Model Selection for High-Dimensional Data

Arash Owrang, and Magnus Jansson

ACCESS Linnaeus Center,
School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm, Sweden

Email: owrang@kth.se, janssonm@kth.se

Abstract—We investigate the task of model selection for high-dimensional data. For this purpose, we propose an extension to the Bayesian information criterion. Our information criterion is asymptotically consistent either as the number of measurements tends to infinity or as the variance of noise decreases to zero. The numerical results provided support our claim. Additionally, we highlight the link between model selection for high-dimensional data and the choice of hyper-parameter in ℓ_1 -constrained estimators, specifically the LASSO.

I. INTRODUCTION

Model selection is the task of selecting a particular model from a set of competitive models, given the measurement. This task is of crucial importance as the development of further statistical inferences relies on the goodness of the modeling. As a result, many methods have been developed, reflecting the importance of this task [1]–[3]. Many of these methods, which inherit the term of information criteria as a part of their name, are the consequence of the seminal work of Akaike [1]. In principle, Akaike’s information criterion chooses a model which fits best the data from a set of hierarchically nested models. Shortly after [1] was introduced, Schwartz proposed an alternative approach to model selection [2]. His approach led to a method, sometimes referred to as Bayesian information criterion (BIC), which is proved to be asymptotically consistent as the sample size increases.

BIC initially has been developed to choose the appropriate dimensionality of a model; a model like the polynomial regression or the auto-regressive model [4], [5]. Precisely speaking, BIC is tailored for the purpose of model selection from a set of hierarchically nested parameter spaces [3]. However, in applications that require subset selection, the competitive parameter spaces are rather structured more general than being hierarchically nested in space. In fact, this generalized set of competitive spaces is a collection of combinatorial parameter spaces up to a certain cardinality. It has been shown that traditional information criteria are too liberal in selecting a model for such a setting [6]. To overcome this issue, an extension to an approximation of BIC (EBIC) is proposed [6].

Consistency is often a desirable property of a statistical procedure. We have investigated the consistency of EBIC

either as the number of the measurements tends to infinity or as the variance of noise converges to zero. Unfortunately, EBIC is inconsistent in selecting the true model as the variance of noise converges to zero. Moreover, the performance of EBIC improves slowly with the number of measurements. To overcome these issues, inspired by the novel work in [6], we propose a consistent extended form of BIC (CEBIC) for high-dimensional settings. Theoretical details in this regard will be provided in a future version of this paper. Here, we provide some numerical simulations to assess the performance of CEBIC. The numerical results in Sec. III support our claims. Furthermore, we briefly address the link between CEBIC and hyper-parameter selection for the LASSO [8], [9].

II. EXTENDING BIC

A. Background

Let $\mathbf{y} \in \mathbb{R}^m$ represent the real measurement vector. In the classical multiple hypothesis testing setup [3], it is assumed that under each $\mathcal{H}_i$, we have

$$\mathcal{H}_i : \mathbf{y} = s(\boldsymbol{\theta}^i) + \sigma\boldsymbol{\epsilon}, \quad (1)$$

for $i \in \{0, 1, \dots, N\}$. The general signal function s , which maps $\mathbb{R}^N$ to $\mathbb{R}^m$, is known. The set of hierarchically nested competing parameter vectors are presented as

$$\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^i, \dots, \boldsymbol{\theta}^N\}, \quad (2)$$

where $\boldsymbol{\theta}^i \in \mathbb{R}^N$ is restricted in space as

$$\boldsymbol{\theta}^i = [\theta_1, \theta_2, \dots, \theta_i, 0, \dots, 0]^T. \quad (3)$$

The elements of the random noise vector $\boldsymbol{\epsilon} \in \mathbb{R}^m$ are independent and identically distributed (i.i.d.), having a Gaussian distribution with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\sigma \geq 0$ is the unknown standard deviation which is treated as a nuisance parameter. Under aforementioned circumstances the probability density function $p(\mathbf{y}; \boldsymbol{\theta}^i)$ is known apart from $\boldsymbol{\theta}^i \triangleq [\sigma^2, \boldsymbol{\theta}^i]^T$.

