ON THE NUMBER OF ITERATIONS FOR THE MATCHING PURSUIT ALGORITHM

Fangyao Li[1], *Christopher M. Triggs*[1], *Bogdan Dumitrescu*[2], *Ciprian Doru Giurcăneanu*[1]

[1]Department of Statistics
University of Auckland
Private Bag 92019, Auckland 1142, New Zealand

[2]Department of Automatic Control and Computers
University Politehnica of Bucharest
313 Spl. Independenţei, 060042 Bucharest, Romania

ABSTRACT

We address the problem of selecting, from a given dictionary, a subset of predictors whose linear combination provides the best description for the vector of measurements. To this end, we apply the well-known matching pursuit algorithm (MPA). Even if there are theoretical results on the performance of MPA, there is no widely accepted rule for stopping the algorithm. In this work, we focus on stopping rules based on information theoretic criteria (ITC). The key point is to evaluate the degrees of freedom (df) for the model produced at each iteration. This is traditionally done by computing the trace of the hat matrix which maps the data vector to its estimate. We prove some theoretical results concerning the hat matrix. One of them provides an upper bound on the increase of df from the m-th to the $(m+1)$-th iteration. Based on the properties of the hat matrix, we propose novel ITC for selecting the number of iterations. All of them are obtained by modifying criteria designed for variable selection in the classical linear model. For assessing the performance of the novel criteria, we conduct a simulation study.

Index Terms— Matching pursuit algorithm, hat matrix, projector, information theoretic criteria, number of iterations

1. INTRODUCTION

The matching pursuit algorithm (MPA) is well-known in the signal processing community mainly due to [1]; it is also applied in statistics and in approximation theory where it is known as L_2-boosting [2] and as the pure greedy algorithm [3], respectively. When MPA is used for high-dimensional linear models, existing theoretical results do not provide a stopping rule for the algorithm (see, for example, [4]). In practical applications the number of iterations is selected either by using cross-validation (CV) or an information theoretic (IT) criterion [2]. The major disadvantage of CV is its computational burden. The IT criteria provide a promising alternative, but until now their use was restricted only to variants of the Akaike Information Criterion [5].

Note that the definition of any information criterion involves the degrees of freedom (df) of the models which compete for the description of the data vector $\mathbf{y}$. For a model that produces an estimate $\hat{\mathbf{y}}$ of $\mathbf{y}$, df is traditionally taken to be the trace of the hat matrix, i.e. the trace of the linear operator which maps $\mathbf{y}$ to $\hat{\mathbf{y}}$ [6]. However, empirical studies have demonstrated that the trace-based evaluation may underestimate the value of df [7].

We prove theoretical results concerning the properties of the hat matrix generated at each iteration of MPA. To this end, we review MPA in Sec. 2 , then present in Sec. 3 the outcome of our theoretical analysis. One important result is an upper bound on the increase of df from the m-th to the $(m+1)$-th iteration ($m \geq 1$). As part of the analysis we show that, in general, the hat matrix is not idempotent, which means that it is not a projector. Based on the properties of the hat matrix, we propose in Sec. 4 novel IT criteria for selecting the number of iterations. All of them are obtained by modifying criteria designed for variable selection in the classical linear model. We conduct a simulation study for assessing the performance of the novel criteria, and the results are reported in Sec. 5. Sec. 6 concludes the paper.

2. MATCHING PURSUIT ALGORITHM

Notation: We employ bold letters to denote both vectors and matrices; upper case letters are used only for matrices. $\mathbf{I}$ denotes the identity matrix of appropriate size, while $\mathbf{0}$ denotes the vector/matrix for which all the entries are equal to zero. The Euclidean norm of a vector $\mathbf{v}$ is $||\mathbf{v}||$. For an arbitrary matrix $\mathbf{M}$, $\mathbf{M}^\top$ is the transpose matrix and $\mathrm{Sp}(\mathbf{M})$ denotes the linear subspace spanned by the columns of $\mathbf{M}$.

