^{T}} & \frac{\partial^{2} \ln p\left(\mathbf{y} ; \boldsymbol{\theta}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)}{\partial \mathbf{x}_{\mathcal{I}} \partial \sigma^{2}}  \tag{3}\\
\frac{\partial^{2} \ln p\left(\mathbf{y} ; \boldsymbol{\theta}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)}{\partial \sigma^{2} \partial \mathbf{x}_{\mathcal{I}}^{T}} & \frac{\partial^{2} \ln p\left(\mathbf{y} ; \boldsymbol{\theta}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)}{\partial \sigma^{4}}
\end{array}\right)
$$

where $\mathbb{E}$ denotes the expected value operator and the expectation is taken with respect to $p\left(\mathbf{y} ; \underline{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)$. The choice of $d|\mathcal{I}| \ln m$ in (9) is due to approximating $\ln \binom{N}{|\mathcal{I}|}$. To show this, note that

$$
\ln \binom{N}{|\mathcal{I}|}=\sum_{i=0}^{|\mathcal{I}|-1} \ln (N-i)-\ln (|\mathcal{I}|!)
$$

Now, since $|\mathcal{I}| \leq K \ll m$ and $N=m^{d}$, we have

$$
\ln \binom{N}{|\mathcal{I}|}=|\mathcal{I}| \ln N(1+\mathrm{o}(1))=d|\mathcal{I}| \ln m(1+\mathrm{o}(1))
$$

as $m \rightarrow \infty$. The choice of the constant $c$ is discussed in Sec. III. In the following, we will make a detailed analysis of EFIC for the linear regression problem.

## A. EFIC for Linear Regression

Now consider $s\left(\mathbf{x}_{\mathcal{I}}\right)$ as a linear function of $\mathbf{x}_{\mathcal{I}}$. Thus,

$$
\begin{equation*}
\mathscr{H}_{\mathcal{I}}: \mathbf{y}=\mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}}+\sigma \boldsymbol{\epsilon}, \quad \mathcal{I} \in \mathcal{J} \tag{4}
\end{equation*}
$$

where the matrix $\mathbf{A}_{\mathcal{I}}$ is the collection of the columns of the known regressor or sensing matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ with the support $\mathcal{I}$. Then, for the preceding linear regression, the loglikelihood becomes

$$
\begin{equation*}
\ln p\left(\mathbf{y} ; \underline{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)=-\frac{m}{2} \ln \left(2 \pi \sigma^{2}\right)-\frac{1}{2 \sigma^{2}}\left\|\mathbf{y}-\mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}}\right\|_{2}^{2} \tag{5}
\end{equation*}
$$

where $\|\cdot\|_{2}$ represents the Euclidean norm. The ML estimate of $\underline{\theta}_{\mathcal{I}}$ is

$$
\begin{equation*}
\hat{\boldsymbol{\theta}}_{\mathcal{I}} \triangleq\left[\hat{\sigma^{2}}, \hat{\mathbf{x}}_{\mathcal{I}}\right]^{T}=\left[\frac{1}{m}\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}, \mathbf{A}_{\mathcal{I}}^{\dagger} \mathbf{y}\right]^{T} \tag{6}
\end{equation*}
$$

where the matrix $\mathbf{A}_{\mathcal{I}}^{\dagger}=\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right)^{-1} \mathbf{A}_{\mathcal{I}}^{T}$ is the Moore-Penrose pseudo-inverse of $\mathbf{A}_{\mathcal{I}}$ and the matrix $\boldsymbol{\Pi}_{\mathcal{I}}^{\perp}=\mathbf{I}-\mathbf{A}_{\mathcal{I}} \mathbf{A}_{\mathcal{I}}^{\dagger}$ is the orthogonal projection matrix onto the null-space of $\mathbf{A}_{\mathcal{I}}^{T}$. Next, take the second derivative of $\ln p\left(\mathbf{y} ; \underline{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)$ with respect to the corresponding unknown parameters,

$$
\begin{aligned}
& \frac{\partial^{2} \ln p\left(\mathbf{y} ; \underline{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)}{\partial \mathbf{x}_{\mathcal{I}} \partial \mathbf{x}_{\mathcal{I}}^{T}}=-\frac{1}{\sigma^{2}} \mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}} \\
& \frac{\partial^{2} \ln p\left(\mathbf{y} ; \underline{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)}{\partial \mathbf{x}_{\mathcal{I}} \partial \sigma^{2}}=-\frac{1}{\sigma^{4}} \mathbf{A}_{\mathcal{I}}^{T}\left(\mathbf{y}-\mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}}\right) \\
& \frac{\partial^{2} \ln p\left(\mathbf{y} ; \boldsymbol{\theta}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)}{\partial \sigma^{4}}=\frac{m}{2 \sigma^{4}}-\frac{1}{\sigma^{6}}\left\|\mathbf{y}-\mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}}\right\|_{2}^{2}
\end{aligned}
$$

By taking the expectation of the preceding identities with respect to $p\left(\mathbf{y} ; \underline{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)$, we obtain the Fisher information matrix of $\underline{\boldsymbol{\theta}}_{\mathcal{I}}$ as

$$
\mathbf{F}\left(\underline{\boldsymbol{\theta}}_{\mathcal{I}}\right)=\frac{1}{\sigma^{2}}\left(\begin{array}{cc}
\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}} & \mathbf{0} \\
\mathbf{0} & \frac{m}{2 \sigma^{2}}
\end{array}\right)
$$

Next use the properties of the determinant function to get

$$
\operatorname{det} \mathbf{F}\left(\underline{\boldsymbol{\theta}}_{\mathcal{I}}\right)=\frac{m}{2}\left(\frac{1}{\sigma^{2}}\right)^{|\mathcal{I}|+2} \operatorname{det}\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right)
$$

Inserting $\hat{\sigma^{2}}$ from (6) into the preceding identity and taking logarithm from both sides leads to

$$
\begin{align*}
\ln \operatorname{det} \mathbf{F}\left(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{I}}\right)= & \ln (m / 2)-(|\mathcal{I}|+2)\left[\ln \left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}-\ln m\right] \\
& +\ln \operatorname{det}\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right) \tag{7}
\end{align*}
$$

From (6) and (5),

$$
\begin{equation*}
-2 \ln p\left(\mathbf{y} ; \hat{\boldsymbol{\theta}}_{\mathcal{I}} \mid \mathscr{H}_{\mathcal{I}}\right)=m \ln \left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}+m\left(1+\ln \frac{2 \pi}{m}\right) \tag{8}
\end{equation*}
$$

Eventually, by inserting (7) and (8) into (2) and ignoring the constant terms we get

$$
\begin{equation*}
\hat{\mathcal{I}}=\underset{\mathcal{I} \in \mathcal{J}}{\arg \min } g(\mathcal{I}) \tag{9}
\end{equation*}
$$

where

$$
\begin{align*}
g(\mathcal{I})= & \overbrace{(m-|\mathcal{I}|-2) \ln \left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}}^{g_{1}(\mathcal{I})} \\
& +\underbrace{\ln \operatorname{det}\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right)+(1+2 c d)|\mathcal{I}| \ln m}_{g_{2}(\mathcal{I})} . \tag{10}
\end{align*}
$$

To gain intuition for the behavior of $g(\mathcal{I})$, it is helpful to break it into two terms as $g_{1}(\cdot)$ and $g_{2}(\cdot)$, and consider the behavior of each term with respect to $\mathcal{I}$ 's that deviate from the true model. In this regard, first split $\mathcal{J} \backslash \mathcal{S}$ into two classes of overfitted and misfitted models as

$$
\mathcal{J} \backslash \mathcal{S}=\left\{\bigcup_{k=|\mathcal{S}|+1}^{K} \dot{\mathcal{I}}^{k}\right\} \bigcup\left\{\bigcup_{k=0}^{K} \breve{\mathcal{I}}^{k}\right\}
$$

where

$$
\begin{aligned}
& \stackrel{\circ}{\mathcal{I}}^{k} \triangleq\{\mathcal{I}| | \mathcal{I} \mid=k, \mathcal{S} \subset \mathcal{I}\} \\
& \breve{\mathcal{I}}^{k} \triangleq\{\mathcal{I}| | \mathcal{I} \mid=k, \mathcal{S} \not \subset \mathcal{I}\}
\end{aligned}
$$

Next consider the term $g_{1}(\cdot)$. Observe that $g_{1}(\cdot)$, which is a function of the estimate of $\sigma^{2}$, inflates as $\mathcal{I}$ deviates from $\mathcal{S}$ such that $\mathcal{I} \in \breve{\mathcal{I}}^{k}$, since

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

for $\mathcal{I} \in \breve{\mathcal{I}}^{k}$. However, in contrast to the misfitted models, for overfitted models of the form

$$
\begin{equation*}
\mathcal{S} \subset \mathcal{I}_{\mathcal{S} \cup i_{k+1}} \subset \mathcal{I}_{\mathcal{S} \cup\left\{i_{k+1}, i_{k+2}\right\}} \subset \cdots \subset \mathcal{I}_{\mathcal{S} \cup\left\{i_{k+1}, \ldots, i_{K}\right\}} \tag{11}
\end{equation*}
$$

$g_{1}(\cdot)$ monotonically decreases as

$$
\begin{aligned}
\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \mathbf{y}\right\|_{2}^{2}>\left\|\boldsymbol{\Pi}_{\mathcal{S} \cup i_{k+1}}^{\perp} \mathbf{y}\right\|_{2}^{2} & >\left\|\boldsymbol{\Pi}_{\mathcal{S} \cup\left\{i_{k+1}, i_{k+2}\right\}}^{\perp} \mathbf{y}\right\|_{2}^{2} \\
>\cdots> & >\boldsymbol{\Pi}_{\mathcal{S} \cup\left\{i_{k+1}, \ldots, i_{K}\right\}}^{\perp} \mathbf{y} \|_{2}^{2}
\end{aligned}
$$



Fig. 1. Projecting $y$ into the orthogonal subspaces of the nested models of $\mathcal{S}$ and $\mathcal{I}_{\left\{\mathcal{S} \cup i_{1}\right\}}$ makes $g_{1}(\mathcal{S}) \geq g_{1}\left(\mathcal{I}_{1}\right)$.
for $i_{k+l} \notin \mathcal{S}$ with $l \geq 1$. See Fig. 1 for a geometrical illustration. By this argument we can see that $g_{1}(\cdot)$ measures the goodness of the fit and gets smaller as the complexity of the model increases in $\mathcal{I} \in \dot{\mathcal{I}}^{k}$. Unlike $g_{1}(\cdot), g_{2}(\cdot)$ measures the complexity of the model and, generally speaking, it increases as the cardinality of the model increases. Intuitively, by this increase, $g_{2}(\cdot)$ attempts to counterbalance the decrease in $g_{1}(\cdot)$ and therefore preventing (9) from overfitting.