The interest of the multiple hypothesis testing problem is to determine the true order of the underlying signal. It is well known that BIC selects the $\mathcal{H}_i$ by solving

$$\min_i \left\{ -2 \ln p(\mathbf{y}; \hat{\boldsymbol{\theta}}^i | \mathcal{H}_i) + \ln \det \mathbf{F}(\hat{\boldsymbol{\theta}}^i) \right\} \quad (4)$$

over $N + 1$ restricted parameters [2], [3]. Here, $\hat{\underline{\theta}}^i$ is the maximum likelihood estimate of the unknown parameter vector $\underline{\theta}^i$ under hypothesis $\mathcal{H}_i$, $\det(\cdot)$ denotes the determinant of a matrix and the matrix $\mathbf{F}(\cdot)$ is the Fisher information matrix defined as

$$\mathbf{F}(\underline{\theta}^i) \triangleq -\mathbb{E} \begin{pmatrix} \frac{\partial^2 \ln p(\mathbf{y}; \underline{\theta}^i | \mathcal{H}_i)}{\partial \underline{\theta}^i \partial \underline{\theta}^{iT}} & \frac{\partial^2 \ln p(\mathbf{y}; \underline{\theta}^i | \mathcal{H}_i)}{\partial \underline{\theta}^i \partial \sigma^2} \\ \frac{\partial^2 \ln p(\mathbf{y}; \underline{\theta}^i | \mathcal{H}_i)}{\partial \sigma^2 \partial \underline{\theta}^{iT}} & \frac{\partial^2 \ln p(\mathbf{y}; \underline{\theta}^i | \mathcal{H}_i)}{\partial \sigma^4} \end{pmatrix}. \quad (5)$$

BIC typically concerns data models such as polynomial regression and the auto-regressive model where the parameter spaces are nested as defined in (2) and (3). It has been proven that this criterion is consistent in selecting the true model order as $m \rightarrow \infty$ [3]. However, in settings like subset selection, the competing parameter vectors are rather required to have more general structure than (2). In fact, the nested parameter spaces should be generalized to the combinatorial parameter spaces. This can be presented as the union of all the sparse vectors with the maximum cardinality of K or

$$\Theta = \cup_{j=1}^K \Theta_j, \quad (6)$$

where each subset Θ_j is the set of all j -sparse parameter vectors

$$\Theta_j = \{\underline{\theta} \in \mathbb{R}^N \mid \|\underline{\theta}\|_0 = j\}. \quad (7)$$

Here, $\|\underline{\theta}\|_0$ denotes the cardinality of the support set of $\underline{\theta}$ defined as $\mathcal{I} = \{j \mid \theta_j \neq 0\}$ and in what follows we will use $\underline{\theta}_{\mathcal{I}}$ to represent $\{\theta_j \mid j \in \mathcal{I}\}$. Note that the integer $K \ll m$ denotes an upper bound on the degrees of freedom of the model. Accordingly, under each $\mathcal{H}_{\mathcal{I}}$ we have

$$\mathcal{H}_{\mathcal{I}} : \mathbf{y} = s(\underline{\theta}_{\mathcal{I}}) + \sigma \epsilon,$$

for $|\mathcal{I}| \leq K$, where $|\mathcal{I}|$ denotes the cardinality of $\mathcal{I}$.

To extend the framework of the multiple hypothesis testing problem, from the nested parameter spaces (2) to the combinatorial parameter spaces (6), EBIC has been proposed [6]. In this respect, EBIC adds a binomial coefficient penalty to an approximation of (4), by considering $\det \mathbf{F}(\hat{\underline{\theta}}_{\mathcal{I}}) \approx m^{|\mathcal{I}|+1}$, so that it becomes

$$\min_{\mathcal{I} \in \Theta} \left\{ -2 \ln p(\mathbf{y}; \hat{\underline{\theta}}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}}) + (|\mathcal{I}| + 1) \ln m + 2 \ln \binom{N}{|\mathcal{I}|} \right\}, \quad (8)$$

where $\hat{\underline{\theta}}_{\mathcal{I}}$ is the ML estimate of $\underline{\theta}_{\mathcal{I}} = [\sigma^2, \underline{\theta}_{\mathcal{I}}]^T$. Intuitively, the binomial coefficient penalty in (8) counterbalance the effect of the large dimensional model space. Unfortunately, EBIC suffers from inconsistency as $\sigma \rightarrow 0$. Moreover, while EBIC was proven to be consistent as $m \rightarrow \infty$ [6], the empirical performance for problems of practical size is not always good, as e.g. illustrated in Fig. 1 and Fig. 2 in Section III.

Inspired by [6], we propose a consistent extended form of BIC for the high-dimensional settings $N > m$, for which the

competitive vector parameters are defined as (6). CEBIC is defined as

$$\min_{\mathcal{I} \in \Theta} \left\{ -2 \ln p(\mathbf{y}; \hat{\underline{\theta}}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}}) + \ln \det \mathbf{F}(\hat{\underline{\theta}}_{\mathcal{I}}) + 2c \ln \binom{N}{|\mathcal{I}|} \right\}, \quad (9)$$

for some $c > 1$. We highlight the facts that CEBIC does not approximate $\det \mathbf{F}(\hat{\underline{\theta}}_{\mathcal{I}})$ and c is introduced to improve the performance of CEBIC for finite m . Selecting a suitable value for c will be discussed in the extended version of this paper.

