IEEE Open Journal of Signal Processing

Received 2 June 2022; revised 26 September 2022; accepted 4 October 2022. Date of publication 10 October 2022; date of current version 8 November 2022. The review of this article was arranged by Associate Editor Usman A Khan.

Digital Object Identifier 10.1109/OJSP.2022.3213128

False Discovery Rate (FDR) and Familywise Error Rate (FER) Rules for Model Selection in Signal Processing Applications

PETRE STOICA [1] AND PRABHU BABU [2]

[1]Division of Systems and Control, Department of Information Technology, Uppsala University, 75237 Uppsala, Sweden
[2]Centre for Applied Research in Electronics, Indian Institute of Technology Delhi, New Delhi, Delhi 110016, India

CORRESPONDING AUTHOR: PRABHU BABU (e-mail: prabhubabu@care.iitd.ac.in)

The work of Petre Stoica was supported by Swedish Research Council (VR) under Grants 2017-04610, 2016-06079, and 2021-05022.

ABSTRACT Model selection is an omnipresent problem in signal processing applications. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are the most commonly used solutions to this problem. These criteria have been found to have satisfactory performance in many cases and had a dominant role in the model selection literature since their introduction several decades ago, despite numerous attempts to dethrone them. Model selection can be viewed as a multiple hypothesis testing problem. This simple observation makes it possible to use for model selection a number of powerful hypothesis testing procedures that control the false discovery rate (FDR) or the familywise error rate (FER). This is precisely what we do in this paper in which we follow the lead of the proposers of the said procedures and introduce two *general* rules for model selection based on FDR and FER, respectively. We show in a numerical performance study that the FDR and FER rules are serious competitors of AIC and BIC with significant performance gains in more demanding cases, essentially at the same computational effort.

INDEX TERMS Model order selection, FDR, FER, AIC, BIC.

I. INTRODUCTION

Model selection is an essential problem in many signal processing applications [1], [2], [3]. Examples of such applications include selecting the order of an autoregressive predictor, the number of source signals impinging on an array of sensors, the order of a polynomial trend, the number of components of a nuclear magnetic resonance signal, the dimension of a linear regression model, the length and paths of the impulse response of a multi-path communication channel, the number of components of a sinusoidal signal, and the rank of the solution of a matrix approximation problem (the last four applications form the nucleus of the numerical example section, and will be described in detail there).

A successful class of rules for model selection is based on penalizing the model complexity (basically its number of parameters, let us say k) by adding a penalty term to the negative log-likelihood [1], [3]:

$$-2\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k) + ck \tag{1}$$

where $\mathbf{y}$ is the data vector and $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimate for the model with k parameters. Different values of c are obtained from different types of statistical or information theory considerations. For example, $c = 2$ for the Akaike information criterion (AIC) [4], and $c = \ln N$ for the Bayesian information criterion (BIC) [5]. The AIC is minimax optimal in a predictive sense, and the BIC is consistent in a model selection sense, but as alluded by a discussion in [6] (see also [7]) neither is optimal (in the sense of maximizing the probability of selection) and therefore, in principle, they both can be outperformed by other rules.

The main goal of the present paper is to propose model selection rules that have the well-established penalized likelihood form in (1), like AIC and BIC, but which perform better than both BIC and AIC especially in demanding cases (such as for high-dimensional models). The derivation of the new rules is based on formulating the model selection as a multi-hypothesis testing problem and using FDR and FER procedures to solve it [8], [9], [10].

TABLE 1. The Number of Times H^0 was Correctly Accepted/Incorrectly Accepted/Correctly Rejected/Incorrectly Rejected is Denoted *CA/IA/CR/IR*, Respectively. M is the Total (Known) Number of Null Hypotheses and $\tilde{M}$ Denotes the (Unknown) Number of True Null Hypotheses. An *IR* Decision is a False Discovery/Alarm, *IA* is a Miss, and the Other Two are Correct Detections

Ground Truth \ Decision	H^0 = accepted	H^0 = rejected	Total
H^0 = true	CA (detection)	IR (false discovery/alarm)	$\tilde{M}$
H^0 = false	IA (miss)	CR (detection)	$M-\tilde{M}$
Total	$M-R$	R	M

Multiple hypothesis testing based on FDR and FER has found numerous applications in many fields from genomics and biomedicine to financial economics. However the use of FDR or FER for model selection in signal processing applications is almost nonexistent. We begin this paper with a short introduction of the basic FDR and FER principles [8], [11]. Then we explain how FDR and FER can be used for selecting the structure (or order) of both linear and nonlinear models with both sparse and dense parameter vectors. We end-up the paper with a numerical performance study of the FDR and FER rules, as well as a comparison with the classical rules of AIC (Akaike Information Criterion) [2], [4] and BIC (Bayesian Information Criterion) [2], [5] in the four signal processing applications mentioned above: linear regression, communication channel estimation, sinusoidal parameter estimation, and low-rank matrix approximation.

The following sections provide details, including further references, on the technical aspects of the plan laid down above. The main contribution of this paper is the proposal of FDR and FER rules based on *general* penalized-likelihood criteria similar to AIC and BIC and therefore directly usable in the many signal processing applications that need model selection. The main pragmatical finding of the paper is that the FDR and FER rules perform much better than AIC and BIC in applications in which the models under consideration have a wide range of possible orders, and not worse in other cases in which the classical rules are known to be consistent (BIC) or possess an optimal minimax prediction property (AIC). Given the good performance of the FDR and FER criteria and the fact that they have the same well-established penalized-likelihood form as AIC and BIC, it is our opinion that the former can be considered to be serious competitors of the latter.

II. FALSE DISCOVERY RATE (FDR)

Consider a set of M null hypotheses : $H_1^0, \ldots, H_M^0$. Let $T_k > 0$ be the test statistic for H_k^0, and let p_k denote the false discovery probability:

$$p_k = \text{prob}(\text{reject } H_k^0 | H_k^0 = \text{true}) \tag{2}$$

(p_k is also called the probability of false alarm or false positive, or simply the significance level). Table 1 summarizes the four possible outcomes when testing one of the above null hypotheses (denoted H^0 in the table).

The FDR is defined as the expected proportion of IR out of the total R:

$$\text{FDR} = \mathbb{E}[IR/R] \tag{3}$$

(FDR $= 0$ if $R = 0$). In many practical applications, for instance in genomics, it is important to keep the ratio IR/R under a pre-specified level, to avoid wasting lab time on investigating an unnecessarily large number of "false discoveries". Intuitively, keeping this ratio small is also important for model selection to prevent selecting unduly complex models too often. We will explain the way in which FDR can be used for model selection in the next sections. In the rest of this section we discuss a procedure for controlling the FDR in the sense that [8]:

$$\text{FDR} \leq \frac{\tilde{M}}{M}\alpha \leq \alpha \tag{4}$$

where α is a pre-specified level. Following the recommendation in [9], [10], and to simplify the exposition, we will use $\alpha = 0.01$ in what follows (but note that the choice of α in (4) is an interesting aspect, see e.g [10]).

Assume that the distribution of T_k under H_k^0 is known. Also, assume that the hypotheses $H_k^0 (k = 1, \ldots, M)$ have been *ordered* [8], [9], [10], [12] so that

$$T_1 \geq T_2 \geq \cdots \geq T_M \tag{5}$$

Let,

$$p_k = \alpha \frac{k}{M \eta_M}, \tag{6}$$

where η_M is the so-called harmonic number

$$\eta_M = 1 + \frac{1}{2} + \cdots + \frac{1}{M} \ \left(\approx \ln M + 0.577 \text{ for large } M\right). \tag{7}$$

Also let q_{p_k} be the following quantile of the distribution of T_k:

$$q_{p_k} : \text{prob}(T_k \geq q_{p_k} | H_k^0) = p_k \tag{8}$$

Finally, let

$$\hat{k} = \max\,[k : T_k \geq q_{p_k}]. \tag{9}$$

Then reject the hypotheses $H_1^0, \ldots, H_{\hat{k}}^0$ and accept $H_{\hat{k}+1}^0, \ldots, H_M^0$, or reject no hypothesis if there is no k that satisfies the inequality in (9). The above procedure ensures that (4) is satisfied under *general conditions* on the statistics $\{T_k\}$ (see, e.g., [13]). If the test statistics can be assumed to be *independent or positively correlated* then the following larger significance levels,

$$p_k = \frac{\alpha k}{M}, \tag{10}$$

can be shown to ensure the control of FDR as in (4) [8], [13]. A compact self-contained proof of this result for independent statistics is presented in the Appendix. Note that if we used (6) in the case of independent statistics $\{T_k\}$ then FDR would be controlled at level $\frac{\alpha}{\eta_m}$.

Remark 1: Equivalently $\hat{k}$ can be defined as the maximum value of k for which the p-value corresponding to the observed T_k is less than the significance level p_k. In this paper we prefer to directly compare $\{T_k\}$ to the quantiles $\{q_k\}$ as in (9).

III. MODEL SELECTION USING FDR

We will separately consider two types of models that are frequently encountered in signal processing applications: linear-in-the parameters models and nonlinear-in-the parameters models.