## B. Identifiability of The Model

A model is considered to be identifiable if no other model of the same or smaller size can describe the given (noise free) measurements equally well. In the linear regression setup, this is equivalent to say $\mathbf{y}=\tilde{\mathbf{A}}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}} \neq \tilde{\mathbf{A}}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}}$ for $\{\mathcal{I}||\mathcal{I}| \leq|\mathcal{S}|, \mathcal{I} \neq \mathcal{S}\}$. Here, $\tilde{\mathbf{A}}$ denotes the normalized version of the sensing matrix $\mathbf{A}$ in the sense that $\tilde{\mathbf{A}}$ has unit-norm columns. The identifiability of the true model in the highdimensional linear regression setup is uniformly maintained if the minimal eigenvalue of all restricted sub-matrices, $\tilde{\mathbf{A}}_{\mathcal{I}}^{T} \tilde{\mathbf{A}}_{\mathcal{I}}$ for $\{\mathcal{I}||\mathcal{I}| \leq 2 K\}$, is bounded away from zero [17]-[19]. This is a reasonable constraint; for instance, when $\tilde{\mathbf{A}}$ is a sub-Gaussian random matrix, one can roughly conclude from Chapter 9 in [19] that the minimal restricted eigenvalue of $\tilde{\mathbf{A}}$ is bounded away from zero, even if $N$ grows exponentially with $m$.

Having this stated, we need to lower bound the minimal eigenvalues of all restricted sub-matrices of at most size $2 K$. Now, inspired by [20], we introduce our slightly more general assumption in comparison with what is traditionally assumed, such as e.g. the restricted isometry property in [21], as we allow the minimal restricted eigenvalues to slowly converge to zero.
Restricted eigenvalue property: The normalized matrix $\tilde{\mathbf{A}}$ satisfies the restricted eigenvalue property if any restricted sub-
matrix $\tilde{\mathbf{A}}_{\mathcal{I}}^{T} \tilde{\mathbf{A}}_{\mathcal{I}}$ obeys

$$
\min _{|\mathcal{I}| \leq 2 K} \Lambda_{\min }\left(\tilde{\mathbf{A}}_{\mathcal{I}}^{T} \tilde{\mathbf{A}}_{\mathcal{I}}\right) \geq \frac{C_{\min }}{\ln m}
$$

for some constant $C_{\min }>0$. Here, $\Lambda_{\min }(\cdot)$ denotes the minimum eigenvalue of the corresponding matrix. Clearly, for a fixed $m$, the value of $C_{\text {min }}$ indicates the degree of correlation between $\tilde{\mathbf{a}}_{i}$ 's.

## III. Performance Analysis of EFIC for the Linear REGRESSION

Desirably, a statistical inference method provides the true unknown parameter in an asymptotic regime. To examine the asymptotic properties of EFIC, we evaluate the probability of (9) selecting the correct model as $\sigma \rightarrow 0$, as well as when $m \rightarrow \infty$.

## A. Deterministic Optimality Conditions for EFIC

Prior to investigating the asymptotic properties of EFIC, we first present the sufficient non-asymptotic conditions, under which EFIC selects the true model. Later, in Theorem 1 and 2, we show that these conditions are satisfied with high probability in the corresponding asymptotic regimes.
Lemma 1. Let $\mathcal{S} \in \mathcal{J}$ be the true model and suppose that $\mathcal{I} \neq \mathcal{S}$ denotes any other competing subset in $\mathcal{J}$. Additionally, assume that the matrix $\mathbf{A}$ satisfies the restricted eigenvalue property. Then, for a particular realization of $\epsilon$, the minimizer of (9) obeys $\hat{\mathcal{I}}=\mathcal{S}$, if

$$
\begin{array}{ll}
\mathscr{I}_{1}: \ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2(m-|\mathcal{S}|-2)}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2(m-|\mathcal{I}|-2)}}<\Upsilon(\mathcal{I})-\ln \sigma^{2 \Delta}, & \mathcal{I} \in \bigcup_{k=|\mathcal{S}|+1}^{K} \dot{\mathcal{I}}^{k} \\
\mathscr{I}_{2}: \ln \frac{\left(\sigma^{2}\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}\right)^{(m-|\mathcal{S}|-2)}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2(m-|\mathcal{I}|-2)}}<\Upsilon(\mathcal{I}), & \mathcal{I} \in \bigcup_{k=0}^{K} \breve{\mathcal{I}}^{k}
\end{array}
$$

where $\Delta=||\mathcal{I}|-|\mathcal{S}||$ and

$$
\Upsilon(\mathcal{I})=\ln \frac{\operatorname{det}\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right)}{\operatorname{det}\left(\mathbf{A}_{\mathcal{S}}^{T} \mathbf{A}_{\mathcal{S}}\right)}+(|\mathcal{I}|-|\mathcal{S}|)(1+2 c d) \ln m
$$

Proof. The strict minimizer of (9) coincides with the subset $\mathcal{S}$ if $g(\mathcal{S})<g(\mathcal{I})$ for any $\mathcal{I} \in \mathcal{J} \backslash \mathcal{S}$. Thus, using (10), it is sufficient to show that

$$
\begin{align*}
& (m-|\mathcal{S}|-2) \ln \left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \mathbf{y}\right\|_{2}^{2}+\ln \operatorname{det}\left(\mathbf{A}_{\mathcal{S}}^{T} \mathbf{A}_{\mathcal{S}}\right) \\
& +(1+2 c d)|\mathcal{S}| \ln m<(m-|\mathcal{I}|-2) \ln \left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2} \\
& +\ln \operatorname{det}\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right)+(1+2 c d)|\mathcal{I}| \ln m \tag{12}
\end{align*}
$$

is equivalent to the $\mathscr{I}_{1}$ and $\mathscr{I}_{2}$ inequalities. In this regard, by the definition of the projection matrix, we have $\boldsymbol{\Pi}_{\mathcal{I}} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}=$ $\mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}$ for $\mathcal{I} \in \dot{\mathcal{I}}^{k}$. Thus,

$$
\begin{equation*}
\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}=\boldsymbol{\Pi}_{\mathcal{I}}^{\perp}\left(\mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}+\sigma \boldsymbol{\epsilon}\right)=\sigma \boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}, \quad \forall \mathcal{I} \in \dot{\mathcal{I}}^{k} \tag{13}
\end{equation*}
$$

Moreover,

$$
\begin{equation*}
\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \mathbf{y}=\sigma \boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon} \tag{14}
\end{equation*}
$$

If (13) and (14) are inserted into (12), then, with some straightforward manipulations, we obtain $\mathscr{I}_{1}$. Next, use (14) and (12) to obtain $\mathscr{I}_{2}$. Also note that, because of the restricted eigenvalue property, $\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}$ and $\mathbf{A}_{\mathcal{S}}^{T} \mathbf{A}_{\mathcal{S}}$ are full-rank and therefore $\ln \operatorname{det}(\cdot)$ is finite.

## B. Model Selection as $\sigma \rightarrow 0$

As mentioned before, it is of course desirable that EFIC selects the true model for high signal to noise ratio data. This motivates us to examine the performance of EFIC as $\sigma \rightarrow 0$.

Theorem 1. Let $m$ be the fixed number of measurements and assume that $N=m^{d}$. Then, under the restricted eigenvalue property, the estimate of (9), $\hat{\mathcal{I}}$, obeys $\hat{\mathcal{I}}=\mathcal{S}$ with a probability approaching one as $\sigma \rightarrow 0$.