B. CEBIC for Linear Model

Now we assume that $s(\underline{\theta}_{\mathcal{I}}) = \mathbf{A}_{\mathcal{I}} \underline{\theta}_{\mathcal{I}}$, where the matrix $\mathbf{A}_{\mathcal{I}}$ is the collection of columns of the known matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ with the support $\mathcal{I}$. Additionally, it is assumed that $\mathbf{A}_{\mathcal{I}}$ is a full-rank matrix for any $\mathcal{I} \in \Theta$. Then the log-likelihood function for this particular model is given by

$$\ln p(\mathbf{y}; \underline{\theta}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}}) = -\frac{m}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{A}_{\mathcal{I}} \underline{\theta}_{\mathcal{I}}\|_2^2. \quad (10)$$

To evaluate (9) we first need to find the ML estimate of $\underline{\theta}_{\mathcal{I}}$. A straightforward calculation yields

$$\hat{\underline{\theta}}_{\mathcal{I}} \triangleq [\hat{\sigma}^2, \hat{\underline{\theta}}_{\mathcal{I}}]^T = \left[\frac{1}{m} \|\Pi_{\mathcal{I}}^{\perp} \mathbf{y}\|_2^2, \quad \mathbf{A}_{\mathcal{I}}^{\dagger} \mathbf{y} \right]^T, \quad (11)$$

where $\Pi_{\mathcal{I}}^{\perp}$ denotes the projection matrix which is defined as $\Pi_{\mathcal{I}}^{\perp} = \mathbf{I} - \mathbf{A}_{\mathcal{I}} \mathbf{A}_{\mathcal{I}}^{\dagger}$ and the matrix $\mathbf{A}_{\mathcal{I}}^{\dagger} = (\mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}})^{-1} \mathbf{A}_{\mathcal{I}}^T$ is the Moore–Penrose pseudo-inverse of $\mathbf{A}_{\mathcal{I}}$. Next, we take the second derivatives of the log-likelihood function (10) with respect to the corresponding variables:

$$\begin{aligned} \frac{\partial^2 \ln p(\mathbf{y}; \underline{\theta}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}})}{\partial \underline{\theta}_{\mathcal{I}} \partial \underline{\theta}_{\mathcal{I}}^T} &= -\frac{1}{\sigma^2} \mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}}, \\ \frac{\partial^2 \ln p(\mathbf{y}; \underline{\theta}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}})}{\partial \underline{\theta}_{\mathcal{I}} \partial \sigma^2} &= -\frac{1}{\sigma^4} \mathbf{A}_{\mathcal{I}}^T (\mathbf{y} - \mathbf{A}_{\mathcal{I}} \underline{\theta}_{\mathcal{I}}), \\ \frac{\partial^2 \ln p(\mathbf{y}; \underline{\theta}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}})}{\partial \sigma^4} &= \frac{m}{2\sigma^4} - \frac{1}{\sigma^6} \|\mathbf{y} - \mathbf{A}_{\mathcal{I}} \underline{\theta}_{\mathcal{I}}\|_2^2. \end{aligned}$$

To determine the Fisher information we take the expectation of the preceding derivatives with respect to $p(\mathbf{y}; \underline{\theta}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}})$, thus

$$\mathbf{F}(\underline{\theta}_{\mathcal{I}}) = \frac{1}{\sigma^2} \begin{pmatrix} \mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \frac{m}{2\sigma^2} \end{pmatrix}.$$

Replacing σ^2 by $\hat{\sigma}^2$ and exploiting properties of the determinant function yields

$$\begin{aligned} \det \mathbf{F}(\hat{\underline{\theta}}_{\mathcal{I}}) &= \left(\frac{1}{\hat{\sigma}^2} \right)^{|\mathcal{I}|+1} \det \begin{pmatrix} \mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \frac{m}{2\hat{\sigma}^2} \end{pmatrix} \\ &= \frac{m}{2} \left(\frac{1}{\hat{\sigma}^2} \right)^{|\mathcal{I}|+2} \det(\mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}}). \end{aligned}$$