Algorithm: Assume that the response vector $\mathbf{y} = [y_1, \dots, y_n]^\top$ is given, as well as the matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_{p_n}]$ of potential predictors for $\mathbf{y}$. The number of predictors, p_n, can grow very fast when n increases (see the discussion in [8, Sec. 3]). MPA can be useful for selecting the most significant predictors which should be included in the linear model for $\mathbf{y}$. If $\mathbf{X}\hat{\boldsymbol{\beta}}$ is the fitted linear model, then all non-zero entries of $\hat{\boldsymbol{\beta}}$ correspond to the selected predictors. The residuals are given

E-mails: lfan523@aucklanduni.ac.nz, cm.triggs@auckland.ac.nz, bogdan.dumitrescu@acse.pub.ro, c.giurcaneanu@auckland.ac.nz.

ISBN 978-0-9928626-7-1 © EURASIP 2017

by $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. In the initialization phase of the algorithm, $\hat{\boldsymbol{\beta}}$ is set to $\mathbf{0}$. At each iteration, MPA selects the column of $\mathbf{X}$ leading to the largest reduction of the residual sum of squares. Assume that, at the j-th step of the algorithm, the column of $\mathbf{X}$ indexed by $s(j)$ is selected, where $1 \leq s(j) \leq p_n$. Then, only the $s(j)$-th entry of $\hat{\boldsymbol{\beta}}$ is updated by using the formula $\hat{\beta}_{s(j)} \leftarrow \hat{\beta}_{s(j)} + \nu(\mathbf{x}_{s(j)}^\top \mathbf{x}_{s(j)})^{-1}\mathbf{x}_{s(j)}^\top \mathbf{e}$.

The parameter $\nu \in (0, 1]$ is the step size, also known as the shrinkage parameter. Note that all other entries of $\hat{\boldsymbol{\beta}}$ remain unchanged. This is a major difference from Orthogonal Matching Pursuit (OMP) as OMP re-estimates all the entries of the vector of linear parameters at each step of the algorithm. The two algorithms have been already compared in [2, Sec. 12.7.1.1].

In general, the value of the shrinkage parameter in MPA is taken to be small, for example, $\nu = 0.1$. This is justified in [2, Sec. 12.6.2.1] by emphasizing the relationship between MPA and the well-known Lasso algorithm [9]. Another peculiarity of MPA is that the same predictor can be selected not only once, but multiple times during the iterations of the algorithm even when $\nu = 1$. This makes it difficult to evaluate the complexity of the linear model produced at each step of MPA. We discuss this aspect below.

Hat matrix: Let $\hat{\mathbf{y}}_m = \mathbf{X}\hat{\boldsymbol{\beta}}_m$ be the estimate of $\mathbf{y}$ obtained after the m-th step of the algorithm. We denote by $\mathbf{B}_m$ the linear operator, equivalently the hat-matrix, which maps $\mathbf{y}$ to $\hat{\mathbf{y}}_m$: $\hat{\mathbf{y}}_m = \mathbf{B}_m\mathbf{y}$. Recalling that $\mathbf{x}_{s(j)}$ denotes the predictor selected at the j-th iteration of MPA, $\mathbf{B}_m$ is expressed as [8]:

$$\mathbf{B}_m = \mathbf{I} - \mathbf{A}_m, \text{ where} \tag{1}$$

$$\mathbf{A}_m = \left(\mathbf{I} - \nu\mathbf{P}_{s(m)}\right) \cdots \left(\mathbf{I} - \nu\mathbf{P}_{s(1)}\right), \tag{2}$$

$\mathbf{P}_{s(j)} = \bar{\mathbf{x}}_{s(j)}\bar{\mathbf{x}}_{s(j)}^\top$ and $\bar{\mathbf{x}}_{s(j)} = \mathbf{x}_{s(j)}/||\mathbf{x}_{s(j)}||$ for $1 \leq j \leq m$. It can be shown by mathematical induction that

$$\mathbf{A}_m = \sum_{k=0}^{m} \mathbf{S}_{m,k}, \text{ where } \mathbf{S}_{m,0} = \mathbf{I} \tag{3}$$

and we have for $1 \leq k \leq m$:

$$\mathbf{S}_{m,k} = (-\nu)^k \sum_{m \geq j_k > j_{k-1} > \cdots > j_1 \geq 1} \mathbf{P}_{s(j_k)}\mathbf{P}_{s(j_{k-1})} \cdots \mathbf{P}_{s(j_1)}. \tag{4}$$

The matrix $\mathbf{B}_m$ is important in evaluating the complexity of the linear model produced at the m-th step. More precisely, the degrees of freedom for the fitted model are estimated by $\mathrm{df}_m = \mathrm{tr}(\mathbf{B}_m)$. This formula has been used, for example, in [6]. It follows from Stein's theory on unbiased risk estimation [10] that for the case when the design matrix is fixed and the residuals are i.i.d. normal, with zero-mean and known variance σ^2, $\mathrm{df} = \sum_{j=1}^{n} \mathrm{Cov}(\hat{y}_j, y_j)/\sigma^2$ [11, 12]. It is a simple exercise to demonstrate that this expression equals the trace of the hat matrix (see [2, Eq. (2.34)]).