Proof. Based on Lemma 1, the minimizer of (9) coincides with $\mathcal{S}$ if $\mathscr{I}_{1}$ and $\mathscr{I}_{2}$ are satisfied. Having this stated, we begin by showing that as $\sigma \rightarrow 0$ the inequality $\mathscr{I}_{1}$ holds with a probability approaching one. In this regard, consider the event of the form

$$
\AA_{\mathcal{I}}: \rho(\mathcal{I})<e^{\Upsilon(\mathcal{I})} / \sigma^{2 \Delta}, \quad \mathcal{I} \in \grave{\mathcal{I}}^{k}
$$

for $|\mathcal{S}|+1 \leq k \leq K$, where $\rho(\mathcal{I})$ is defined by

$$
\rho(\mathcal{I})=\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2(m-|\mathcal{S}|-2)} /\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2(m-|\mathcal{I}|-2)}
$$

Observe that $\left\|\Pi_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}$ is a Chi-square random variable with $m-|\mathcal{I}|$ degrees of freedom. Therefore, $\operatorname{Pr}\left\{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}=\mathbf{0}\right\}=$ 0 . This implies that $\rho(\mathcal{I})$ is a well defined random variable [22], and therefore we can consider the probability of $\AA_{\mathcal{I}}$ happening. Now we look at the term $e^{\Upsilon(\mathcal{I})} / \sigma^{2 \Delta}$. It is clear that, due to the restricted eigenvalue property, there exists some constant $\alpha \in \mathbb{R}$ for which $\ln \left[\operatorname{det}\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right) / \operatorname{det}\left(\mathbf{A}_{\mathcal{S}}^{T} \mathbf{A}_{\mathcal{S}}\right)\right]>\alpha$, and therefore $e^{\Upsilon(\mathcal{I})} / \sigma^{2 \Delta} \rightarrow \infty$ as $\sigma \rightarrow 0$. Hence, we have $\operatorname{Pr}\left\{\AA_{\mathcal{I}}^{c}\right\} \rightarrow 0$ as $\sigma \rightarrow 0$, where $\AA_{\mathcal{I}}^{c}$ denotes the complement event of $\AA_{\mathcal{I}}$. Finally, by using Boole's inequality, we have

$$
\operatorname{Pr}\left\{\mathscr{I}_{1}\right\}=\operatorname{Pr}\left\{\bigcap_{\substack{k==\\|\mathcal{S}|+1}}^{K} \bigcap_{\mathcal{I} \in \dot{\mathcal{I}}^{k}} \AA_{\mathcal{I}}\right\} \geq 1-\sum_{\substack{k==\\|\mathcal{S}|+1}}^{K} \sum_{\mathcal{I} \in \mathcal{I}^{k}} \AA_{\mathcal{I}}^{c} \rightarrow 1
$$

as $\sigma \rightarrow 0$.
Continuing with $\mathscr{I}_{2}$, let the event $\breve{A}_{\mathcal{I}}$ be

$$
\breve{A}_{\mathcal{I}}:\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}<\beta(\mathcal{I}) / \sigma^{2}, \quad \mathcal{I} \in \breve{\mathcal{I}}^{k}
$$

where $\beta(\mathcal{I})$ is defined by

$$
\beta(\mathcal{I})=\left[e^{\Upsilon(\mathcal{I})}\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2(m-|\mathcal{I}|-2)}\right]^{\frac{1}{m-|\mathcal{S}|-2}}
$$

Now observe that $\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2} \rightarrow\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}^{2}$ in probability as $\sigma \rightarrow 0$. Moreover, we know from Lemma (4)-(i) in the Appendix that $\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}^{2}>0$. Because of this and the fact that $e^{\Upsilon(\mathcal{I})}>0$ we can conclude that $\beta(\mathcal{I})>\alpha^{\prime}$ as $\sigma \rightarrow 0$, for some positive constant $\alpha^{\prime}$. As a result, $\beta(\mathcal{I}) / \sigma^{2} \rightarrow \infty$ when $\sigma \rightarrow 0$ for any $\mathcal{I} \in \breve{\mathcal{I}}^{k}$. Therefore, the probability of the event $\breve{A}_{\mathcal{I}}$ happening converges to one as $\sigma \rightarrow 0$. Again, we use Boole's inequality to show that $\operatorname{Pr}\left\{\mathscr{I}_{2}\right\} \rightarrow 1$ as $\sigma \rightarrow 0$.

Theorem 1 implies that, when the noise power is negligible, $\mathcal{S}$ is the global minimizer of $g(\cdot)$ if the restricted eigenvalue property is satisfied. Now, to show the inconsistency of EBIC as $\sigma \rightarrow 0$, let us to restate the equivalent of $\mathscr{I}_{1}$ for EBIC. In this regard, first insert (8) into (1) and ignore the constant terms to get

$$
g_{\text {EBIC }}(\mathcal{I})=m \ln \left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}+(1+2 c ́ d)|\mathcal{I}| \ln m
$$

Next, by imitating the argument in Lemma 1, we can say

$$
\ln \left(\frac{\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}\right)^{m}<(|\mathcal{I}|-|\mathcal{S}|)(1+2 \dot{c} d) \ln m, \quad \mathcal{I} \in \bigcup_{k=|\mathcal{S}|+1}^{K} \dot{\mathcal{I}}^{k}
$$

which is the equivalent form of $\mathscr{I}_{1}$ for EBIC. Clearly, for a fixed $m$, EBIC cannot assure $\operatorname{Pr}\{\hat{\mathcal{I}}=\mathcal{S}\} \rightarrow 1$ as $\sigma \rightarrow 0$.

## C. Model Selection as $m \rightarrow \infty$

In the following, we state our second main result which presents the asymptotic behavior of EFIC as $m \rightarrow \infty$.

Theorem 2. Suppose that the matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$, with $N=m^{d}$, satisfies the restricted eigenvalue property. Moreover, assume that the columns of $\mathbf{A}$ fulfill

$$
\begin{equation*}
\left\|\mathbf{a}_{i}\right\|_{2}^{2}=\Omega\left(m^{a}\right) \tag{15}
\end{equation*}
$$

for some constant $a>0$. Then, the EFIC's estimate obeys $\hat{\mathcal{I}}=\mathcal{S}$ with probability one as $m \rightarrow \infty$, if $c$ is chosen such that

$$
c>1-\frac{a}{2 d}+\frac{1}{d}
$$

Proof. The proof of Theorem 2 consists of three main parts. We start by finding the asymptotic behavior of $\Upsilon$. Then, we establish the probability of $\mathscr{I}_{1}$ happening as $m \rightarrow \infty$ and, finally, we do the same for $\mathscr{I}_{2}$.

The asymptotic behavior of $\Upsilon(\mathcal{I})$ : Normalize the columns of $\mathbf{A}_{\mathcal{I}}$ as $\tilde{\mathbf{A}}_{\mathcal{I}}=\mathbf{A}_{\mathcal{I}} \mathbf{W}_{\mathcal{I}, \mathcal{I}}^{-1}$, where $\mathbf{W}_{\mathcal{I}, \mathcal{I}}$ is a diagonal matrix with the diagonal elements defined by $w_{i i}=\left\|\mathbf{a}_{i}\right\|_{2}$ for $i \in \mathcal{I}$. Now using the properties of the determinant function, we have

$$
\begin{aligned}
\ln \operatorname{det}\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right) & =\ln \operatorname{det}\left(\mathbf{W}_{\mathcal{I}, \mathcal{I}}^{2}\right)+\ln \operatorname{det}\left(\tilde{\mathbf{A}}_{\mathcal{I}}^{T} \tilde{\mathbf{A}}_{\mathcal{I}}\right) \\
& =\sum_{i \in \mathcal{I}} \ln w_{i i}^{2}+\sum_{i=1}^{|\mathcal{I}|} \ln \Lambda_{i}\left(\tilde{\mathbf{A}}_{\mathcal{I}}^{T} \tilde{\mathbf{A}}_{\mathcal{I}}\right)
\end{aligned}
$$

where $\Lambda_{i}(\cdot)$ denotes the $i$-th eigenvalue of the corresponding matrix. Then, observe that, by Gerschgorin's Theorem,

$$
\left|\Lambda_{i}\left(\tilde{\mathbf{A}}_{\mathcal{I}}^{T} \tilde{\mathbf{A}}_{\mathcal{I}}\right)-1\right| \leq \sum_{j \in \mathcal{I}, j \neq i}\left|\tilde{\mathbf{a}}_{i}^{T} \tilde{\mathbf{a}}_{j}\right| \leq|\mathcal{I}|-1
$$

Therefore, because of (15), we can say

$$
\begin{equation*}
\ln \operatorname{det}\left(\mathbf{A}_{\mathcal{I}}^{T} \mathbf{A}_{\mathcal{I}}\right)=\sum_{i \in \mathcal{I}} \ln w_{i i}^{2}(1+\mathrm{o}(1)) \tag{16}
\end{equation*}
$$

as $m \rightarrow \infty$. Now, after inserting (16) into $\Upsilon(\mathcal{I})$ and some straightforward simplifications, we conclude that for $\mathcal{I} \in \dot{\mathcal{I}}^{k}$

$$
\begin{align*}
\Upsilon(\mathcal{I}) & =\left[\sum_{i \in \mathcal{I} \backslash \mathcal{S}} \ln w_{i i}^{2}\right](1+\mathrm{o}(1))+\Delta(1+2 c d) \ln m \\
& \geq \Delta(1+a+2 c d) \ln m(1+\mathrm{o}(1)) \tag{17}
\end{align*}
$$

and for $\mathcal{I} \in \breve{\mathcal{I}}^{k}$

$$
\begin{align*}
\Upsilon(\mathcal{I})= & {\left[\sum_{i \in \mathcal{I} \backslash \mathcal{S}} \ln w_{i i}^{2}-\sum_{i \in \mathcal{S} \backslash \mathcal{I}} \ln w_{i i}^{2}\right](1+\mathrm{o}(1)) } \\
& +(|\mathcal{I}|-|\mathcal{S}|)(1+2 c d) \ln m \tag{18}
\end{align*}
$$

as $m \rightarrow \infty$.