A. LINEAR MODELS

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ be a set of $N \times 1$ known vectors, and let the observed data vector $\mathbf{y} \in \mathbb{R}^N$ be given by:

$$\mathbf{y} = \sum_{j=1}^{M} c_j \mathbf{x}_j + \mathbf{e}, \tag{11}$$

where an unknown number of the coefficients $\{c_j\}$ are different from zero, and the noise $\mathbf{e}$ is normally distributed with zero mean and covariance matrix $\sigma^2\mathbf{I}$: $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The problem is to decide which of the M possible regressors $\{\mathbf{x}_j\}$ in (11) have really contributed to the data vector $\mathbf{y}$, or equivalently which of the parameters $\{c_j\}$ are zero and which are different from zero. The maximum likelihood estimates of the unknown parameters in (11) are given by (see, e.g., chap 4 in [1]):

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_M \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{12}$$

$$\hat{\sigma}^2 = \frac{1}{N}||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}||^2 \tag{13}$$

where

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M] \tag{14}$$

(assuming that the inverse in (12) exists). Moreover, the covariance matrix of the estimation errors in $\hat{\boldsymbol{\theta}}$ has the following simple expression:

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{15}$$

It is well known (see, e.g., [1], [7]) that the statistics :

$$T_k = \frac{|\hat{c}_k|^2}{\hat{\sigma}^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{kk}} \tag{16}$$

are asymptotically chi-square distributed with one degree of freedom:

$$T_k \sim \chi^2(1) \quad k = 1, \ldots, M \quad (\text{Under } H_k^0) \tag{17}$$

To use the above $\{T_k\}$ as test statistics in the procedure of Section II we must order them so that (see (5)):

$$T_1 \geq T_2 \geq \cdots \geq T_M \tag{18}$$

The thresholds $\{q_{p_k}\}$ in (8) can be obtained from a table/calculator of the $\chi^2(1)$ distribution, and then we can use (9) to find the regressors that should be included in the model.

Remark 2: Alternatively we can get the quantiles $\{q_{p_k}\}$, for given $\{p_k\}$, from a table (or calculator) of the standard normal distribution. Indeed $t \sim \chi^2(1)$ if $t = z^2$ with $z \sim \mathcal{N}(0, 1)$. Consequently,

$$\mathrm{prob}(t \geq q_p) = \mathrm{prob}(z \geq q_p^{1/2} \cup z \leq -q_p^{1/2}) = 2\phi(-q_p^{1/2}) \tag{19}$$

where $\phi(z)$ is the cumulative distribution function of $\mathcal{N}(0, 1)$. Hence, we can get γ_p that satisfies

$$\mathrm{prob}(z \geq \gamma_p) = \frac{p}{2} \tag{20}$$

from a table of $\mathcal{N}(0, 1)$ and use $q_p = \gamma_p^2$ in (9).

We can relate the above FDR rule to the classical AIC and BIC rules. To do so we consider the following hypotheses:

H_k : the model (11) contains exactly k regressors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$

for $k = 0, \ldots, M$ (with H_0: $\mathbf{y} = \mathbf{e}$). Note that, because the regressors are ordered according to (18), the hypotheses H_k are *nested* (i.e., under H_k the regressors are $\mathbf{x}_1, \ldots, \mathbf{x}_k$, the first $(k-1)$ of which are the regressors under H_{k-1}). Also note that the ordering of the regressors is a departure from the traditional way in which AIC and BIC are applied to (11). Finally we note that, in the terminology of the previous section, H_k is equivalent to $\{H_1^0, \ldots, H_k^0\}$ being false and $\{H_{k+1}^0, \ldots, H_M^0\}$ being true.

Under H_k the negative log-likelihood function of $\mathbf{y}$ in (11) is given by :

$$-\ln p\left(\mathbf{y}|\boldsymbol{\theta}_k, \sigma_k^2\right) = \text{constant} + \frac{N}{2}\ln(\sigma_k^2) + \frac{1}{2\sigma_k^2}||\mathbf{y} - \mathbf{X}_k\boldsymbol{\theta}_k||^2, \tag{21}$$

where

$$\mathbf{X}_k = [\mathbf{x}_1, \ldots, \mathbf{x}_k] \text{ for } k = 1, \ldots, M \ (\mathbf{X}_0 = \mathbf{0}) \tag{22}$$

$$\boldsymbol{\theta}_k = [c_1, \ldots, c_k]^T \tag{23}$$

The above function is minimized by (compare with (12) and (13)):

$$\hat{\boldsymbol{\theta}}_k = \left(\mathbf{X}_k^T\mathbf{X}_k\right)^{-1}\mathbf{X}_k^T\mathbf{y} \tag{24}$$

$$\hat{\sigma}_k^2 = \frac{||\mathbf{y} - \mathbf{X}_k\hat{\boldsymbol{\theta}}_k||^2}{N} \tag{25}$$

and the minimum value is (to within a constant) :

$$-2\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2) = N\ln\hat{\sigma}_k^2 \tag{26}$$

The *likelihood ratio* is defined as follows:

$$\begin{aligned} T_k &= -2\ln\frac{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{k-1}, \hat{\sigma}_{k-1}^2)}{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2)} \\ &= 2\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2) - 2\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{k-1}, \hat{\sigma}_{k-1}^2) \end{aligned} \tag{27}$$

Inserting (26) in (27) yields the following simple expression for T_k:

$$T_k = N \ln\left(\frac{\hat{\sigma}_{k-1}^2}{\hat{\sigma}_k^2}\right) \text{ for } (k = 1, \ldots, M) \tag{28}$$

We have used the same notation for the right-hand sides of (16) and (28) because these two quantities are asymptotically equivalent (see Remark 3 below). Consequently, under H_{k-1} the T_k in (28), similarly to (16), also has an asymptotic chi-square distribution with one degree of freedom. In fact it is well known that the likelihood ratio has an asymptotic χ^2 distribution, viz.

$$-2 \ln \frac{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{k-1}, \hat{\sigma}_{k-1}^2)}{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2)} \sim \chi^2(\dim \boldsymbol{\theta}_k - \dim \boldsymbol{\theta}_{k-1}) \quad (\text{under } H_{k-1} \subset H_k) \tag{29}$$

not only for the linear model in (11) but under much more general conditions (see, e.g., [14] and chapter 11 in [1]). We will make use of this result in the next sub-section.

Remark 3: It is well known and easy to show that if σ^2 is given and the regressors are orthogonal then the equivalence between (16) and (28) holds in finite samples. This equivalence also holds even if the regressors are not orthogonal but the proof is a bit more complicated. First, observe that:

$$\begin{aligned}
&2 \ln p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \sigma^2\right) - 2 \ln p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_{k-1}, \sigma^2\right) \\
&= \left(\left\|\mathbf{y} - \mathbf{X}_{k-1}\hat{\boldsymbol{\theta}}_{k-1}\right\|^2 - \left\|\mathbf{y} - \mathbf{X}_k\hat{\boldsymbol{\theta}}_k\right\|^2\right)/\sigma^2 \\
&= \mathbf{y}^T\left(\mathbf{I} - \mathbf{X}_{k-1}\left(\mathbf{X}_{k-1}^T\mathbf{X}_{k-1}\right)^{-1}\mathbf{X}_{k-1}^T\right)\mathbf{y}/\sigma^2 \\
&\quad - \mathbf{y}^T\left(\mathbf{I} - \mathbf{X}_k\left(\mathbf{X}_k^T\mathbf{X}_k\right)^{-1}\mathbf{X}_k^T\right)\mathbf{y}/\sigma^2 \\
&= \mathbf{r}_k^T\left\{\mathbf{R}_k^{-1} - \begin{bmatrix}\mathbf{I} \\ \mathbf{0}\end{bmatrix}\left([\mathbf{I} \quad \mathbf{0}]\,\mathbf{R}_k\begin{bmatrix}\mathbf{I} \\ \mathbf{0}\end{bmatrix}\right)^{-1}[\mathbf{I} \quad \mathbf{0}]\right\}\mathbf{r}_k/\sigma^2,
\end{aligned} \tag{30}$$

where

$$\mathbf{r}_k = \mathbf{X}_k^T\mathbf{y} \tag{31}$$

$$\mathbf{R}_k = \mathbf{X}_k^T\mathbf{X}_k \tag{32}$$

and we used the fact that

$$\mathbf{X}_{k-1} = \mathbf{X}_k\begin{bmatrix}\mathbf{I} \\ \mathbf{0}\end{bmatrix}. \tag{33}$$

Next note that the numerator in (30) can be rewritten as:

$$\mathbf{r}_k^T\mathbf{R}_k^{-1/2}\left\{\mathbf{I}_k - \mathbf{R}_k^{1/2}\begin{bmatrix}\mathbf{I} \\ \mathbf{0}\end{bmatrix}\left([\mathbf{I} \quad \mathbf{0}] \times \mathbf{R}_k\begin{bmatrix}\mathbf{I} \\ \mathbf{0}\end{bmatrix}\right)^{-1}[\mathbf{I} \quad \mathbf{0}]\,\mathbf{R}_k^{1/2}\right\}\mathbf{R}_k^{-1/2}\mathbf{r}_k. \tag{34}$$

The matrix between curly brackets in (34) is the orthogonal projector onto the null space of the $(k-1) \times k$ matrix $[\mathbf{I} \quad \mathbf{0}]\mathbf{R}_k^{1/2}$. Because

$$[\mathbf{I} \quad \mathbf{0}]\,\mathbf{R}_k^{1/2}\mathbf{R}_k^{-1/2}\mathbf{u} = \mathbf{0} \quad \left(\mathbf{u} = \begin{bmatrix}\mathbf{0} \\ 1\end{bmatrix}\right), \tag{35}$$

it follows that the said null space is spanned by the vector $\mathbf{R}_k^{-1/2}\mathbf{u}$. Consequently (34), and hence (30), can be written as:

$$\mathbf{r}_k^T\mathbf{R}_k^{-1}\mathbf{u}\mathbf{u}^T\mathbf{R}_k^{-1}\mathbf{r}_k/\sigma^2\mathbf{u}^T\mathbf{R}_k^{-1}\mathbf{u} = |\hat{c}_k|^2/\sigma^2\left[\mathbf{R}_k^{-1}\right]_{kk} \tag{36}$$

which coincides with (16) (for known σ^2).