In practice, the user chooses an upper bound, m_{ub}, for the number of iterations. It is often recommended to use an IT criterion for selecting the best model from the m_{ub} different models which were produced during these iterations. Because of the particularities of MPA, the IT criteria which have been previously derived for the classical linear model cannot be applied in their original form [8]. The modifications of the criteria are discussed in Sec. 4. They are based on the properties of the hat matrix outlined in the next section.

3. SOME PROPERTIES OF THE HAT MATRIX

We give some theoretical results whose proofs are detailed in the supplemental material [13]. Firstly we show that, at each step of the MPA, the increase of df is at most ν. This can be recast as a property of the hat matrix:

Proposition 3.1. *For $m \geq 1$, we have*

$$\mathrm{tr}(\mathbf{B}_{m+1}) - \mathrm{tr}(\mathbf{B}_m) \leq \nu. \tag{5}$$

The equality holds if and only if $\bar{\mathbf{x}}_{s(m+1)}^\top \bar{\mathbf{x}}_{s(j)} = 0$ for all $j \in \{1, \ldots, m\}$.

Remark 1. *For all $m \geq 1$, one can show that $\mathrm{tr}(\mathbf{B}_{m+1}) - \mathrm{tr}(\mathbf{B}_m) \geq -\nu$ by using Result 3 and inequality (3) from [13]. In practice, it is observed that $\mathrm{tr}(\mathbf{B}_{m+1}) - \mathrm{tr}(\mathbf{B}_m)$ can be negative, hence is not guaranteed that df increases at each iteration of MPA.*

$\mathbf{B}_m$ is not, in general, a projection matrix. As a square matrix is a projector if and only if it is idempotent (see, for example, [14, Th. 2.1]), we check when $\mathbf{B}_m^2 = \mathbf{B}_m$.

Proposition 3.2. *(i) If $\nu \in (0, 1)$, then $\mathbf{B}_m$ is not idempotent for all $m \geq 1$.*
(ii) Consider the following conditions: (c_1) $m \geq 2$; (c_2) $\nu = 1$; (c_3) $\bar{\mathbf{x}}_{s(i)}^\top \bar{\mathbf{x}}_{s(j)} = 0$ for all $i, j \in \{1, \ldots, m\}$ with property $i > j$. If all these conditions are satisfied, then $\mathbf{B}_m$ is idempotent and symmetric.

Remark 2. *The second part of the proposition above can be understood in connection with the result from [14, p. 44] which says that a sufficient condition for (see (1)-(2))*

$$\mathbf{B}_m = \sum_{k=1}^{m} (-1)^{k+1} \sum_{m \geq j_k > j_{k-1} > \cdots > j_1 \geq 1} \mathbf{P}_{s(j_k)}\mathbf{P}_{s(j_{k-1})} \cdots \mathbf{P}_{s(j_1)}$$

to be the orthogonal projector onto $\mathrm{Sp}\left(\bar{\mathbf{x}}_{s(1)}, \ldots, \bar{\mathbf{x}}_{s(m)}\right)$ *is:*

$$\mathbf{P}_{s(i)}\mathbf{P}_{s(j)} = \mathbf{P}_{s(j)}\mathbf{P}_{s(i)} \text{ for all } i, j \in \{1, \ldots, m\}. \tag{6}$$

Remark 3. *In order for (c_3) to be fulfilled, we need to have $m \leq n$.*

At the end of this analysis, we prove the following result:

Proposition 3.3. *(i) If $\nu \in (0, 1)$, then $\mathbf{A}_m^\top \mathbf{A}_m + \mathbf{B}_m^\top \mathbf{B}_m \neq \mathbf{I}$ for all $m \geq 1$.*
(ii) If the conditions $(c_1) - (c_3)$ from Prop. 3.2(ii) are satisfied, then $\mathbf{A}_m^\top \mathbf{A}_m + \mathbf{B}_m^\top \mathbf{B}_m = \mathbf{I}$.