Establishing the probability of $\mathscr{I}_{1}$ as $m \rightarrow \infty$ : Rewrite the left-hand side of $\mathscr{I}_{1}$ as

$$
\begin{align*}
& \ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2(m-|\mathcal{S}|-2)}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2(m-|\mathcal{I}|-2)}} \\
& \quad=(m-|\mathcal{I}|-2) \ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}+\Delta \ln \left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2} \tag{19}
\end{align*}
$$

Observe, now, that

$$
\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}=\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}+\left(\boldsymbol{\Pi}_{\mathcal{S}}^{\perp}-\boldsymbol{\Pi}_{\mathcal{I}}^{\perp}\right) \boldsymbol{\epsilon}\right\|_{2}^{2}=\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}+\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}
$$

where $\bar{\Pi}_{\mathcal{I} \backslash \mathcal{S}}$ is defined as $\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}}=\boldsymbol{\Pi}_{\mathcal{I}}-\mathbf{\Pi}_{\mathcal{S}}$. Moreover, note that, $\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\Pi}_{\mathcal{S}}=\mathbf{0}$ since $\mathcal{S} \subset \mathcal{I}$. Thus, we rewrite

$$
\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}=\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}+\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}
$$

Then, by using the properties of the logarithm function, we can say

$$
\ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}=\ln \left(1+\frac{\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}\right) \leq \frac{\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}
$$

Next we insert the preceding inequality into (19) to get

$$
\begin{align*}
& \ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2(m-|\mathcal{S}|-2)}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2(m-|\mathcal{I}|-2)}} \\
& \quad \leq(m-|\mathcal{I}|-2) \frac{\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}+\Delta \ln \left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2} \tag{20}
\end{align*}
$$

Considering (20) together with $\mathscr{I}_{1}$, it is clear that on the event

$$
\begin{aligned}
\stackrel{\circ}{E}_{k}: \max _{\mathcal{I} \in \mathcal{I}^{k}} & \left\{(m-|\mathcal{I}|-2) \frac{\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}+\Delta \ln \left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}\right\} \\
& <\min _{\mathcal{I} \in \overline{\mathcal{I}}^{k}}\{\Upsilon(\mathcal{I})\}-\ln \sigma^{2 \Delta}
\end{aligned}
$$

for all $k \in\{|\mathcal{S}|+1, \ldots, K\}, \mathscr{I}_{1}$ is satisfied. Consequently, when $\boldsymbol{\epsilon}$ is a random vector

$$
\begin{equation*}
\operatorname{Pr} \mathscr{I}_{1} \geq \operatorname{Pr}\left\{\bigcap_{k=1+|\mathcal{S}|}^{K} \stackrel{\circ}{E}_{k}\right\} \tag{21}
\end{equation*}
$$

As $m \rightarrow \infty$, the event $\stackrel{\circ}{E}_{k}$ can be restated by its asymptotic equivalent. In this regard, using the law of large numbers, $\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2} \rightarrow m-|\mathcal{S}|$, together with (17) yields

$$
\begin{equation*}
\stackrel{\circ}{E}_{k}: m \frac{\max _{\mathcal{I} \in \grave{\mathcal{I}}^{k}}\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}}{m-\max _{\mathcal{I} \in \dot{\mathcal{I}}^{k}}\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I}} \boldsymbol{\epsilon}\right\|_{2}^{2}}<\Delta(a+2 c d) \ln m \tag{22}
\end{equation*}
$$

as $m \rightarrow \infty$. Now we continue the argument with the goal of showing, under some suitable condition for $c, \operatorname{Pr} \stackrel{\circ}{\circ}_{k}=1$ as $m \rightarrow \infty$. The idea is to use the extreme value theory to show that $\max _{\mathcal{I} \in \dot{\mathcal{I}}^{k}}\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}$ and $\max _{\mathcal{I} \in \dot{\mathcal{I}}^{k}}\left\|\boldsymbol{\Pi}_{\mathcal{I}} \boldsymbol{\epsilon}\right\|_{2}^{2}$ are of the order of $\ln m$ in probability, i.e. $\max _{\mathcal{I} \in \dot{\mathcal{I}}^{k}}\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}=$ $\mathrm{O}_{p}(\ln m)$ [23]. To develop this idea, consider $\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}$ 's as a sequence of $\binom{N}{\Delta}$ random variables, identically distributed, having Chi-square distribution with $\Delta$ degrees of freedom, $\left\|\bar{\Pi}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2} \sim \chi_{\Delta}^{2}$. Therefore, by applying Lemma 2 and Lemma 3, we can conclude that, for any $t>0$, the centered and normalized maximum of this sequence obeys

$$
\operatorname{Pr}\left\{\left[\max _{\mathcal{I} \in \mathcal{I}^{k}}\left\{\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}\right\}-h\right] \leq 2 t\right\} \geq \exp \left(-e^{-t}\right)
$$

as $m \rightarrow \infty$, where $h$ is defined by

$$
h=2 \ln \binom{N}{\Delta}+(\Delta-2) \ln \ln \binom{N}{\Delta}
$$

Next, replace $h$ with its asymptotic approximation

$$
h=2 d \Delta \ln m(1+\mathrm{o}(1))
$$

and let $t=\gamma \ln m$. Thus, as $m \rightarrow \infty$,

$$
\begin{equation*}
\operatorname{Pr}\left\{\max _{\mathcal{I} \in \dot{\mathcal{I}}^{k}}\left\{\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}\right\} \leq 2(\Delta d+\gamma) \ln m\right\} \geq 1-m^{-\gamma} \tag{23}
\end{equation*}
$$

where $\gamma>0$ is a constant. Then, the Borel-Cantelli lemma implies that for $\gamma>1$

$$
\begin{equation*}
\max _{\mathcal{I} \in \dot{I}^{k}}\left\{\left\|\overline{\boldsymbol{\Pi}}_{\mathcal{I} \backslash \mathcal{S}} \boldsymbol{\epsilon}\right\|_{2}^{2}\right\} \leq 2(\Delta d+\gamma) \ln m \tag{24}
\end{equation*}
$$

with probability one as $m \rightarrow \infty$. Continuing, one can use the same argument to obtain

$$
\begin{equation*}
m-\max _{\mathcal{I} \in \mathcal{I}^{k}}\left\|\boldsymbol{\Pi}_{\mathcal{I}} \boldsymbol{\epsilon}\right\|_{2}^{2}=m(1+\mathrm{o}(1)) \tag{25}
\end{equation*}
$$

with probability one. We conclude from (22), (24) and (25) that under the condition of

$$
2(\Delta d+\gamma)<\Delta(a+2 c d)
$$

for $\Delta=1, \ldots, K-|\mathcal{S}|$, the event $\stackrel{\circ}{E}_{k}$ occurs with probability one. Finally, because of (21), one can say that $\mathscr{I}_{1}$ holds with probability one.

Rearranging the preceding condition as

$$
\begin{equation*}
c>1+\frac{\gamma}{\Delta d}-\frac{a}{2 d} \tag{26}
\end{equation*}
$$

provides the guideline for choosing $c$ upon setting $\Delta=1$.
Establishing the probability of $\mathscr{I}_{2}$ : Rewrite $\mathscr{I}_{2}$ as
$(m-|\mathcal{I}|-2) \ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}}{\sigma^{2}\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}}-(|\mathcal{I}|-|\mathcal{S}|) \ln \left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}+\Upsilon(\mathcal{I})>0$.
Then, it is clear that on the event

$$
\begin{align*}
\breve{E}_{k}: \min _{\mathcal{I} \in \breve{\mathcal{I}}^{k}}\{ & (m-|\mathcal{I}|-2) \ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}}{\sigma^{2}\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}} \\
& \left.-(|\mathcal{I}|-|\mathcal{S}|) \ln \left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}+\Upsilon(\mathcal{I})\right\}>0 \tag{27}
\end{align*}
$$

for all $k$ such that $0 \leq k \leq K, \mathscr{I}_{2}$ holds. Thus, when $\boldsymbol{\epsilon}$ is a random vector, we can say

$$
\begin{equation*}
\operatorname{Pr} \mathscr{I}_{2} \geq \operatorname{Pr}\left\{\bigcap_{k=0}^{K} \breve{E}_{k}\right\} \tag{28}
\end{equation*}
$$

The next step is to establish a lower bound on the asymptotic equivalent of $(m-|\mathcal{I}|-2) \ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{1} \mathbf{y}\right\|_{2}^{2}}{\sigma^{2}\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{1} \epsilon\right\|_{2}^{2}}$. In this respect, expand $\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}$ as

$$
\begin{equation*}
\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{y}\right\|_{2}^{2}=A_{1}+\sigma^{2}\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2} \tag{29}
\end{equation*}
$$

where $A_{1}$ is defined as

$$
A_{1}=\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}^{2}\left(1+2 \sigma \frac{\boldsymbol{\epsilon}^{T} \boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}}{\left\|\boldsymbol{\Pi}_{\frac{\mathcal{I}}{}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}^{2}}\right)
$$

However,

$$
\begin{aligned}
A_{1} \geq & \left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}^{2} \\
& \times(1-\underbrace{\frac{2 \sigma}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}} \max _{\mathcal{I} \in \overline{\mathcal{I}}^{k}}\left|\boldsymbol{\epsilon}^{T} \frac{\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}}\right|}_{\zeta_{1}})
\end{aligned}
$$