We now insert $\hat{\sigma}^2$ from (11) into $\det \mathbf{F}(\hat{\underline{\theta}}_{\mathcal{I}})$, so that

$$\det \mathbf{F}(\hat{\underline{\theta}}_{\mathcal{I}}) = \frac{m}{2} \left(\frac{m}{\|\Pi_{\mathcal{I}}^{\perp} \mathbf{y}\|_2^2} \right)^{|\mathcal{I}|+2} \det(\mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}}).$$

Lastly we insert (11) into (10) and then combine the result with $\ln \det \mathbf{F}(\hat{\boldsymbol{\theta}}_{\mathcal{I}})$. A straightforward calculation eventually provides CEBIC for the linear model as

$$\min_{\mathcal{I} \in \Theta} \left\{ (m - |\mathcal{I}| - 2) \ln \left\| \Pi_{\mathcal{I}}^{\perp} \mathbf{y} \right\|_2^2 + \ln \det(\mathbf{A}_{\mathcal{I}}^T \mathbf{A}_{\mathcal{I}}) + (|\mathcal{I}| + 2) \ln m + 2c \ln \binom{N}{|\mathcal{I}|} \right\}. \quad (12)$$

III. NUMERICAL RESULTS

In this section, numerical results are provided to assess the performance of CEBIC for the linear model. The probability that (12) correctly selects the true model is investigated in three different setups. In each setup, this probability is measured versus a varying factor, namely, *number of measurements*, *standard deviation of the noise* and *cardinality*. The numerical results of CEBIC are also compared with those of some existing information criteria, EBIC and combined BIC (COBIC). To extend the framework of COBIC, the original formulation in [4] is extended with the binomial coefficient penalty, similar to EBIC in [6], so that it becomes

$$\min_{\mathcal{I} \in \Theta} \left\{ m \ln \hat{\sigma}^2 + h + 2 \ln \binom{N}{|\mathcal{I}|} \right\},$$

where h is defined as

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

The reader should note that solving the combinatorial minimization (12) involves an exhaustive search through Θ . It is also known that the size of Θ increases exponentially with m . Consequently, solving the minimization (12) through an exhaustive search even for relatively small dimensions might be infeasible, let alone for the high-dimensional settings. To overcome this issue, we exploit some estimator with ℓ_1 -constraint in order to get the set of competing models. A viable ℓ_1 estimator is the LASSO defined as

$$\hat{\boldsymbol{\theta}}_{\mathcal{L}} = \arg \min_{\boldsymbol{\theta}} \left\{ \lambda \|\boldsymbol{\theta}\|_1 + \frac{1}{2} \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|_2^2 \right\}, \quad (13)$$

where $\lambda \in [0, \|\mathbf{A}^T \mathbf{y}\|_{\infty}]$ is a hyper-parameter that indirectly controls the support of $\hat{\boldsymbol{\theta}}_{\mathcal{L}}$, $\mathcal{I} = \text{supp}(\hat{\boldsymbol{\theta}}_{\mathcal{L}})$. By this we mean, setting $\lambda = \|\mathbf{A}^T \mathbf{y}\|_{\infty}$ in the LASSO results in $\mathcal{I} = \emptyset$ and as λ decreases, $\mathcal{I}$ gradually evolves (not strictly, but mostly such that $|\mathcal{I}|$ gradually increases) [10], [11]. Thus, there are pivotal λ 's for which $\mathcal{I}$ changes accordingly. Therefore, for each pivotal λ , we plug in the corresponding $\mathcal{I}$ into the objective function of the information criterion, e.g. (12).

The general setting for the numerical experiments is as follows. The columns of $\mathbf{A}$ are the instances of i.i.d. Gaussian random vectors $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The dimension of the problem is a function of the number of measurements as $\lceil N = m^d \rceil$ with $d = 1.3$, where $\lceil \cdot \rceil$ is the ceiling function. The true parameter $\boldsymbol{\theta}^*$ is specified in each Monte Carlo trial by first choosing the support set $\mathcal{S}$ with cardinality equal to five randomly. Then $\boldsymbol{\theta}_{\mathcal{S}}^*$ is drawn from the binomial distribution of the sample space of $\{1, -1\}$. The additive perturbation is a white Gaussian noise with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Fig. 1. The probability of correct model selection versus m when $\sigma^2 = 10^{-0.3}$, $\|\boldsymbol{\theta}^*\|_0 = 5$ and $N = \lceil m^d \rceil$ for $d = 1.3$.