Remark 4. *At the m-th step of MPA, we obtain the estimate $\hat{\mathbf{y}}_m = \mathbf{B}_m\mathbf{y}$ and the error $\mathbf{e}_m = \mathbf{y} - \hat{\mathbf{y}}_m = \mathbf{A}_m\mathbf{y}$. In general, $||\hat{\mathbf{y}}_m||^2 + ||\mathbf{e}_m||^2 \neq ||\mathbf{y}||^2$.*

4. MODIFICATIONS OF IT CRITERIA

Model selection rules like the Akaike Information Criterion (AIC) [15] or the Bayesian Information Criterion (BIC) [16] depend on the norm of the vector of residuals and the number of parameters. When they are used in conjunction with MPA, the only alteration in their formula replaces the number of parameters with df. It is more difficult to modify the criteria in which $||\hat{\mathbf{y}}||$ appears explicitly.

Consider the classical linear regression problem for which the additive noise is i.i.d. zero-mean normal, with unknown variance. Let $\hat{\boldsymbol{\beta}}_\gamma$ denote the estimated vector of linear parameters for a model whose set of regressor variables is γ. We denote the cardinality of γ by $|\gamma|$, and assume that $|\gamma| > 0$. This means that we exclude the possibility that $\mathbf{y}$ is pure noise. The definition of stochastic complexity (SC) is [17]:

$$\begin{aligned}\mathrm{SC}(\mathbf{y};\gamma) &= n\ln\frac{||\mathbf{e}_\gamma||^2}{n} \\ &+|\gamma|\ln\frac{||\hat{\mathbf{y}}_\gamma||^2/|\gamma|}{||\mathbf{e}_\gamma||^2/(n-|\gamma|)}+\ln\frac{|\gamma|}{(n-|\gamma|)^{n-1}}, \quad (7)\end{aligned}$$

where $\hat{\mathbf{y}}_\gamma = \mathbf{X}\hat{\boldsymbol{\beta}}_\gamma$, $\mathbf{e}_\gamma = \mathbf{y} - \hat{\mathbf{y}}_\gamma$ and $\ln(\cdot)$ is the natural logarithm. The "best" model is the one which minimizes $\mathrm{SC}(\mathbf{y};\gamma)$.

A similar criterion, called generalized Minimum Description Length (gMDL), was introduced in [18] :

$$\mathrm{gMDL}(\mathbf{y};\gamma) = \\ \begin{cases}\frac{n}{2}\ln\frac{||e_\gamma||^2}{n-|\gamma|}+\frac{|\gamma|}{2}\ln\frac{||\hat{\mathbf{y}}_\gamma||^2/|\gamma|}{||\mathbf{e}_\gamma||^2/(n-|\gamma|)}+\ln n, & \text{if } \frac{||\hat{\mathbf{y}}_\gamma||^2}{||\mathbf{y}||^2}\geq\frac{|\gamma|}{n} \\ \frac{n}{2}\ln\frac{||\mathbf{y}^2||}{n}+\frac{1}{2}\ln n, & \text{otherwise}\end{cases} \quad (8)$$

Reference [19] reports the results of an empirical study on model selection for high-dimensional linear models, where gMDL is used together with the forward selection algorithm (FSA). A comparison of FSA and OMP can be found in [2, Sec. 12.7.1.1]. It is mentioned in [20, Sec. 5.4] that gMDL can be used for selecting the models produced at various steps of MPA.

Before discussing the modification of the two criteria, we mention that the superiority of SC and gMDL to other selection rules comes from the term $|\gamma|\ln F_\gamma = |\gamma|\ln\frac{||\hat{\mathbf{y}}_\gamma||^2/|\gamma|}{||\mathbf{e}_\gamma||^2/(n-|\gamma|)}$ [21, 22]. One remarkable property is that F_γ is F-distributed.