To proceed with the discussion on the connection between FDR and penalized likelihood criteria (such as AIC and BIC), observe that the inequality $T_k \geq q_{p_k}$ (see (9) in the description of FDR) is equivalent to (see (27)):

$$2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2) - 2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{k-1}, \hat{\sigma}_{k-1}^2) \geq q_{p_k} \tag{37}$$

which in turn holds if and only if

$$C_k^{\text{FDR}} \leq C_{k-1}^{\text{FDR}}$$

where

$$C_k^{\text{FDR}} = -2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2) + \sum_{j=1}^{k} q_{p_j} \tag{38}$$

The conclusion is that evaluating (38) for $k = 1, 2, \ldots$ and selecting the $\hat{k}$ at which the last minimum of the criterion occurs is equivalent to the FDR procedure based on the likelihood ratio (and also asymptotically equivalent to the FDR based on the statistics in (16)). To relate (38) to AIC and BIC (see e.g. [2], [4] and [5] for details on these two rules) it only remains to observe that for $q_p = 2$, we obtain

$$C_k^{\text{AIC}} = -2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2) + 2k \tag{39}$$

and for $q_p = \ln N$, we get

$$C_k^{\text{BIC}} = -2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2) + k \ln N \tag{40}$$

As we will show later, the above AIC and BIC rules applied to (11) after ordering the regressors (which can be viewed as modifications of the classical rules using FDR ideas, see e.g. [10] and references therein) perform reasonably well whenever their basic assumptions hold. The main advantage of this modified way of using AIC and BIC is computational. The traditional use of AIC and BIC requires testing all 2^M possible combinations of zero and non-zero coefficients in (11), which quickly would become prohibitive as M increases.

B. NONLINEAR MODELS

A common type of model in signal processing applications has the same basic form as (11) but with the essential difference that the "regressors" $\{\mathbf{x}_j\}$ depend on *unknown* parameters.

This means that, under H_k, the model is given by :

$$\mathbf{y} = \sum_{j=1}^{k} c_j \mathbf{x}_j(\mathbf{b}_j) + \mathbf{e} \tag{41}$$

where both $\{c_j\}$ and $\{\mathbf{b}_j\}$ (as well as σ^2) are unknown. Therefore, in the present case the parameter vector of the "signal" term in (41) comprises both the linear parameters $\{c_j\}$ and the nonlinear ones $\{\mathbf{b}_j\}$:

$$\boldsymbol{\theta}_k = [c_1, \ldots, c_k, \mathbf{b}_1^T, \ldots, \mathbf{b}_k^T]^T \tag{42}$$

The likelihood ratio property in (29) is valid in this general case as well, and thus we can use FDR either in the hypothesis testing form, (9), or in the model selection criterion form (38), to obtain an estimate $\hat{k}$ of the integer-valued parameter of (41). An important difference from the procedure described in the previous section is that in the present case we do not need to do anything to order the test statistics $\{T_k\}$ as in (18). Indeed, for (41) we can expect that an analog ranking will hold in most cases without any intervention by the user. This means that in the case of (41) both AIC and BIC as well as FDR can be used without any concern about ordering the test statistics. To explain why this is so, note that for (41) we estimate not only the coefficients $\{c_j\}$ by minimizing the fitting criterion but also the parameters $\{\mathbf{b}_j\}$ that define the regressors. Therefore, for example for $k = 1$, we determine both the optimal c_1 and the optimal regressor $\mathbf{x}_1$. The greedy procedure described in the previous section chooses (for $k = 1$) the regressor for which T_1 in (16) is larger than $T_2, T_3,$ etc. However, this is nothing but a computationally convenient surrogate for the complete procedure which would fit the regressors one by one to $\mathbf{y}$ in order to find the best regressor (out of all possible regressors) for the one-regressor ($k = 1$) model.

An important aspect concerning the application of FDR (or FER, see the next section) to nonlinear models described by (41) is the choice of M: how many null hypotheses $\{H_k^0\}$ do we (implicitly) test when taking a decision about the possible inclusion in (41) of an additional regressor? In the real-valued parameter case, which we consider in this paper, the set of possible regressors spanned by the nonlinear parameters in (41) is a manifold, and hence M is theoretically infinite. However for any given regressor vector there are many regressors in its vicinity that are only infinitesimally different from it and thus can be neglected. Consequently, from a practical standpoint, we can "sample" the said manifold using a finite number M of regressors that cover it well. The selection of these regressors, and therefore of M, is a problem whose solution is *application dependent* (see Section IV for details and examples).

IV. MODEL SELECTION USING FER

Consider, once again, the hypotheses $\{H_k^0\}_{k=1}^M$, which have been ordered according to (5). The so-called familywise error rate is defined as (see Table 1):

$$\text{FER} = \text{prob}(IR \geq 1) \tag{43}$$

Bonferroni rule is the oldest and simplest procedure for controlling the FER. It uses the following significance levels:

$$p_k = \alpha/M \tag{44}$$

and it guarantees that

$$\text{FER} \leq \alpha \tag{45}$$

Because the significance levels in (44) do not depend on k, the Bonferroni rule does not require the hypotheses to be ordered. However, it turns out that in many cases these significance levels are "too small" and hence the Bonferroni rule is too conservative (i.e., it has a high probability of miss).

A more powerful rule, that controls the FER as in (45), uses the following sequence of significance levels [11]:

$$p_k = \alpha/(M + 1 - k). \tag{46}$$

A simple proof of the fact that the inequality in (45) is satisfied runs as follows. Let

$$C = \text{the sub-set of true null hypotheses, hence } |C| = \tilde{M} \tag{47}$$

$$H_h^0 = \text{the first incorrectly rejected null hypothesis in the sequence } H_1^0, H_2^0, \ldots H_M^0. \tag{48}$$

Then it must hold that:

$$T_h \geq q_{\alpha/(M+1-h)} \tag{49}$$

and

$$H_1^0, \ldots, H_{h-1}^0 = \text{correctly rejected, hence } h - 1 \leq M - \tilde{M} \tag{50}$$

It follows that the total false alarm probability of the rule based on (46) satisfies:

$$\begin{aligned} \text{prob}\left(\bigcup_{h\in C} T_h \geq q_{\alpha/(M+1-h)}\right) &\leq \sum_{h\in C} \text{prob}\left(T_h \geq q_{\alpha/(M+1-h)}\right) \\ &= \sum_{h\in C} \frac{\alpha}{M+1-h} \leq \sum_{h\in C} \frac{\alpha}{\tilde{M}} = \alpha, \end{aligned} \tag{51}$$

where the first inequality follows from Boole's union bound and the second from (50) (the equality follows from the definition of the quantiles). With this observation, the proof of (45) is concluded.

Note that, similarly to the control of FDR using (6), the control of FER as in (45) does not require any assumption on the statistics $\{T_k\}$. However, like for FDR, the upper bound in (45) can be loose if $\{T_k\}$ are highly correlated. Also note that, the expression for the significance levels in (46) is quite intuitive: let us say that we have rejected $H_1^0, \ldots, H_{k-1}^0$ and are currently considering H_k^0, then there are $M - k + 1$ hypotheses left to test and consequently we can change the denominator M in the Bonferroni significance levels (44) to $M + 1 - k$.

We can directly use (46) to obtain a model selection criterion analog to the FDR criterion in (28), which we will call

FER for short:

$$C_k^{\text{FER}} = -2\ln p\left(\mathbf{y}|\hat{\theta}_k, \hat{\sigma}_k^2\right) + \sum_{j=1}^{k} q_{p_j} \tag{52}$$

(like for FDR, we choose $\alpha = 0.01$ in (46)).