Continuing, when $\boldsymbol{\epsilon}$ is a random vector, because of Lemma 4-(ii) and the Borel-Cantelli lemma, we have

$$
\begin{equation*}
\max _{\mathcal{I} \in \breve{\mathcal{I}}^{k}}\left|\boldsymbol{\epsilon}^{T} \frac{\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}}\right| \leq 2 \sqrt{\ln \binom{N}{|\mathcal{I}|}} \tag{30}
\end{equation*}
$$

with probability one. Moreover, by considering Lemma 4(i), the restricted eigenvalue property and $\left\|\mathbf{a}_{i}\right\|_{2}^{2}=\Omega\left(m^{a}\right)$, we can say $\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}=\Omega\left(\sqrt{m^{a} / \ln m}\right)$. Because of this and (30), we can conclude that $\zeta_{1}=0$ with probability one as $m \rightarrow \infty$. Again, use Lemma 4-(i), so that

$$
A_{1} \geq C_{\min }\left\|\mathbf{x}_{\dot{\mathcal{S}}}\right\|_{2}^{2} \frac{\sum_{i \in \mathcal{S}} w_{i i}^{2}}{\ln m}
$$

where $\mathcal{S}$ is defined by $\mathcal{S} \triangleq \mathcal{S} \backslash \mathcal{I}$. Next, consider (29) together with the preceding inequality and that $\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2} \rightarrow m-|\mathcal{S}|$. Thus, as $m \rightarrow \infty$, we have

$$
\begin{aligned}
&(m-|\mathcal{I}|-2) \ln \frac{\left\|\boldsymbol{\Pi}_{\mathcal{I}} \mathbf{y}\right\|_{2}^{2}}{\sigma^{2}\left\|\boldsymbol{\Pi}_{\mathcal{S}}^{\perp} \boldsymbol{\epsilon}\right\|_{2}^{2}} \\
& \geq m \ln (1+\frac{A_{1}}{\sigma^{2} m}-\underbrace{\frac{\max _{\mathcal{I} \in \breve{\mathcal{I}}^{k}}\left\|\boldsymbol{\Pi}_{\mathcal{I}} \boldsymbol{\epsilon}\right\|_{2}^{2}}{m}}_{\zeta_{2}}) \\
& \quad \geq m \ln \left(1+\frac{C_{\min }\left\|\mathbf{x}_{\dot{\mathcal{S}}}\right\|_{2}^{2}}{\sigma^{2} m \ln m} \sum_{i \in \mathcal{S}} w_{i i}^{2}-2(d|I|+\dot{\gamma}) \frac{\ln m}{m}\right)
\end{aligned}
$$

Note that the inequality $T$ is due to the fact that $\zeta_{2} \leq 2(d|I|$ $+\dot{\gamma}) \frac{\ln m}{m}$ with probability one for some constant $\dot{\gamma}>1$, cf. (24). Now, by exploiting the preceding inequality and (18), we can bound the inner expression in $E_{k}$ from below as

$$
\begin{equation*}
A_{2}+\mathrm{O}(\ln m) \tag{31}
\end{equation*}
$$

where
$A_{2}=m \ln \left(1+\frac{C_{\min }\left\|\mathbf{x}_{\mathcal{S}}\right\|_{2}^{2}}{\sigma^{2} m \ln m} \sum_{i \in \mathcal{S}} w_{i i}^{2}(1+\mathrm{o}(1))\right)-\sum_{i \in \mathcal{S}} \ln w_{i i}^{2}$.
It is straightforward to show that $A_{2}$ is an increasing function of $w_{i i}^{2}$. Therefore, since $w_{i i}^{2}=\Omega\left(m^{a}\right), A_{2}$ has the smallest growth if $0<a \leq 1$, so that

$$
A_{2} \geq b|\mathcal{S}| \frac{\left\|\mathbf{x}_{\mathcal{\mathcal { S }}}\right\|_{2}^{2}}{\sigma^{2}} \frac{m^{a}}{\ln m}, \quad 0<a \leq 1
$$

for some constant $b>0$. This implies that $A_{2}$ is the dominant term in (31) and therefore, the event $\breve{E}_{k}$ occurs with probability one as $m \rightarrow \infty$. Finally, since (28), we obtain $\operatorname{Pr} \mathscr{I}_{2}=1$ as $m \rightarrow \infty$.

In summary, because $\operatorname{Pr} \mathscr{I}_{1}=1$ under the condition (26) and that $\operatorname{Pr} \mathscr{I}_{2}=1$, Lemma 1 provides that $\hat{\mathcal{I}}=\mathcal{S}$ with probability one.

We would like to remark that the condition for choosing $c$ in Theorem 2 can be restated as $c>1-a /(2 d)+\gamma / d$ for
$\gamma>0$ where $\gamma$ affects the probability of success. In fact, our proof indicates that when $\gamma \leq 1$ one can only assure the event $\hat{\mathcal{I}}=\mathcal{S}$ in probability as $m \rightarrow \infty$, whereas when $\gamma>1$, this event occurs with probability one as $m \rightarrow \infty$. Having this stated, we also would like to emphasize on the effect of $\gamma>1$ on the performance of EFIC for practical size of $m$. In this regard, (23) indicates that when $\gamma$ is large, the probability of success of $\mathscr{I}_{1}$ is elevated. However, at the same time, a large $\gamma$ lowers the probability of success of (31), and consequently $\operatorname{Pr} \mathscr{I}_{2}$, for small size $m$. The latter happens since a large $\gamma$ may cause the $\mathrm{O}(\ln m)$ term to dominate $A_{2}$. In conclusion, we do not recommend to use a large $\gamma$. In the simulations in Section V we use $\gamma=2$.

## IV. Computing EFIC for Linear Regressions

Finding the optimal solution of (9) essentially involves an exhaustive search over $\mathcal{J}$. However, the size of $\mathcal{J}$ grows as $|\mathcal{J}|=\mathrm{O}\left(N^{K}\right)$, and therefore, solving (9) with an exhaustive search is impractical, even for a moderate size $N$. To overcome this issue, motivated by the intrinsic nature of the Lasso estimator as a model selection method, we exploit Lasso to improve the computational performance of (9). The Lasso estimator is a well-known model selection method that estimates the unknown model by solving

$$
\begin{equation*}
\hat{\mathbf{x}}^{\mathrm{L}}(\lambda)=\underset{\tilde{\mathbf{x}} \in \mathbb{R}^{N}}{\arg \min } \frac{1}{2}\|\mathbf{y}-\tilde{\mathbf{A}} \tilde{\mathbf{x}}\|_{2}^{2}+\lambda\|\tilde{\mathbf{x}}\|_{1} \tag{32}
\end{equation*}
$$

where $\lambda \geq 0$ is a regularization parameter that controls $\hat{\mathcal{I}}^{\mathrm{L}}(\lambda)=\operatorname{supp}\left(\hat{\mathbf{x}}^{\mathrm{L}}(\lambda)\right)$, the vector $\tilde{\mathbf{x}}$ is linked to $\mathbf{x}$ as $\mathbf{x}=\mathbf{W}^{-1} \tilde{\mathbf{x}}$ and $\mathbf{W}$ is a diagonal matrix with the diagonal elements defined as $w_{i i}=\left\|\mathbf{a}_{i}\right\|_{2}$ [24]. If $\lambda$ is properly chosen, then under some suitable conditions, Lasso can estimate the model correctly [25], [26]. To elaborate on the effect of $\lambda$ on the solution of Lasso, one should note that for the choice of $\lambda \geq \lambda_{1}=\left\|\mathbf{A}^{T} \mathbf{y}\right\|_{\infty}$, Lasso's estimate is an empty set, then as $\lambda$ decreases, $\hat{\mathcal{I}}^{\mathrm{L}}(\lambda)$ evolves at some pivotal $\lambda_{j}$ 's, providing the solution set of $\left\{\hat{\mathcal{I}}^{\mathrm{L}}\left(\lambda_{j}\right)\right\}_{j=1}^{K}$ [14], [27]. Having this stated, we use the modified-LARS algorithm to obtain $\left\{\hat{\mathcal{I}}^{\mathrm{L}}\left(\lambda_{j}\right)\right\}_{j=1}^{K}$ [14], and then we apply (10) on the solution set of Lasso to solve (9); see Alg. 1. Note that we use the normalized regressor matrix $\tilde{\mathbf{A}}$ in the Lasso estimator, but the un-normalized $\mathbf{A}$ in computing the EFIC criterion in (9). This is because EFIC is derived under the assumption that the dependence on $m$ should be reflected in A and not in x. However, when using Lasso it is recommended to normalize the regressors.

```
\(\overline{\text { Algorithm } 1 \text { Model selection by combining EFIC and the modified- }}\)
LARS algorithm.
    for \(j=1\) to \(K\) do
        \(\left\{\hat{\mathcal{I}}^{\mathrm{L}}\left(\lambda_{j}\right)\right\} \leftarrow\) execute modified-LARS at step \(j\)
        evaluate \(g\left(\hat{\mathcal{I}}^{\mathrm{L}}\left(\lambda_{j}\right)\right)\)
    end for
    \(\hat{\mathcal{I}}=\arg \min _{\left\{\hat{\mathcal{I}}^{\mathrm{L}}\left(\lambda_{j}\right)\right\}_{j=1}^{K}} g\left(\hat{\mathcal{I}}^{\mathrm{L}}\left(\lambda_{j}\right)\right)\)
```

Compared to performing an exhaustive search, we reduce the search cost in solving EFIC from $\mathrm{O}\left(m^{d K}\right)$ to the computational cost of modified-LARS, which is $\mathrm{O}\left(\mathrm{m}^{2}\right)$ at each


Fig. 2. The behavior of a few model selection criteria versus the indices of the solution set of Lasso. The setting is $\sigma^{2}=10^{-1},|\mathcal{S}|=5, m=100$ and $N=\left\lceil m^{d}\right\rceil$, for $d=1.3$. Label six corresponds to the true model.
step [14]. Another observation is that we can find the ideal $\lambda$ value for the Lasso estimator without knowing $\sigma$. In details, when $\sigma$ is known, by choosing $\lambda=(1+\tau) \sigma \sqrt{2 \ln N}$ for some $\tau>0$, Lasso performs near-ideal in estimating the model [26]. However, the proper choice of $\lambda$ when $\sigma$ is unknown is rather challenging [28] and to the best of our knowledge, selecting $\lambda$ in such scenarios is still a practical issue. We would like to remark that the choice of Lasso is not a necessity and the operator can potentially use any other viable variable selection method for high-dimensional data.