For adapting SC and gMDL to the context of MPA, we propose two different sets of modifications:

$$|\gamma| \mapsto \mathrm{df}_m, ||\mathbf{e}_\gamma||^2 \mapsto ||\mathbf{e}_m||^2, ||\hat{\mathbf{y}}_\gamma||^2 \mapsto ||\hat{\mathbf{y}}_m||^2, \quad (9)$$

$$|\gamma| \mapsto \mathrm{df}_m, ||\mathbf{e}_\gamma||^2 \mapsto ||\mathbf{e}_m||^2, ||\hat{\mathbf{y}}_\gamma||^2 \mapsto ||\mathbf{y}||^2 - ||\mathbf{e}_m||^2. \quad (10)$$

The modifications presented in (10) are suggested by the result of Prop. 3.3. Even if we apply the modifications in (9), the resulting $F_m = \frac{||\hat{\mathbf{y}}_m||^2/\mathrm{df}_m}{||\mathbf{e}_m||^2/(n-\mathrm{df}_m)}$ is not F-distributed because $\mathbf{B}_m$ is not idempotent. This can be seen from the conditions of a well-known theorem, which can be found, for example, in [22, Th. A1]. Use of an approximate F-distribution for the terms of this type, when the hat matrix is not idempotent, is addressed in [5]. Their work is of special interest because it is done in the context of modifying the AIC selection rule for the case when the hat matrix is neither idempotent nor symmetric. The resulting criterion is dubbed AIC_{C_1}. Interestingly enough, two other criteria are proposed in [5]: AIC_{C_0}, which is difficult to evaluate because it involves some integrals and AIC_C, which is very popular in the statistical literature:

$$\mathrm{AIC}_C(\mathbf{y};m) = \ln\frac{||\mathbf{e}_m||^2}{n}+\frac{1+\mathrm{df}_m/n}{1-(\mathrm{df}_m+2)/n}.$$

AIC_C is derived by simply replacing the number of linear parameters with df_m in the formula of bias-corrected AIC from [23]. This encourages us to apply the modifications in (9) and (10) to both SC and gMDL.

Another criterion which has been modified by using df instead of the number of linear parameters is BIC [2, Sec. 2.11]:

$$\mathrm{BIC}(\mathbf{y};m) = n\ln\frac{||\mathbf{e}_m||^2}{n}+\mathrm{df}_m\ln n. \quad (11)$$

The expression above can be further modified by adding one more term:

$$\mathrm{EBIC}(\mathbf{y};m) = \mathrm{BIC}(\mathbf{y};m)+2\zeta\ln\binom{p_n}{s_m}, \quad (12)$$

where $\zeta \in (0,1]$ and s_m is the number of non-zero entries of $\hat{\boldsymbol{\beta}}_m$. This Extended Bayesian Information Criterion (EBIC) is inspired by the work in [24]. The additional term represents the cost of listing the indexes of variables included in the model. It can be neglected when p_n is small, but becomes important when the original set of predictors is large. The only difference between the formula in (12) and the one in [24, p. 761] stems from the second term in (11): $\mathrm{df}_m\ln n$ instead of $s_m\ln n$.

In the case of the classical linear model for which the number of predictors (p_n) is large, the authors of [25] proposed to add the term $2\ln\binom{p_n}{|\gamma|}$ to the expression of SC in (7). This suggests that, in our case, the criteria in (7)-(8) can be further modified as follows:

$$\mathrm{ESC}_{\mathrm{alt}}(\mathbf{y};m) = \mathrm{SC}_{\mathrm{alt}}(\mathbf{y};m)+2\zeta\ln\binom{p_n}{s_m}, \quad (13)$$

$$\mathrm{EgMDL}_{\mathrm{alt}}(\mathbf{y};m) = \mathrm{gMDL}_{\mathrm{alt}}(\mathbf{y};m)+\zeta\ln\binom{p_n}{s_m}, \quad (14)$$

where $\mathrm{alt} = 1, 2$. The notation $\mathrm{SC}_1(\cdot)$ is employed for the criterion which results after applying the alteration in (9) to the formula in (7). Similarly, $\mathrm{SC}_2(\cdot)$ is the criterion produced by altering (7) with (10). The same notional conventions are applied for $\mathrm{gMDL}_{\mathrm{alt}}(\cdot)$.

ISBN 978-0-9928626-7-1 © EURASIP 2017

It is clear that a larger value for ζ in (13)-(14) reduces the probability of including spurious predictors in the model, but at the same time might diminish the probability of selecting the "true" predictors. As we know from Prop. 3.1 that $\mathrm{df}_m \leq m\nu$, it is reasonable to impose the condition that $\zeta \in (0, \nu]$. In our experiments, we take $\zeta = \nu$.