Next we remark on a difference between FER and FDR, which is quite relevant to model selection. Let C be defined as in (47), and let:

$\bar{C}$ = the subset of false null hypotheses, $\left|\bar{C}\right| = M - \tilde{M} \triangleq k_0$

P_{FA} = probability of false alarm (the probability of rejecting all hypotheses in $\bar{C}$ and at least one in C)

P_{D} = probability of detection (the probability of rejecting all hypotheses in $\bar{C}$ and nothing else)

We can write P_{FA} as

$$\text{P}_{\text{FA}} = \text{prob}(\text{rejecting all } H_k^0 \in \bar{C} | IR \geq 1) \underbrace{\text{prob}(IR \geq 1)}_{\text{FER}}$$

$$\leq \text{FER} \tag{53}$$

Because for a sound selection procedure P_{D} is close to one, it follows that in such a case

$$\text{P}_{\text{FA}} \approx \text{FER} \tag{54}$$

Consequently α can be expected to be a tight upper bound on the P_{FA} of a sound rule that controls FER at level α. Interestingly such a rule also controls the FDR at level α because FDR$\leq$ FER:

$$\text{FDR} = \mathbb{E}(IR/R) \leq \mathbb{E}(\text{I}(IR \geq 1)) = \text{prob}(IR \geq 1) = \text{FER} \tag{55}$$

On the other hand, the converse is not necessarily true: a procedure that controls FDR at level α may have a substantially larger FER (and hence P_{FA}) than α. Intuitively this is so since, out of the total number of R rejections in a data realization, an FDR control at level α allows on the average $IR = \alpha R$ of them to be incorrect; and αR incorrect rejections per data realization yield a P_{FA} significantly larger than α. We can use Markov's inequality to lend some support to the above intuitive argument:

$$\text{FDR} \geq \lambda \text{prob}(IR/R \geq \lambda) \text{ for any } \lambda > 0. \tag{56}$$

Note that we cannot use (56) to upper bound FER by FDR$\leq \alpha$ because FER corresponds to $\lambda \to 0$ in (56), in which case the bound in (56) becomes infinite and hence useless. However, we can heuristically make use of (56) in the following manner. For a sound procedure with P_{D} close to 1, R can be expected to be near k_0 (a bit smaller if there are false negatives, or a bit larger when there are false positives). Using the approximation $R = k_0$ in (56) yields the following approximate upper bound on FER in terms of FDR:

$$\text{FER} = \text{prob}\left(\frac{IR}{k_0} \geq \frac{1}{k_0}\right) \approx \text{prob}\left(\frac{IR}{R} \geq \frac{1}{k_0}\right) \leq k_0 \text{FDR}$$

$$\leq \alpha k_0 \tag{57}$$

Seemingly, therefore, the P_{FA} of an FDR controlling procedure can be significantly larger than that of a procedure that controls the FER at the same level. However, while we have empirically observed higher P_{FA} values for the FDR rule (especially in cases with many false null hypotheses, i.e. $k_0 >> 1$), the increase was not as drastic as suggested by (57). A possible explanation is that the control of FDR may often be loose and therefore the achieved FDR might be considerably smaller than the bound $(M - k_0)\alpha/M$ in (4), which decreases as k_0 increases; also the procedure based on (6) in fact may control FDR at a level much smaller than α, for instance at level α/η_m whenever the regressors are (nearly) orthogonal.

Both FDR and FER rules can be used as in (9) to select the null hypotheses to be rejected. Alternatively, they can select these hypotheses using:

$$\hat{\hat{k}} = \min_k \left[k : \; T_k < q_{p_k}\right] - 1 \tag{58}$$

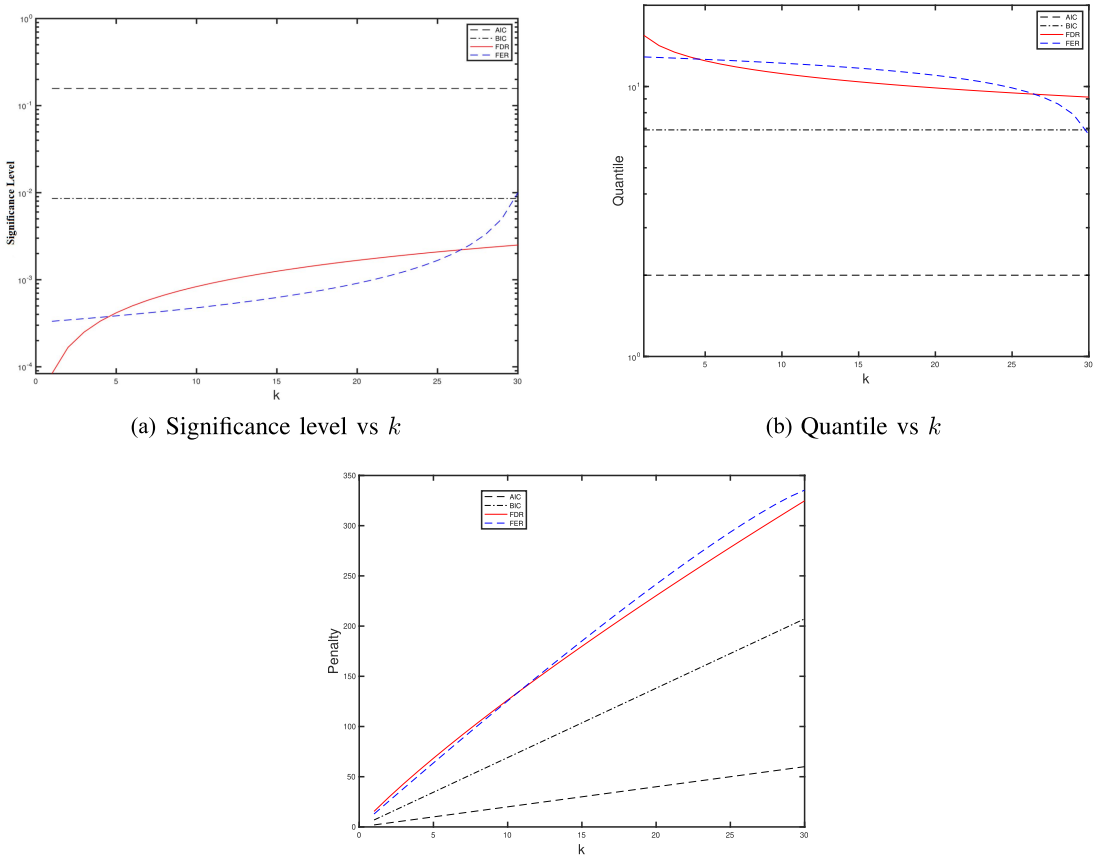
The rules based on (9) are called "stepup rules " because they proceed forward to test T_k for $k = 1, 2, \ldots,$ up to the largest k for which $T_k \geq q_{p_k}$. Different from this, the rules based on (58) can be implemented in a backward fashion to test T_k for $k = M, M-1, \ldots,$ up to the smallest k for which $T_k < q_{p_k}$ and so they are called "stepdown rules".

With reference to the associated model selection criterion C_k, the stepdown and stepup rules correspond to picking up the first ($\hat{\hat{k}}$) and, respectively, the last ($\hat{k}$) minimum point of the criterion. Clearly $\hat{k} \geq \hat{\hat{k}}$ and thus the stepup rules have a larger probability of false alarm (and, correspondingly, a smaller probability of miss) than the stepdown rules using the same sequence of significance levels. However, significant differences between the stepup and stepdown rules usually occur only for small values of N or high noise levels, a regime in which they have rather small P_{D} and hence are not really useful anyway. As N increases (or the noise power decreases), the criterion C_k tends to be unimodal and therefore the two types of rules yield same estimate $\hat{k} = \hat{\hat{k}}$. To avoid choosing between stepup and stepdown rules the user may decide to pick up the *global minimum* of C_k, which is what we will do in the numerical experiments of the next section. It is our experience that the global minimum of C_k coincides with the better one of $\hat{k}$ and $\hat{\hat{k}}$ in most realizations.

To illustrate the differences between the significance levels, quantiles and the corresponding penalties used by FDR, FER, AIC and BIC we consider an example with $M = 30$ and $N = 1000$ ($\alpha = 0.01$). From the results presented in Fig. 1 we can see that the FDR and FER penalties are larger than those of BIC and AIC.

V. SIGNAL PROCESSING APPLICATIONS AND NUMERICAL ILLUSTRATIONS

We will present four signal processing applications of the FDR and FER rules, and compare their performance with that of AIC and BIC.



(a) Significance level vs k

(b) Quantile vs k

(c) Penalty vs k

FIGURE 1. Comparison of significance levels, quantiles and penalties of FDR, FER, AIC and BIC, $M = 30$, $N = 1000$, $\alpha = 0.01$.