A numerical example of the behavior of the objective function of a few model selection criteria versus $\left\{\hat{\mathcal{I}}^{\mathrm{L}}\left(\lambda_{j}\right)\right\}_{j=1}^{20}$ is depicted in Fig. 2. The setup for this numerical evaluation is as $\sigma^{2}=10^{-1},|\mathcal{S}|=5, m=100, N=\left\lceil m^{d}\right\rceil$, for $d=1.3$. The abrupt decrease in the value of the objective functions at label six is due to the perfect model selection. After label six, $g(\cdot)$ monotonically increases. As was mentioned before, this increase is caused by $g_{2}(\cdot)$ in order to offset $g_{1}(\cdot)$ from overfitting. EFIC has estimated the model correctly whereas EBIC and BIC fail in doing so. Here and in the empirical section, we set $\dot{c}=1$ in EBIC, as it is also applied in [11]. One should note that, by choosing $c$ c $>1-1 /(2 d)$, EBIC's estimate theoretically satisfies $\operatorname{Pr}\{\hat{\mathcal{I}}=\mathcal{S}\} \rightarrow 1$ as $m \rightarrow \infty$. However, when $m$ is of practical size, EBIC's performance is sensitive to the value of $c$. That is to say, small $\dot{c}$ causes EBIC to overestimate the size of the model, whereas large $\dot{c}$ would cause the opposite. Therefore, by setting $\dot{c}$ just above the threshold level $1-1 /(2 d)$, EBIC's performance would deteriorate in the numerical simulations presented herein. Fig. 2 also shows that BIC is too liberal to find the true model.

## V. Empirical Results

In this section, we provide some numerical results to illustrate the empirical performance of EFIC in selecting the correct model. Additionally, we compare the performance of EFIC with the following existing model selection methods.


Fig. 3. The empirical probability of $\{\hat{\mathcal{I}}=\mathcal{S}\}$ versus $m$ when $\mathbf{A}$ has an uncorrelated structure. Here, $\mu=0, \sigma^{2}=10^{-0.3},|\mathcal{S}|=5$ and $N=\left\lceil\mathrm{m}^{d}\right\rceil$ for $d=1.3$.

## A. Some Existing Model Selection Methods for HighDimensional Data

1) Square Root (SR)-Lasso: Replacing the least squares term in (32) with its square-root results in the SR-Lasso estimator [29],

$$
\min _{\tilde{\mathbf{x}} \in \mathbb{R}^{m}}\|\mathbf{y}-\tilde{\mathbf{A}} \tilde{\mathbf{x}}\|_{2}+\lambda_{\mathrm{SR}}\|\tilde{\mathbf{x}}\|_{1}
$$

The parameter $\lambda_{\mathrm{SR}}$ can be chosen by

$$
\lambda_{\mathrm{SR}}=c_{1} \sqrt{m} \mathrm{~F}_{n}^{-1}\left(1-\frac{\alpha_{1}}{2 N}\right)
$$

Here, $\mathrm{F}_{n}^{-1}(\cdot)$ is the quantile function for a normal distribution, and $c_{1}$ and $\alpha_{1}$ are constants that are recommended to set to 1.1 and 0.05 , respectively. Note that SR-Lasso does not need to know $\sigma$ to choose its regularization parameter $\lambda_{\text {SR }}$.
2) Combined BIC (COBIC): The objective function of BIC is presented in an elegant simple form [3]. This simplicity is achieved under some suitable conditions, one of which concerns the signal to noise power ratio. On this subject, it is shown that as the noise power converges to zero, a modified version of BIC, namely COBIC, is needed [16]. Here, an extended version of COBIC, where we have added the binomial coefficient penalty similar to EBIC, is considered as

$$
\min _{\mathcal{I} \in \mathcal{J} \mathrm{JR}}\left\{m \ln \hat{\sigma^{2}}+f+2 \ln \binom{N}{|\mathcal{I}|}\right\},
$$

where $f$ is defined as

$$
f=\max \left(-(|\mathcal{I}|+2) \ln \hat{\sigma^{2}},(|\mathcal{I}|+1) \ln m\right)
$$

## B. Results

Here, to support our theoretical findings, we measure the empirical probability of $\hat{\mathcal{I}}=\mathcal{S}$. The general setting for our numerical simulation is as follows. In each Monte Carlo trial, the true support, $\mathcal{S}$, is chosen randomly from $\binom{N}{|\mathcal{S}|}$ possible choices, when $|\mathcal{S}|$ is fixed to five. The elements of the true parameter $\mathbf{x}_{\mathcal{S}}$ is drawn from the Bernoulli distribution of the


Fig. 4. The empirical probability of $\{\hat{\mathcal{I}}=\mathcal{S}\}$ versus $\ln \left(1 / \sigma^{2}\right)$ when $\mathbf{A}$ has an uncorrelated structure. Here, $\mu=0, m=120,|\mathcal{S}|=5$ and $N=\left\lceil m^{d}\right\rceil$ for $d=1.3$.
sample space $\{1,-1\}$. The rows of the measurement matrix, $\mathbf{a}_{* i}$ 's, are chosen as i.i.d. multivariate Gaussian random vectors from $\mathcal{N}(\mathbf{0}, \mathbf{C})$, where the matrix $\mathbf{C} \in \mathbb{R}^{N}$ is structured as

$$
\mathbf{C}=\left(\begin{array}{cccccc}
1 & \mu & \mu & & \ldots & \mu \\
\mu & 1 & \mu & \mu & \ldots & \mu \\
\mu & \mu & 1 & \mu & \ldots & \mu \\
\vdots & \vdots & & & & \vdots \\
\mu & \mu & \ldots & & \mu & 1
\end{array}\right)
$$

and the constant $\mu$ determines the degree of correlation between $\mathbf{a}_{i}$ 's. The structure of $\mathbf{C}$ also implies that $\mathbf{a}_{i}$ 's are statistically equiangular. The dimension of the parameter space is linked to the number of measurements as $N=\left\lceil m^{d}\right\rceil$, where $\lceil\cdot\rceil$ is the ceiling function and $d$ is fixed to 1.3 . The vector $\epsilon$ is a white Gaussian noise with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To select parameter $a$ for EFIC, recall that the smallest $\left\|\mathbf{a}_{i}\right\|_{2}^{2}$ for $1 \leq i \leq N$ determines the constant $a$. Furthermore, note that $\left\|\mathbf{a}_{i}\right\|_{2}^{2}=m$ with very high probability for a Gaussian random A. This motivates us to set $a=1$. Finally, as mentioned at the end of Section IV, we let $\gamma=2$ which results in $c=1+3 /(2 d)$.

Fig. 3 shows the empirical probability of correct model selection versus $m$ when $\mathbf{A}$ has an uncorrelated structure, i.e. $\mu=0$. Here, the noise variance is fixed as $\sigma^{2}=10^{-0.3}$. The measured probabilities are the result of calculating the success rate over 500 Monte Carlo trials. Here, Lasso-oracle represents the performance of Lasso when $|\mathcal{S}|$ is known. Recall that, in our algorithm, the exhaustive search over $\mathcal{J}$ is replaced by Lasso; therefore, it is natural to have Lassooracle as the performance benchmark. As can be seen, for $m \geq 80$ EFIC coincides with Lasso-oracle and both achieve $\operatorname{Pr}(\hat{\mathcal{I}}=\mathcal{S})=1$ for $m \geq 120$. The numerical results conform that EFIC is consistent in selecting the true model as $m$ grows. Moreover, the numerical simulation shows that $\operatorname{Pr}\{\hat{\mathcal{I}}=\mathcal{S}\}$ for EBIC improves very slowly for $m \geq 100$. The slack pace in the improvement is due to the conservative choice of $\dot{c}=1$. COBIC is practically identical to EBIC for this example and therefore its curve is not displayed in Fig. 3. The performance gap between SR-Lasso and Lasso-oracle is very


Fig. 5. The empirical probability of $\{\hat{\mathcal{I}}=\mathcal{S}\}$ versus $m$ when $\mathbf{A}$ has a correlated structure. Here, $\mu=0.25, \sigma^{2}=10^{-0.3},|\mathcal{S}|=5$ and $N=\left\lceil m^{d}\right\rceil$ for $d=1$.3.
big for $m<140$, nevertheless, SR-Lasso manages to achieve ideal performance as $m$ grows. Finally, Lasso- $\sigma$ shows how Lasso behaves if $\lambda=7 \sigma \sqrt{2 \ln N}$ when $\sigma$ is known [26].

We now consider the effect of a decrease in $\sigma$ on $\operatorname{Pr}(\hat{\mathcal{I}}=$ $\mathcal{S})=1$ when $m=120$ and $\mathbf{A}$ has an uncorrelated structure. Fig. 4 illustrates the empirical probability of correct model selection versus $\ln 1 / \sigma^{2}$ over 5000 Monte Carlo trials. As can be seen, all information criteria perform poorly for large $\sigma$. This can be linked to the poor performance of Lasso in estimating the model in presence of strong noise. As $\sigma$ slightly decreases, EFIC achieves the ideal performance. However, in contrast to EFIC, the success rate of EBIC always stays below 0.95. This indicates that EBIC is inconsistent in finding $\mathcal{S}$ as $\sigma \rightarrow 0$. Fig. 4 also shows that COBIC achieves $\operatorname{Pr}(\mathcal{I}=\mathcal{S})=1$ at a very slow pace. Finally, note the unsatisfactory performance of SR-Lasso. This is because SRLasso requires a larger $m$ than Lasso to achieve the same performance in estimating models.