First, we provide a numerical simulation to address the dependency of the problem of correct model selection to the number of measurements. In this simulation, the noise variance is fixed as $\sigma^2 = 10^{-0.3}$ and $c = 2$. Fig. 1 shows the estimated probability of correct model selection versus m . For each m , this probability is estimated over 200 Monte Carlo trials. The numerical result shows that CEBIC achieves $\Pr(\mathcal{I} = \mathcal{S}) = 1$, indicating that CEBIC is consistent in selecting the correct model as m increases. The result also shows that EBIC fails to achieve the same probability value for this range of m values. For the parameter settings of this example, COBIC is essentially equivalent to EBIC. We emphasize that using an information criterion, e.g. (12), in combination with the LASSO twists the performance of model selection with the LASSO's ability to identify $\mathcal{S}$. Clearly, for successful model selection, we need that the LASSO regularization path includes the true model structure as well as that the information criterion attains its minimum at that point. Hence, the relatively poor performance for small m may be due to the failure of either the LASSO or the information criterion.

Next, we consider the effect of σ on the performance of the information criteria. Fig. 2 illustrates the estimated probability of correct model selection versus $\ln 1/\sigma^2$ over 300 Monte Carlo trials when m is fixed to 120 and $c = 1.5$. The probability of $\mathcal{I} = \mathcal{S}$ for CEBIC converges to one quickly, whereas this probability for EBIC stays below 0.95. Hence, the numerical result indicates that EBIC is inconsistent in finding $\mathcal{S}$ as $\sigma \rightarrow 0$. Fig. 2 also shows that in comparison to CEBIC, COBIC achieves $\Pr(\mathcal{I} = \mathcal{S}) = 1$ at a much smaller σ . All information criteria perform poorly for large σ . This can be linked to the poor performance of LASSO's model selection in presence of strong noise.

Lastly, we let the cardinality of the true support vary while the other factors are fixed. Fig. 3 illustrates $\Pr(\mathcal{I} = \mathcal{S})$ versus $|\mathcal{S}|$ when m is fixed to 120, $\sigma^2 = 10^{-1}$ and $c = 2$. This probability is estimated over 300 Monte Carlo trials. It is,

Fig. 2. The probability of correct model selection versus $\ln 1/\sigma^2$ when $m = 120$, $\|\theta^*\|_0 = 5$ and $N = \lceil m^d \rceil$, for $d = 1.3$.

again because of the LASSO's properties, expected that the probability of successful model selection decreases as $|\mathcal{S}|$ increases. This is confirmed by Fig. 3.

Fig. 3. The probability of correct model selection versus $|\mathcal{S}|$ when $\sigma^2 = 10^{-1}$, $m = 120$ and $N = \lceil m^d \rceil$, for $d = 1.3$.

We end this section by providing a numerical realization of the objective functions related to the information criteria. Fig. 4 illustrates the behavior of different information criteria versus the choice of $\mathcal{I}$ by the LASSO, i.e. $\mathcal{I} = \text{supp}(\hat{\theta}_{\mathcal{L}})$. Here, m is fixed to 100, $\sigma^2 = 10^{-1}$, $\|\theta^*\|_0 = 5$ and $c = 2$. The waterfall behavior of the numerical realization of the objective functions shows that the LASSO has chosen the model correctly for a pivotal λ . Fig. 4 shows that despite the fact that the LASSO's solution-path includes the true model structure, EBIC fails to attain its minimum at that point. Fig. 4 also shows that BIC behaves rather liberal, which results in overfitting.

Fig. 4. The behavior of different information criteria versus the sequence of $\mathcal{I}$'s. The support $\mathcal{I}$ is chosen by the LASSO. The setting is $\sigma^2 = 10^{-1}$, $\|\theta^*\|_0 = 5$, $m = 100$ and $N = \lceil m^d \rceil$, for $d = 1.3$. The small box on the corner magnifies CEBIC and EBIC curves just after the waterfall behavior. Label six corresponds to the true model structure.

IV. CONCLUSION

Model selection is a crucial task in statistical signal processing. Accomplishing this task in a high-dimensional setting is challenging. Recently, researchers have shown an increasing interest in the problem of model selection in high-dimensional settings. In this study we proposed a new information criterion, CEBIC, which benefits from the desirable property of consistency. The numerical results support the claimed consistency of CEBIC either as $m \rightarrow \infty$ or as $\sigma \rightarrow 0$.

REFERENCES

- [1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.
- [2] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.
- [3] 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.
- [4] 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.
- [5] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1998.
- [6] 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.
- [7] A. Owring and M. Jansson, "Hyper-parameter selection for lasso," *to be submitted to IEEE Trans. Signal Process.*
- [8] E. Candes and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Ann. Statist.*, vol. 37, no. 5A, pp. 2145–2177, Oct. 2009.
- [9] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
- [10] J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3601–3608, Oct. 2005.
- [11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, Apr. 2004.