A. LINEAR REGRESSION

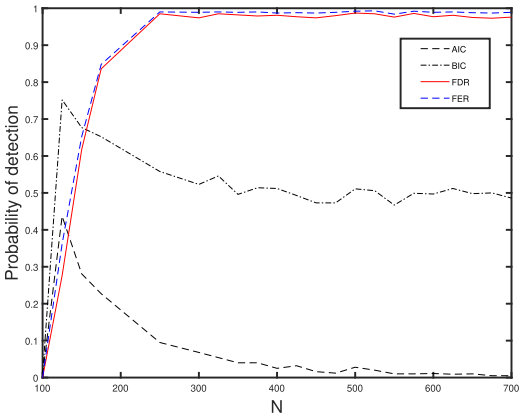
We generate the data vector $\mathbf{y}$ using (11) with $k_0 = 10$ (the true number of regressors), an $\mathbf{X}$ matrix with elements independently drawn from $\mathcal{N}(0, 1)$, $M = 100$, $\{c_j\}_{j=1}^{5} = 5$, $\{c_j\}_{j=6}^{8} = 3$ and $\{c_j\}_{j=9}^{10} = 1$; the indexes of the 10 non-zero coefficients are independently drawn from $\mathcal{U}[1, 100]$ and then fixed, and the noise variance is set to 1. The detection probability (which is the probability of selecting the 10 correct regressors) is shown as a function of N in Fig. 2(a). It can be seen from the figure that FDR and FER perform similar to each other and their performance is much better than that of AIC or BIC. The false alarm probability vs N is shown in Fig. 2(b). The false alarm probabilities of FDR and FER are quite low for all values of N; on the other hand, the false alarm probabilities of AIC and BIC are rather large and they tend to increase with N (the reason behind this behavior will be explained in the discussion at the end of the second example). In Fig. 2(c), the average model order selected by each method is shown for different values of N. It can be seen that as N increases both FDR and FER choose the correct model order. AIC, on the other hand, significantly overestimates the model order, and to some extent BIC does the same as well.

B. COMMUNICATION CHANNEL

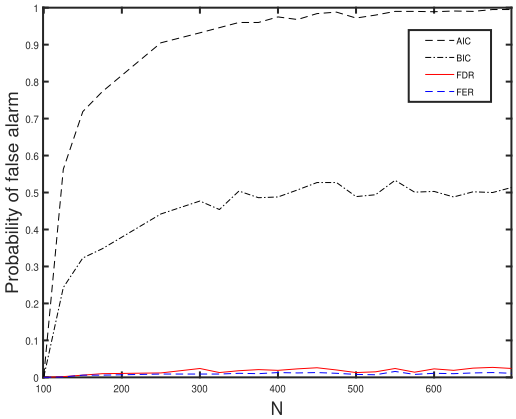
We consider a finite-memory channel with input denoted $\{u(k)\}_{k=1}^{N}$. Then the output of the channel can be written as in (11) with

$$\mathbf{X} = \begin{bmatrix} u(1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ \vdots & & u(1) \\ \vdots & & \vdots \\ u(N) & \cdots & u(N-M) \end{bmatrix} \tag{59}$$

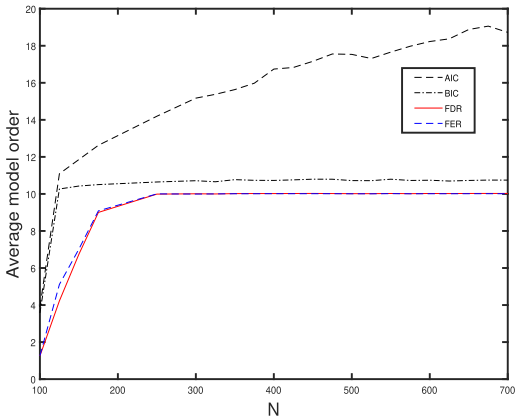
We generate $\{u(k)\}$ as a white normal sequence with zero mean and unit variance. The noise variance is $\sigma^2 = 1$. Different from the linear regression example, here we consider a case in which the sparsity of the parameter vector is more pronounced (i.e. we assume a multi-path channel): $k_0 = 5$, $M = 100$, $c_j = 2$ for $j = 1, 10, 20, 40$ and 60 and N is varying in $[100, 300]$. The detection probability, false alarm probability, and average model order are displayed versus N in Fig. 3. Similar to the linear regression case, the detection performances of FDR and FER are similar to each



(a) Probability of detection vs N



(b) Probability of false alarm vs N



(c) Average model order vs N

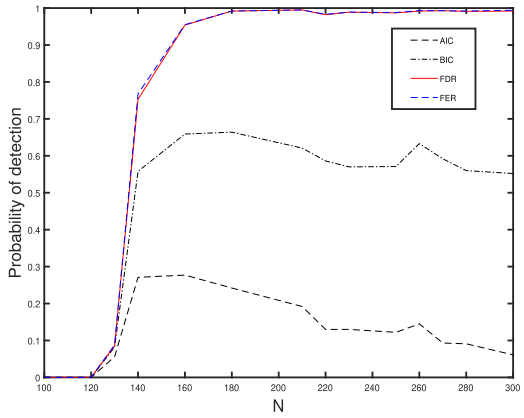
FIGURE 2. **Linear regression application. The values of different parameters: $k_0 = 10$, $M = 100$, $\{c_j\}_{j=1}^{5} = 5$, $\{c_j\}_{j=6}^{8} = 3$, $\{c_j\}_{j=9}^{10} = 1$, and $\sigma^2 = 1$.**

other and significantly better than those of AIC and BIC. Also both FDR and FER choose the correct model order as N increases.

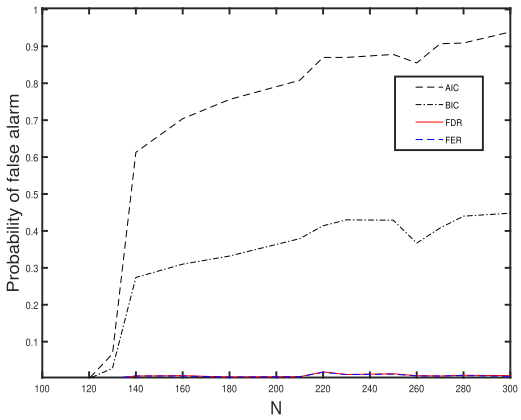
An explanation of the poor performance of AIC and BIC in the two examples above, in particular their high probability of false alarm/discovery, runs as follows. Let us assume that the k_0 true regressors have been selected by the rule. The next step is to decide if the number of regressors should be increased to $k_0 + 1$. Ideally, making this decision would require comparing the likelihoods of the model with order k_0 and all models with order $k_0 + 1$. There are $M - k_0$ models with $k_0 + 1$ regressors that we should compare with, and it is clear that if the comparison were done with each of them then the false alarm probability, which is p for one test (see Fig. 1), would significantly increase beyond p. While we make only one comparison, we compare with the model which includes the $(k_0 + 1)-$th regressor selected in the first step of the procedure (according to the ranking in (18)) and that model is likely to produce a larger increase of the likelihood function than any of the other $(M - k_0 - 1)$ models. Consequently the single comparison that we make is basically equivalent to comparing the k_0-order model with all $(M - k_0)$ models of order $(k_0 + 1)$.

The somewhat counter intuitive increase of the probability of false alarm of AIC and BIC as N increases, see the above figures, also begs an explanation. Consider AIC as an example: it will produce a false alarm if $C_{k_0} \geq C_{k_0+1}$ (k_0 being the true number of regressors). Let us say that for $N >> 1$ this happens in 100% of cases (i.e. noise realizations), as in Fig. 2. For a much smaller value of N, however, the random fluctuations of the negative-log-likelihood in the AIC criterion are much larger and therefore it can happen that by chance $C_{k_0} \leq C_{k_0+1}$ in a number of cases, which explains why the probability of false alarm can be smaller for a smaller value of N. Obviously the probability of false alarm of both AIC and BIC can be kept at bay by increasing their penalties. However there are no clear rules for how to do that in the general case: only heuristic suggestions, which lack theoretical motivation, are available.

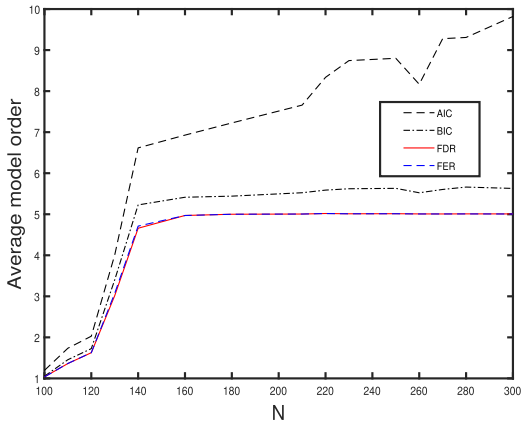
An interesting aspect, related to the above discussion, concerns the application of selection rules to the nonlinear model in (41) which will be considered in the next two examples.



(a) Probability of detection vs N



(b) Probability of false alarm vs N



(c) Average model order vs N

FIGURE 3. **Communication channel application. The values of different parameters: $k_0 = 5$, $M = 100$, $c_1 = c_{10} = c_{20} = c_{40} = c_{60} = 2$, and $\sigma^2 = 1$.**

As explained in Section III-B, in the case of this model, it is the parameter estimation method that implicitly selects the "best regressor" to be added to the model when we increase the order by one ($k \to k+1$). However, in the case of (41), the regressor set is a manifold (spanned by the parameters $\{\mathbf{b}_j\}$) and hence M is theoretically infinite. Consequently, the situation seems to be worse than in the linear model case, unless the rule has a built-in feature that increases the penalty term to prevent a blow-up of the false-alarm probability. This feature does exist: when we increase the order from k to $k+1$, the number of parameters (and hence the number of degrees of freedom of the $\mathcal{X}^2$ distribution of the likelihood ratio statistics) increases by $1+\dim(\mathbf{b})$ instead of just by 1 (assuming, for simplicity, that $\dim(\mathbf{b}_j)$ does not depend on j). This leads to a proportionate increase of the penalties of AIC and BIC. For FDR and FER the penalties also increase due to the increase of the quantiles that compose the penalty terms of the latter rules. In addition to this (as mentioned above) in the case of nonlinear models the regressor set is a manifold, which can be accurately sampled (or covered) only if M is sufficiently large. A larger value of M leads to smaller significance levels for both FDR and FER, which in turn leads to an increase of the penalty terms of these rules and therefore a reduction of the false-alarm probability.