Next, we examine the effect of the correlation in the structure of $\mathbf{A}$ on the performances of the model selection criteria. Fig. 5 illustrates the empirical probability of correct model selection versus $m$ over 500 Monte Carlo trials when $\mu=0.25$. Here, the rest of the setting is identical to the corresponding uncorrelated case. As can be seen, EFIC's and Lasso-oracle's performances are identical and both achieve $\operatorname{Pr}\{\hat{\mathcal{I}}=\mathcal{S}\}=1$ for $m \geq 200$. Note that based on Algorithm 1, the performance of EFIC is tied to the solution set of Lasso. Thus, EFIC cannot outperform Lasso-oracle. Predictively, due to the correlation between $\mathbf{a}_{i}$ 's, Lasso-oracle requires relatively larger number of measurements to achieve ideal model selection. The numerical simulation again shows that the performance of EBIC improves very slowly for $m \geq 200$. Interestingly, the performance gap between Lasso$\sigma$ and Lasso-oracle is tightened. Finally, SR-Lasso performs rather poorly which indicates that SR-Lasso is more sensitive to the correlation in the structure of $\mathbf{A}$ than Lasso.

Fig. 6 shows the empirical probability of $\hat{\mathcal{I}}=\mathcal{S}$ versus $\ln 1 / \sigma^{2}$ over 5000 Monte Carlo trials. Here, $\mu=0.25$ and $m$ is fixed to 200. EFIC performs as good as the benchmark


Fig. 6. The empirical probability of $\{\hat{\mathcal{I}}=\mathcal{S}\}$ versus $\ln \left(1 / \sigma^{2}\right)$ when $\mathbf{A}$ has a correlated structure. Here, $\mu=0.25, m=200,|\mathcal{S}|=5$ and $N=\left\lceil m^{d}\right\rceil$ for $d=1.3$.
for $\ln 1 / \sigma^{2} \geq 0$ and it achieves ideal model selection for $\ln 1 / \sigma^{2} \geq \overline{1}$. In contrast, EBIC's performance does not improve as $\ln 1 / \sigma^{2}$ gets larger than one, indicating the inconsistency of EBIC as $\sigma \rightarrow 0$. The numerical simulation also shows that COBIC's estimate achieves $\operatorname{Pr}\{\hat{\mathcal{I}}=\mathcal{S}\}=1$ for a very small $\sigma$. Lasso- $\sigma$ shows an abrupt change in its performance. This is due to setting $\tau$ to a relatively large value [26]. Predictively, SR-Lasso again performs poorly.

## VI. Conclusion

Many real life applications face the challenging task of model selection when the number of accessible measurements is much smaller than the dimension of the parameter space. To accomplish this task, we have proposed a new criterion for high-dimensional linear regression. Additionally, we have analyzed the behavior of the proposed criterion as $m \rightarrow \infty$ as well as $\sigma \rightarrow 0$. In both cases, we have shown that the probability that our criterion gives the true model approaches one. Moreover, we propose a computationally affordable algorithm to practically perform the model selection with the proposed criterion. This algorithm implicitly determines the regularization parameter in the Lasso estimator for precise variable selection. Finally, some numerical simulations are conducted to support our theoretical findings. Moreover, we would like to remark that our focus in this paper has been on perfect variable selection. Arguably, there are also other broader aspects to the problem of model selection. In some applications, selecting relevant variables is indeed a major interest whereas in others the predictive ability of the model is the main focus. Hence, other measures of performance such as positive discovery rate, false discovery rate, and prediction mean square error should also be considered in future evaluations of EFIC and related approaches.

## VII. Appendix

Lemma 2. Let $M_{n}=\max _{i}\left\{X_{1}, X_{2}, \ldots, X_{n}\right\}$, where $X_{1}, X_{2}, \ldots, X_{n}$ is a sequence of i.i.d. random variables having Chi-square distribution with $r$ degrees of freedom.

Then, the centered and normalized maxima of this sequence has the Gumbel distribution such that

$$
\operatorname{Pr}\left\{\frac{M_{n}-h_{n}}{2} \leq x\right\} \xrightarrow{d} \exp \left(-e^{-x}\right) \quad \text { as } n \rightarrow \infty
$$

where

$$
h_{n}=2 \ln n+(r-2) \ln \ln n(1+\mathrm{o}(1))
$$

and $\xrightarrow{d}$ denotes convergence in distribution.
Proof. Based on Fisher-Tippet theorem, if there exists a sequence of norming constants $\left\{c_{n}>0\right\}$ and $\left\{d_{n}\right\}$ such that

$$
\operatorname{Pr}\left\{\frac{M_{n}-d_{n}}{c_{n}} \leq x\right\} \xrightarrow{d} G(x) \quad \text { as } n \rightarrow \infty
$$

where $G(x)$ is a non-degenerate distribution function, then $G$ belongs to one of the extreme value distributions [23], [30]. Knowing that the Chi-square distribution is a special case of the Gamma distribution, verifies that $\chi^{2}$ random variables belong to the maximum domain of attraction of the Gumbel distribution, so that $G(x)=\exp \left(-e^{x}\right)$ [23]. Accordingly, Proposition 3.3.25 in [23] provides a possible choice of the norming constants. Exploiting the provided constants on page 159 in [23], we have

$$
d_{n}=2 \ln n+(r-2) \ln \ln n(1+\mathrm{o}(1)), \quad c_{n}=2
$$

Lemma 3. Let $M_{\kappa}=\max _{i}\left\{X_{i}\right\}_{i=1}^{\kappa}$ where $X_{1}, X_{2}, \ldots, X_{\kappa}$ is a sequence of i.i.d. random variables having Chi-square distribution with $r<m$ degrees of freedom. Further, let $\bar{M}_{\kappa}=\max _{i}\left\{\left\|\boldsymbol{\Pi}_{i} \boldsymbol{\epsilon}\right\|_{2}^{2}\right\}_{i=1}^{\kappa}$ where $\left\{\boldsymbol{\Pi}_{i}\right\}_{i=1}^{\kappa}$ is a set of orthogonal projection matrices projecting onto $r$-dimensional subspaces of $\mathbb{R}^{m}$, and $\epsilon \in \mathbb{R}^{m}$ is a normalized zero-mean Gaussian vector, $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_{m}\right)$.

If $m<r \kappa$, then, for each $t>0$,

$$
\begin{equation*}
\operatorname{Pr}\left\{\bar{M}_{\kappa} \leq t\right\} \geq \operatorname{Pr}\left\{M_{\kappa} \leq t\right\} \tag{33}
\end{equation*}
$$

Proof. Consider the random vectors $\mathbf{t}$ and $\overline{\mathbf{t}}$, defined as

$$
\begin{aligned}
\mathbf{t} & =\left[\begin{array}{llll}
\mathbf{t}_{1} & \mathbf{t}_{2} & \ldots & \mathbf{t}_{\kappa}
\end{array}\right]^{T} \\
\overline{\mathbf{t}} & =\left[\begin{array}{llll}
\mathbf{U}_{1} & \mathbf{U}_{2} & \ldots & \mathbf{U}_{\kappa}
\end{array}\right]^{T} \boldsymbol{\epsilon}
\end{aligned}
$$

where the $\mathrm{t}_{i}$ 's are $r$-dimensional real Gaussian random row vectors such that $\mathbf{t} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_{r \kappa}\right)$ and $X_{i}=\mathbf{t}_{i} \mathbf{t}_{i}^{T}$. Moreover, $\left\{\mathbf{U}_{i} \in \mathbb{R}^{m \times r}\right\}_{i=1}^{\kappa}$ are orthonormal matrices such that $\boldsymbol{\Pi}_{i}=$ $\mathbf{U}_{i} \mathbf{U}_{i}^{T}$. Observe that $M_{\kappa}$ and $\bar{M}_{\kappa}$ are obtained by applying the function

$$
f\left(\left[\begin{array}{llll}
\mathbf{b}_{1} & \mathbf{b}_{2} & \ldots & \mathbf{b}_{\kappa}
\end{array}\right]^{T}\right)=\max _{j}\left\{\left\|\mathbf{b}_{j}\right\|_{2}^{2}\right\}_{j=1}^{\kappa}
$$

on $\mathbf{t}$ and $\overline{\mathbf{t}}$, respectively, where $\mathbf{b}_{j} \in \mathbb{R}^{r}$. Note that the subvectors of $\overline{\mathbf{t}}, \overline{\mathbf{t}}_{i}=\mathbf{U}_{i}^{T} \boldsymbol{\epsilon}$, are $r$-dimensional real Gaussian random vectors similar to the $\mathbf{t}_{i}$ 's. However,

$$
\overline{\mathbf{t}} \in \operatorname{span}\left(\left[\begin{array}{llll}
\mathbf{U}_{1} & \mathbf{U}_{2} & \ldots & \mathbf{U}_{\kappa}
\end{array}\right]^{T}\right)
$$

whereas $\mathbf{t} \in \mathbb{R}^{r \kappa}$. This implies, when $m<r \kappa, \overline{\mathbf{t}}$ is constrained to a subspace of $\mathbb{R}^{r \kappa}$ and therefore (33) follows.