C. SINUSOIDAL SIGNALS

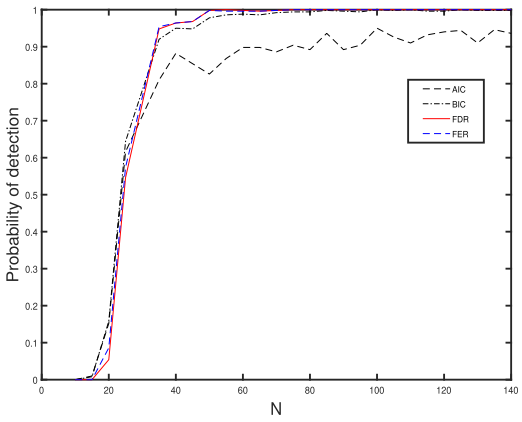
This is a typical signal processing application of the nonlinear model discussed in Section III-B with

$$\mathbf{x}_j = [\sin(2\pi f_j + \phi_j) \ldots \sin(N2\pi f_j + \phi_j)]^T, \tag{60}$$

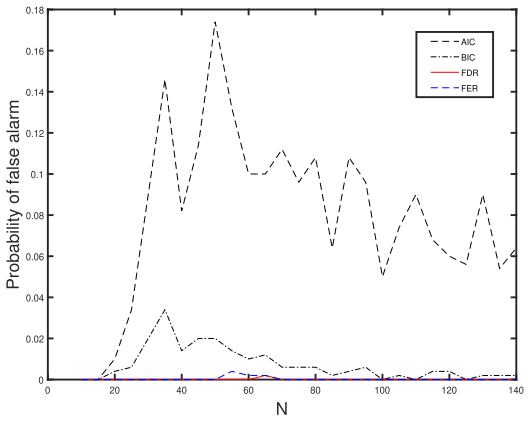
where f_j is the normalized frequency and ϕ_j is the initial phase of $j-$th sinusoid. A generic regressor vector of the above form can be re-written as:

$$\begin{bmatrix} \cos(2\pi f) & \sin(2\pi f) \\ \vdots & \vdots \\ \vdots & \vdots \\ \cos(N2\pi f) & \sin(N2\pi f) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \tag{61}$$

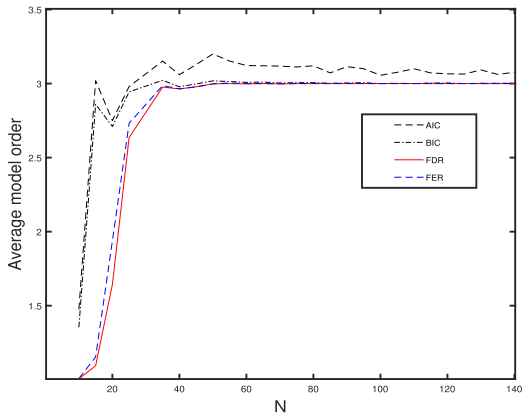
where $a = \sin(\phi)$ and $b = \cos(\phi)$. The nonlinear parameter of (61) is the frequency f whose variation in the interval [0,1] generates the regressor set/manifold. It follows from basic results in spectral analysis (see e.g. [15]) that a sampling of f using a grid with step $1/N$ yields a satisfactory covering of



(a) Probability of detection vs N



(b) Probability of false alarm vs N



(c) Average model order vs N

FIGURE 4. **Sinusoidal signal application. Parameter settings: $k_0 = 3$, $c_1 = 5$, $c_2 = 3$, $c_3 = 2$, $f_1 = 0.1$, $f_2 = 0.25$, $f_3 = 0.42$, $\phi_1 = 0.1\pi$, $\phi_2 = -0.7\pi$, $\phi_3 = -0.32\pi$, and $\sigma^2 = 1$.**

the said manifold. Consequently, in the present case we can use the significance levels of FDR and FER in (6) and (46) with $M = N$:

$$p_k^{\text{FDR}} = \frac{\alpha k}{N(0.577 + \ln N)} \tag{62}$$

$$p_k^{\text{FER}} = \frac{\alpha}{N + 1 - k} \tag{63}$$

From the above significance levels we can calculate the corresponding quantiles (and hence the penalties) of the two rules using a χ^2 calculator for a distribution with 3 degrees of freedom (which is the number of unknown parameters in (61): a, b and f).

We consider the following specific values of the parameters in this example: $k_0 = 3, c_1 = 5,\ c_2 = 3,\ c_3 = 2,\ f_1 = 0.1,\ f_2 = 0.25,\ f_3 = 0.42,\ \phi_1 = 0.1\pi,\ \phi_2 = -0.7\pi,\ \phi_3 = -0.32\pi,\ N$ varying in $[10, 140]$, and $\sigma^2 = 1$. We obtain approximate maximum likelihood estimates of $\{f_j\}_{j=1}^k$ using the periodogram method (see e.g. [15]) and then use $\{\hat{f}_j\}$ to estimate $\{c_j, \phi_j\}_{j=1}^k$ via a simple linear least squares method. The estimate of σ^2 is obtained as in (13). The probability of correct detection, probability of false alarm and average model order selected by the four rules are shown as a function of N in Fig. 4. In this example BIC worked almost as well as FDR and FER and the performance of AIC also improved significantly compared with what we saw in the previous examples. This was expected as BIC is known to be consistent in the present application, and AIC is known to work reasonably well too.

D. LOW-RANK MATRIX APPROXIMATION

We consider a noisy matrix whose "signal part" has a low rank :

$$\mathbf{A} = \mathbf{B}\mathbf{C}\mathbf{D}^T + \mathbf{E} \tag{64}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{m\times k_0}$, $\mathbf{D} \in \mathbb{R}^{n\times k_0}$, and

$$\mathbf{C} = \begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & c_{k_0} \end{bmatrix} \in \mathbb{R}^{k_0\times k_0} \tag{65}$$

with $k_o << \min(m, n)$. The elements of the noise matrix $\mathbf{E}$ in (64) are assumed to be i.i.d. normal variables with zero mean and variance denoted σ^2. The low-rank decomposition in (64) is not unique and, to be more specific, we assume that $\mathbf{BCD}^T$ is the singular value decomposition (SVD) of the signal part of (64), therefore $\mathbf{B}$ and $\mathbf{D}$ are semi-unitary matrices and $\mathbf{C}$ is positive semi-definite ($\mathbf{B}^T\mathbf{B} = \mathbf{D}^T\mathbf{D} = \mathbf{I}$ and $\{c_k \geq 0\}$). Furthermore, to simplify the notation and exposition we focus on the case of square matrices: $m = n$ (the general case of $m \neq n$ can be treated in much the same manner).

We can re-write (64) as in (41) with $\mathbf{y} = \text{vec}(\mathbf{A})$, $\mathbf{E} = \text{vec}(\mathbf{E})$, and the following unknown "regressor" vectors (with a Kronecker-product structure):

$$\mathbf{x}_k = \mathbf{d}_k \otimes \mathbf{b}_k \tag{66}$$

where $\mathbf{b}_k$ and $\mathbf{d}_k$ are the $k-$th columns of $\mathbf{B}$ and $\mathbf{D}$. Under H_k, the negative log-likelihood function associated with (64) is given by (to within a constant):

$$-\ln p(\mathbf{A}|\mathbf{B}_k, \mathbf{C}_k, \mathbf{D}_k, \sigma_k^2) = \frac{m^2}{2}\ln\sigma_k^2 + \frac{1}{2\sigma_k^2}||\mathbf{A} - \mathbf{B}_k\mathbf{C}_k\mathbf{D}_k^T||_F^2 \tag{67}$$

It is well known that the maximum likelihood estimates $\hat{\mathbf{B}}_k, \hat{\mathbf{C}}_k, \hat{\mathbf{D}}_k$ which minimize the above function with respect to $\mathbf{B}_k, \mathbf{C}_k, \mathbf{D}_k$ can be obtained by truncating the SVD of the matrix $\mathbf{A}$ keeping only the k largest singular values and associated singular vectors [16] (note that the SVD provides *all* estimated models for $k = 1, \ldots, m$). The estimate of σ_k^2 is then given by :

$$\hat{\sigma}_k^2 = \frac{1}{m^2}\|\mathbf{A} - \hat{\mathbf{B}}_k\hat{\mathbf{C}}_k\hat{\mathbf{D}}_k^T||_F^2 = \frac{\hat{c}_{k+1}^2 + \cdots + \hat{c}_m^2}{m^2} \tag{68}$$

The likelihood ratio statistics T_k have the same expression as in (28) but with N replaced by m^2 :

$$T_k = m^2 \ln\left(\frac{\hat{\boldsymbol{\sigma}}_{k-1}^2}{\hat{\sigma}_k^2}\right) \tag{69}$$

However in the present case T_k has an asymptotic chi-square distribution with $2\,m$ degrees of freedom (see (23)): $T_k \sim \chi^2(2\,m)$. This means that the quantiles q_{p_k} should be computed for $\chi^2(2\,m)$, with $2\,m$ being possibly much larger than the number of degrees of freedom encountered in the previous cases. The main difference from the previous applications, however, is the fact that finding a satisfactory sampling of the regressor vector set is a bit more complicated in the present case.

Consider the FDR first. A generic vector $\mathbf{b}$ (and similarly for $\mathbf{d}$) lies on the $(m-1)$-dimensional surface of the unit sphere in $\mathbb{R}^m$. For 1D this "surface" consists of 2 points ($+1$ and -1). In 2D the said surface is a circle, and we can use 360 vectors to sample it (with 1 deg between adjacent vectors). In 3D we can cover the surface using 360 deg in longitude and 180 deg in latitude, therefore we can use $180 \times 360 = 2(180)^2$ vectors to sample it. Generalizing to $m-$dimensions leads to the conclusion that a number of $2(180)^{m-1}$ vectors can be used to sample the set of a generic $\mathbf{b}$ regressor (and similarly for $\mathbf{d}$). Because all possible combinations between the $\mathbf{b}$ and $\mathbf{d}$ vectors can occur, the total number of regressor vectors is:

$$4(180)^{2(m-1)} \tag{70}$$

Consequently the significance levels of FDR are given by:

$$p_k^{\text{FDR}} = \frac{\alpha k}{4(180)^{2(m-1)}[0.577 + 2(m-1)\ln(180) + \ln 4]} \tag{71}$$

where we used the approximation in (7) for the harmonic number (which holds well here because (70) is a large number).

Next we consider the FER, for which the only difference from the above discussion is as follows. For $\mathbf{b}_1$ there is no difference as for FER too this vector is only restricted to lie on the surface of the unit sphere in $\mathbb{R}^m$, but $\mathbf{b}_2$ must be orthogonal to $\mathbf{b}_1$ (and similarly for $\mathbf{d}_1$ and $\mathbf{d}_2$); hence $\mathbf{b}_2$ belongs to a sphere in $\mathbb{R}^{m-1}$, $\mathbf{b}_3$ to a sphere in $\mathbb{R}^{m-2}$ (because $\mathbf{b}_3$ must be orthogonal to both $\mathbf{b}_1$ and $\mathbf{b}_2$), and so forth. Making use of this observation and of (70) we can write the significance levels of FER as:

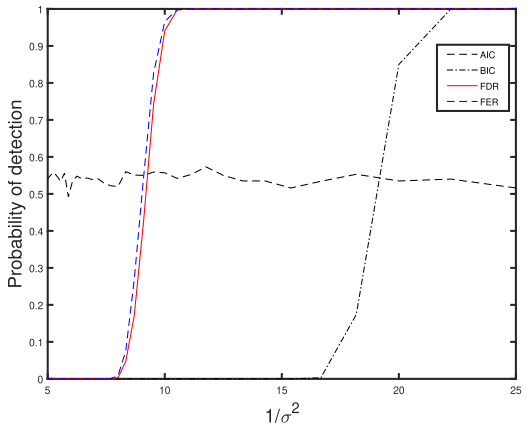
$$p_k^{\text{FER}} = \frac{\alpha}{4(180)^{2(m-k)}} \tag{72}$$

Remark 4: The above sampling of the spherical surface, based on patches of 1 deg in longitude and latitude, is denser close to the poles than around the equator (easy to visualize in 3D). If desired, a uniform patching can be obtained using the expression for the solid angle subtended by the spherical surface in $\mathbb{R}^m$:

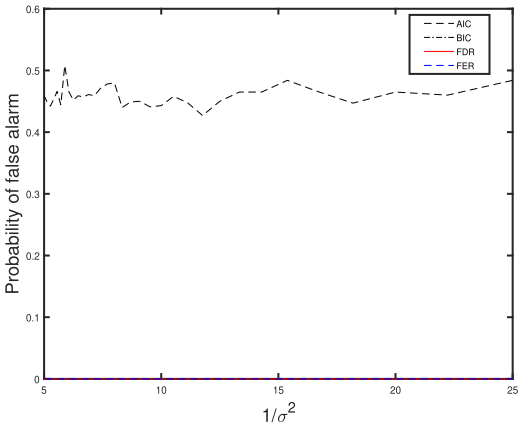
$$\Omega_m = \frac{2\pi^{m/2}}{\Gamma(m/2)} \quad \text{[in steradians]} \tag{73}$$

where $\Gamma(\cdot)$ is the gamma function. The conversion of (73) from steradians to "degrees" (for example, square-degrees for 3D) can be done multiplying (73) by $\left(\frac{180}{\pi}\right)^{m-1}$. For $m = 1$, 2 and 3 this gives $\Omega_m = 2$, 360, and $4(180)^2/\pi$. The first two of these numbers are the same as the values given by our simpler formula, $2(180)^{m-1}$, but the third one is smaller (which was expected as our sampling is denser close to the poles). We have tried using (73) in FDR and FER rules in a limited number of cases but failed to notice any significant difference from the performance of the rules based on our simpler (although less uniform) sampling formula.

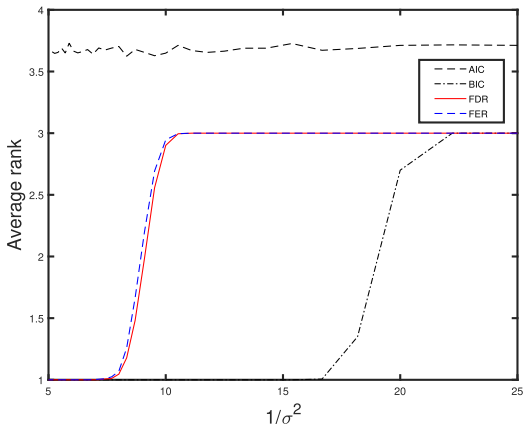
Regarding the implementation of (71) and (72), note that for large values of m ($m > 10$), the chi-square calculator in MATLAB can fail to return meaningful quantiles. To circumvent this numerical problem we suggest the following alternative. We first calculate the quantile corresponding to p_k from the standard normal distribution and then use the approximation $\chi^2(2\,m) \approx \mathcal{N}(2\,m, 4\,m)$, which holds for sufficiently large m, to obtain the corresponding quantile for the chi-square distribution. To compute the quantiles of the normal distribution (as the normal distribution calculator in



(a) Probability of detection vs $1/\sigma^2$



(b) Probability of false alarm vs $1/\sigma^2$



(c) Average rank vs $1/\sigma^2$

FIGURE 5. **Low-rank matrix approximation application. Parameter settings: $k_0 = 3$, $m = n = 100$, $c_1 = 5$, $c_2 = 3$, $c_3 = 3$.**

MATLAB does not work well either for small p values) we used the fact that $\ln(p)$ is well approximated (for sufficiently small p) by the following function of the quantile (q) of the normal distribution [17]:

$$\ln(p) = \ln\left(\frac{e^{-\frac{q^2}{2}}}{\sqrt{2\pi}\,q}\right) \tag{74}$$

$$= -\frac{q^2}{2} - \ln(\sqrt{2\pi}) - \ln(q) \tag{75}$$

In sum for given p, we first solve the above nonlinear equation using the Fsolve function in MATLAB and obtain q and next use q in the normal approximation of the chi-square distribution to obtain the corresponding chi-square quantile.

In the simulation example of (64) we consider the following specifications: $k_0 = 3$, $m(= n) = 100$, $c_1 = 5$, $c_2 = 3$, and $c_3 = 3$. Furthermore $\mathbf{B}$ and $\mathbf{D}$ are $m \times 3$ semi-unitary matrices obtained from the QR decomposition of random matrices with i.i.d. normal elements. Fig. 5 shows the probability of detection, probability of false alarm, and the average model order selected by the four rules for a range of values of $1/\sigma^2$. The detection probabilities of FDR and FER almost coincide with each other and they converge to one much faster than the detection probability of BIC. On the other hand, AIC has a constant detection probability over the range of noise variances considered, which is almost equal to probability of false alarm of this rule.

Remark 5: Asymptotic performance studies of model selection rules typically require that $[\dim(\boldsymbol{\theta}_k)]^2/N \to 0$ as $N \to \infty$. This condition was satisfied in all three previous examples. However, in the present case the above ratio is larger than 1 even for $k = 1$: $(2\,m)^2/m^2 = 4$. Consequently the performance of the rules is not expected to increase with m. That is the reason why in this example we have plotted the performance metrics versus $1/\sigma^2$.