Lemma 4. Let $\mathcal{S}$ be $\mathcal{S}=\{\mathcal{S} \backslash \mathcal{I}\}$ for $\mathcal{I} \in \breve{\mathcal{I}}^{k}$ and constant $B>0$. Then, we have
(i) $\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}^{2} \geq \Lambda_{\min }\left(\tilde{\mathbf{A}}_{\mathcal{I} \cup \mathcal{S}}^{T} \tilde{\mathbf{A}}_{\mathcal{I} \cup \mathcal{S}}\right)\left\|\mathbf{x}_{\mathcal{\mathcal { S }}}\right\|_{2}^{2} \sum_{i \in \mathcal{S}} w_{i i}^{2}$;
(ii) $\operatorname{Pr}\left\{\max _{\mathcal{I} \in \breve{\mathcal{I}}^{k}}\left|\boldsymbol{\epsilon}^{T} \frac{\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}}{\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}}\right|>B \sqrt{2 \ln \binom{N}{|\mathcal{I}|}}\right\}$

$$
\leq(\sqrt{\pi} B)^{-1}\left[\ln \binom{N}{|\mathcal{I}|}\right]^{-1 / 2}\left[\binom{N}{|\mathcal{I}|}\right]^{1-B^{2}}
$$

Proof. For (i), split $\mathcal{S}$ into two disjoint subsets as $\mathcal{S}=\{\mathcal{S} \cap$ $\mathcal{I}\} \cup\{\mathcal{S} \backslash \mathcal{I}\}$. Since $\operatorname{span}\left(\mathbf{A}_{\mathcal{S} \cap \mathcal{I}}\right) \subset \operatorname{span}\left(\boldsymbol{\Pi}_{\mathcal{I}}\right)$, we have

$$
\begin{equation*}
\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}^{2}=\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}^{2} \tag{34}
\end{equation*}
$$

Now consider the closest point theorem to show that

$$
\begin{aligned}
& \left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\dot{\mathcal{S}}} \mathbf{x}_{\mathcal{\mathcal { S }}}\right\|_{2}^{2}=\min _{\boldsymbol{\nu} \in \mathbb{R}^{|\mathcal{I}|}}\left\|\mathbf{A}_{\dot{\mathcal{S}}} \mathbf{x}_{\dot{\mathcal{S}}}-\mathbf{A}_{\mathcal{I}} \boldsymbol{\nu}\right\|_{2}^{2} \\
& \stackrel{T_{1}}{=} \min _{\boldsymbol{\nu} \in \mathbb{R}^{\mid \mathcal{I}}}\left\|\left[\begin{array}{ll}
\tilde{\mathbf{A}}_{\mathcal{\mathcal { S }}} & \tilde{\mathbf{A}}_{\mathcal{I}}
\end{array}\right]\left[\begin{array}{c}
\mathbf{W}_{\dot{\mathcal{S}}, \dot{\mathcal{S}}} \mathbf{x}_{\dot{\mathcal{S}}} \\
-\mathbf{W}_{\mathcal{I}, \mathcal{I}} \boldsymbol{\nu}
\end{array}\right]\right\|_{2}^{2} \\
& \stackrel{T_{2}}{\geq} \Lambda_{\min }\left(\tilde{\mathbf{A}}_{\mathcal{I} \cup \dot{\mathcal{S}}}^{T} \tilde{\mathbf{A}}_{\mathcal{I} \cup \dot{\mathcal{S}}}\right) \min _{\boldsymbol{\nu} \in \mathbb{R}^{|\mathcal{I}|}}\left\|\left[\begin{array}{c}
\mathbf{W}_{\dot{\mathcal{S}}, \dot{\mathcal{S}}} \mathbf{x}_{\dot{\mathcal{S}}} \\
-\mathbf{W}_{\dot{\mathcal{I}}, \dot{\mathcal{I}}} \boldsymbol{\nu}
\end{array}\right]\right\|_{2}^{2} \\
& \geq \Lambda_{\min }\left(\tilde{\mathbf{A}}_{\mathcal{I} \cup \dot{\mathcal{S}}}^{T} \tilde{\mathbf{A}}_{\mathcal{I} \cup \dot{\mathcal{S}}}\right)\left\|\mathbf{x}_{\mathcal{\mathcal { S }}}\right\|_{2}^{2} \sum_{i \in \mathcal{\mathcal { S }}} w_{i i}^{2},
\end{aligned}
$$

where the identity in $T_{1}$ is because of normalizing the corresponding columns of $\mathbf{A}$ and the inequality in $T_{2}$ is due to the definition of the smallest eigenvalue.
For (ii), define $\mathbf{t}$ as $\mathbf{t}=\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}} /\left\|\boldsymbol{\Pi}_{\mathcal{I}}^{\perp} \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\right\|_{2}$ and note that, when $\boldsymbol{\epsilon}$ is a random variable, $\boldsymbol{\epsilon}^{T} \mathbf{t}$ is a standard Gaussian random variable with $\epsilon^{T} \mathbf{t} \sim \mathcal{N}(0,1)$. Exploiting Boole's inequality, for any $\alpha>0$, we have

$$
\operatorname{Pr}\left\{\max _{\mathcal{I} \in \mathcal{I}^{\mathfrak{K}}}\left|\boldsymbol{\epsilon}^{T} \mathbf{t}\right|>\alpha\right\} \leq 2\binom{N}{|\mathcal{I}|} \operatorname{Pr}\left\{\boldsymbol{\epsilon}^{T} \mathbf{t}>\alpha\right\} .
$$

Continuing, by the tail distribution of a Gaussian random variable, we have

$$
\operatorname{Pr}\left\{\boldsymbol{\epsilon}^{T} \mathbf{t}>\alpha\right\}<\frac{1}{\alpha \sqrt{2 \pi}} e^{-\alpha^{2} / 2}
$$

Now letting $\alpha=B \sqrt{2 \ln \binom{N}{|\mathcal{I}|}}$ results in the statement in (ii).

## Acknowledgment

Arash Owrang would like to thank Prof. Alexandre Proutiere for constructive discussions and for offering insightful comments about parts of this paper.

## REFERENCES

[1] S. Konishi and G. Kitagawa, Information criteria and statistical modeling. Springer Science \& Business Media, 2008.
[2] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol. 19, no. 6, pp. 716-723, Dec. 1974.
[3] G. Schwarz, "Estimating the dimension of a model," Ann. Statist., vol. 6, no. 2, pp. 461-464, Mar. 1978.
[4] R. Shibata, "An optimal selection of regression variables," Biometrika, vol. 68, no. 1, pp. 45-54, Apr. 1981.
[5] D. M. A. Haughton, "On the choice of a model to fit data from an exponential family," The Annals of Statistics, vol. 16, no. 1, pp. 342355, Mar. 1988.
[6] 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, Sep. 1987.
[7] J. Rissanen, "Modeling by shortest data description," Automatica, vol. 14, no. 5, pp. 465-471, Sep. 1978.
[8] C. R. Rao and Y. Wu, On model selection, ser. Lecture NotesMonograph Series. Institute of Mathematical Statistics, 2001, vol. Volume 38, pp. 1-57.
[9] M. Woodroofe, "On model selection and the arc sine laws," Ann. Statist., vol. 10, no. 4, pp. 1182-1194, Dec. 1982.
[10] K. W. Broman and T. P. Speed, "A model selection approach for the identification of quantitative trait loci in experimental crosses," J. $R$. Stat. Soc.: Series B (Stat. Methodol.), vol. 64, no. 4, pp. 641-656, Oct. 2002.
[11] J. Chen and Z. Chen, "Extended bayesian information criteria for model selection with large model spaces," Biometrika, vol. 95, no. 3, pp. 759771, Sep. 2008.
[12] S. M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1998.
[13] A. Owrang and M. Jansson, "Model selection for high-dimensional data," in Signals, Systems and Computers, 2016 50th Asilomar Conference on. IEEE, 2016, pp. 606-609.
[14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," Ann. Statist., vol. 32, no. 2, pp. 407-499, Apr. 2004.
[15] R. L. Graham, D. E. Knuth, and O. Patashnik, Concrete mathematics: a foundation for computer science. Pearson Education India, 1994.
[16] P. Stoica and P. Babu, "On the proper forms of bic for model order selection," IEEE Trans. Signal Process., vol. 60, no. 9, pp. 4956-4961, Sep. 2012.
[17] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," Comm. Pure Appl. Math., vol. 59, no. 6, pp. 907 - 934, Jul. 2006.
[18] E. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," Ann. of Statist., vol. 35, no. 6, pp. 2313-2351, Jan. 2007.
[19] S. Foucart and H. Rauhut, A mathematical introduction to compressive sensing. Birkhäuser Basel, 2013, vol. 1, no. 3.
[20] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," Ann. Statist., vol. 37, no. 1, pp. 246270, Jan. 2009.
[21] E. Candes and T. Tao, "Decoding by linear programming," IEEE Trans. Inf. Theory, vol. 51, no. 12, pp. 4203-4215, Dec. 2005.
[22] J. H. Curtiss, "On the distribution of the quotient of two chance variables," Ann. Math. Stat., vol. 12, no. 4, pp. 409-421, Dec. 1941.
[23] P. Embrechts, C. Klüppelberg, and T. Mikosch, Modelling extremal events: for insurance and finance. Springer Science \& Business Media, 2013, vol. 33.
[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," J. R. Stat. Soc. Series B (Methodol.), vol. 58, no. 1, pp. 267-288, Jan. 1996. [Online]. Available: http://www.jstor.org/stable/2346178
[25] P. Zhao and B. Yu, "On model selection consistency of lasso," J. Mach. Learn. Res., vol. 7, pp. 2541-2563, Nov. 2006.
[26] E. Candes and Y. Plan, "Near-ideal model selection by $\ell_{1}$ minimization," Ann. of Statist., vol. 37, no. 5A, pp. 2145-2177, Oct. 2009.
[27] S. Rosset and J. Zhu, "Piecewise linear regularized solution paths," The Ann. of Statist., vol. 35, no. 3, pp. 1012-1030, Jul. 2007.
[28] S. Chretien and S. Darses, "Sparse recovery with unknown variance: A lasso-type approach," IEEE Tran. Inf. Theory, vol. 60, no. 7, pp. 39703988, Jul. 2014.
[29] A. Belloni, V. Chernozhukov, and L. Wang, "Square-root lasso: pivotal recovery of sparse signals via conic programming," Biometrika, vol. 98, no. 4, pp. 791-806, Dec. 2011.
[30] M. Leadbetter and G. Lindgren, Rootz en, H.(1983) Extremes and related properties of random sequences and processes. Springer, Berlin, 1983.


[^0]:    This work was supported in part by the Swedish Research Council under contract 2015-05484. This project has received funding from 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 Department of Information Science and Engineering and the ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Stockholm, Sweden (e-mail: owrang@kth.se; janssonm@kth.se).