VI. CONCLUSION

Most models used in signal processing applications contain both real-valued parameters and integer-valued parameters (which determine the structure, in particular the orders, of the model). The estimation of the former is well studied and understood and there are optimal methods, such as the method

of the maximum likelihood, which attain the ultimate accuracy. On the other hand, the estimation of the latter parameters (i.e., the operation of model selection), while also extensively studied, is less understood and well-established methods that always perform better than any competitor do not exist. AIC and BIC rules, without a doubt, are the workhorses of model selection in the signal processing literature and elsewhere (see, e.g. [2], [3], [4]). Under certain conditions, the former is known to overestimate the true orders with a (false alarm) probability of about 0.16, whereas the latter is consistent (as $N \to \infty$). AIC's tendency of overestimation should not be viewed as a drawback if the model is to be used for prediction. However, if the main purpose of the modeling exercise is acquiring information about the actual data generating mechanism and model interpretability is important, then consistent (or nearly so) model selection becomes a key factor and consequently BIC is to be preferred to AIC.

The problem is that, as illustrated in the previous section, in applications in which the range of possible model orders is considerable (i.e., $M >> 1$), even BIC can perform rather poorly. The primary aim of this paper was to show that model selection methods, which perform much better than BIC in such applications and not worse in regular cases in which BIC is consistent, do exist. We have shown how the FDR and FER procedures of multiple hypothesis testing can be used to derive such methods (or rules) that can perform better in terms of model selection accuracy than both AIC and BIC. A secondary goal of the paper was to briefly introduce the FDR and FER principles to those readers who were less familiar with them.

The FDR and FER rules presented in the previous sections can be directly used in many signal processing applications. Furthermore, the basic ideas of these rules can be employed to develop new rules for model selection based on even more advanced results and methods from the multiple hypothesis testing literature. Indeed there exist several enhanced versions of (6) and (46) as well as more sophisticated methods for controlling the FDR or FER (see, e.g., [18], [19]). Can we use the methodology described in this paper along with those more advanced hypothesis testing methods to obtain new model selection rules with enhanced performance? This is a question that is worthy of further research in our opinion.

APPENDIX
PROOF OF (4) FOR INDEPENDENT STATISTICS AND THE SIGNIFICANCE LEVELS IN (10)

Using the definitions of $\hat{k}$ in (9) and C in (47) we can write FDR as:

$$\text{FDR} = \mathbb{E}\left[\frac{\sum_{j\in C} \mathrm{I}\left(T_j \geq q_{p_{\hat{k}}}\right)}{\hat{k}}\right] \tag{76}$$

where $\mathrm{I}(\cdot)$ denotes the indicator function ($\mathrm{I}(A) = 1$ if $A =$ true and $\mathrm{I}(A) = 0$ if $A =$ false). The expression for FDR in (76) holds because any H_j^0 with $j > \hat{k}$ (and hence $T_j < q_{p_{\hat{k}}}$) is accepted, whereas any H_j^0 with $j \leq \hat{k}$ is rejected and its test statistics satisfies $T_j \geq T_{\hat{k}} \geq q_{p_{\hat{k}}}$. Interestingly there can be T_j's that are under their quantiles, $T_j < q_{p_j}$, but they are rejected too if $j \leq \hat{k}$ (see (9)).

The main difficulty associated with evaluating the expectation in (76) is that T_j and $\hat{k}$ are not independent of each other. To overcome this problem we will try to find a $\tilde{k}_j$ that is independent of T_j and which is such that

$$\frac{\mathrm{I}\left(T_j \geq q_{p_{\hat{k}}}\right)}{\hat{k}} \leq \frac{\mathrm{I}\left(T_j \geq q_{p_{\tilde{k}_j}}\right)}{\tilde{k}_j} \; \forall j \tag{77}$$

To do so, we replace T_j by a value $\tilde{T}_1 >> 1$ (theoretically infinite, but a value $\tilde{T}_1 > T_1$ suffices). Then we re-order the statistics with T_j replaced by $\tilde{T}_1$:

$$[T_1, \ldots, T_{j-1}\tilde{T}_1, T_{j+1}, \ldots, T_M] \to [\tilde{T}_1, \tilde{T}_2, \ldots, \tilde{T}_M] \tag{78}$$

(with $\tilde{T}_1 \geq \tilde{T}_2 \cdots \geq \tilde{T}_M$). Let $\tilde{k}_j$ be defined similarly to $\hat{k}$ but for the statistics $\{\tilde{T}_k\}$. From the definition in (78) it is clear that $\tilde{T}_k = T_{k-1}$ for $k = 2, \ldots, j$ and $\tilde{T}_k = T_k$ for $k = j+1, \ldots, M$. This observation has the following implications:

- If $j > \hat{k}$ then $\mathrm{I}(T_j \geq q_{p_{\hat{k}}}) = 0$ and thus (77) must hold. Note that the inequality in (77) can be strict in such a case because $\tilde{T}_k \geq T_k$ (for $k = 1, \ldots, j$), consequently we might have $\tilde{T}_k \geq q_{p_k}$ even when $T_k < q_{p_k}$. This means that in the present case $\tilde{k}_j \geq \hat{k}$ with strict inequality being possible and we have $q_{p_{\tilde{k}_j}} < q_{p_{\hat{k}}}$ if $\tilde{k}_j > \hat{k}$ and hence the RHS in (77) can be > 0.
- If $j < \hat{k}$ then $\tilde{T}_k = T_k$ for $k = \hat{k}, \ldots, M$ and thus $\tilde{k}_j = \hat{k}$ so (77) holds.
- Finally, If $j = \hat{k}$ then $\tilde{T}_k = T_k$ for $k = \hat{k}+1, \ldots, M$ and $\tilde{T}_{\hat{k}} = T_{\hat{k}-1} \geq T_{\hat{k}} \geq q_{p_{\hat{k}}}$, so we still have $\tilde{k}_j = \hat{k}$ and (77) holds true.

Using (77) along with the fact that $\tilde{k}_j$ and T_j are independent of one another (as $\tilde{k}_j$ depends on $\{T_k\}_{k\neq j}$) we can write:

$$\begin{aligned}
\text{FDR} &\leq \sum_{j\in C} \mathbb{E}\left(\frac{\mathrm{I}(T_j \geq q_{p_{\tilde{k}_j}})}{\tilde{k}_j}\right) \\
&= \sum_{j\in C} \mathbb{E}_{\tilde{k}_j}\left\{\mathbb{E}_{T_j}\left(\frac{\mathrm{I}(T_j \geq q_{p_{\tilde{k}_j}})}{\tilde{k}_j}\right)\right\} \\
&= \sum_{j\in C} \mathbb{E}_{\tilde{k}_j}\left\{\frac{\mathrm{prob}(T_j \geq q_{p_{\tilde{k}_j}})}{\tilde{k}_j}\right\} \qquad (79) \\
&= \sum_{j\in C} \frac{\alpha\tilde{k}_j}{M\tilde{k}_j} = \frac{\tilde{M}}{M}\alpha \qquad (80)
\end{aligned}$$

The proof of (4) is thus concluded.

ACKNOWLEDGMENT

The authors would like to thank Associate Editor (Prof. Usman Khan) and the four reviewers for their constructive and helpful comments.

REFERENCES

[1] T. Soderstrom and P. Stoica, *System Identification*. Upper Saddle River, NJ, USA: Prentice Hall, 1989. [Online]. Available: https://user.it.uu.se/ps/ps.html

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

[3] K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," *Sociol. Methods Res.*, vol. 33, no. 2, pp. 261–304, 2004.

[4] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

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

[6] P. Stoica and Y. Selén, "Cross-validation rules for order estimation," *Digit. Signal Process.*, vol. 14, no. 4, pp. 355–371, 2004.

[7] T. W. Anderson, "The choice of the degree of a polynomial regression as a multiple decision problem," *Ann. Math. Statist.*, vol. 33, pp. 255–265, 1962.

[8] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc. Ser. B Methodol.*, vol. 57, pp. 289–300, 1995.

[9] Y. Benjamini and Y. Gavrilov, "A simple forward selection procedure based on false discovery rate control," *Ann. Appl. Statist.*, vol. 3, pp. 179–198, 2009.

[10] F. Bunea, M. H. Wegkamp, and A. Auguste, "Consistent variable selection in high dimensional regression via multiple testing," *J. Stat. Plan. Inference*, vol. 136, no. 12, pp. 4349–4364, 2006.

[11] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, pp. 65–70, 1979.

[12] E. Spjøtvoll, "Ordering ordered parameters," *Biometrika*, vol. 64, no. 2, pp. 327–334, 1977.

[13] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Statist.*, vol. 29, pp. 1165–1188, 2001.

[14] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statist.*, vol. 9, no. 1, pp. 60–62, 1938.

[15] P. Stoica and R. L. Moses, *Spectral Anal. of Signals*. Upper Saddle River, NJ, USA: Prentice Hall, 2005. [Online]. Available: https://user.it.uu.se/ps/ps.html

[16] C. F. Van Loan and G. Golub, *Matrix Computations*. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.

[17] M. Abramowitz, I. A. Stegun, and R. H. Romer, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. College Park, MD, USA: Amer. Assoc. Phys. Teachers, 1988.

[18] X. Cui, T. Dickhaus, Y. Ding, and J. C. Hsu, *Handbook of Multiple Comparisons*. Boca Raton, FL, USA: CRC Press, 2021.

[19] R. F. Barber and E. J. Candes, "Controlling the false discovery rate via knockoffs," *Ann. Statist.*, vol. 43, no. 5, pp. 2055–2085, 2015.