5. NUMERICAL RESULTS

The experimental settings are similar to those in [4, Sec. 3.3]; the sample size is $n = 100$, and the dictionary is complete ($p_n = 100$). The entries of the $n \times p_n$ matrix which represents the dictionary are i.i.d. standard normal random variables. The response vector is given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the entries of $\boldsymbol{\beta}$ are chosen according to four different models: **Model 1** (low-dimensional): $\beta_1 = \beta_2 = \beta_3 = 1/3$ and $\beta_q = 0$ for $4 \leq q \leq p_n$; **Model 2** (high-dimensional, small equal coefficients): $\beta_q = p_n^{-1}$ for $1 \leq q \leq p_n$; **Model 3** (high-dimensional, decaying coefficients): $\beta_q = q^{-1}$ for $1 \leq q \leq p_n$; **Model 4** (high-dimensional, slowly decaying coefficients): $\beta_q = q^{-1/2}$ for $1 \leq q \leq p_n$.

The $\boldsymbol{\varepsilon}$-vector is obtained by multiplying with κ/ς a column vector $\tilde{\boldsymbol{\varepsilon}}$ whose n entries are i.i.d. standard normal random variables. Note that $\kappa = \left[\frac{\mathrm{Var}(\mathbf{X}\boldsymbol{\beta})}{\mathrm{Var}(\tilde{\boldsymbol{\varepsilon}})}\right]^{1/2}$ and ς is a parameter which controls the signal-to-noise ratio (SNR). Following [4, Sec. 3.3], we take $\varsigma^2 = 8$ for high SNR and $\varsigma^2 = 0.2$ for low SNR. For each model and for each value of ς^2, we simulate $\mathrm{N}_{TR} = 100$ data sets. In order to investigate how the greediness of MPA impacts on the accuracy of the estimation, we use for each data set a large and a small value for ν (0.95 and 0.1). The vector $\mathbf{y}$ and the columns of $\mathbf{X}$ are centred. Additionally, the columns of $\mathbf{X}$ are standardised such that all the diagonal entries of $(\mathbf{X}^\top\mathbf{X})/n$ are equal to one.

In all cases, the upper bound on the number of iterations is $m_{\mathrm{ub}} = 20000$. Because of the way in which the expression of AIC_C depends on df, we end the iterations before df equals $n - 2$. An additional rule is applied such that MPA is stopped after the number of distinct selected predictors becomes equal to p_n.

For testing the predictive power of each IT criterion, we use the same method as in [4, Sec. 3.3]: For each trial, the same algorithm as the one used to generate the dictionary is applied in order to produce a matrix $\mathbf{X}_{\mathrm{out},r}$ whose size is $(10n) \times p_n$. If $\hat{\boldsymbol{\beta}}_r^{\mathrm{ITC}}$ is the vector of linear parameters corresponding to the model selected by a particular IT criterion, in trial r, then we compute the mean integrated square error as follows [4, Sec. 3.3]:

$$\mathrm{MISE} = \frac{\sum_{r=1}^{\mathrm{N}_{TR}} \left\| \mathbf{X}_{\mathrm{out},r}\boldsymbol{\beta} - \mathbf{X}_{\mathrm{out},r}\hat{\boldsymbol{\beta}}_r^{\mathrm{ITC}} \right\|^2}{(10n) \times \mathrm{N}_{TR}}, \quad (15)$$

where $\boldsymbol{\beta}$ is defined for each model in the description above. Note that, for all $1 \leq r \leq \mathrm{N}_{TR}$, the columns of $\mathbf{X}_{\mathrm{out},r}$ are centred. We show in Table 1 the values of MISE, computed for the IT criteria which are evaluated in our study. We have also conducted experiments for which $n = 20$ and all other settings are the same as above. The full results of these experiments are reported in [13, Table 1].

6. FINAL REMARKS

The empirical results confirm what has already been pointed out previously: We recommend choosing a small value for ν, especially when SNR is low. We made efforts to follow as accurately as possible the experimental settings from [4]. The major difference between it and our implementation is the use of IT criteria instead of cross-validation. For Models 1-3, our results are better or comparable with those reported in [4], but are worse in the case of the high-dimensional model with slowly decaying coefficients. However, cross-validation is much more computationally intensive than model selection based on IT criteria.

We can draw the following conclusions about the use of various IT criteria: (a) Applying the alteration in (9) to SC leads to similar results as in the case when SC is altered with (10). The same is true for gMDL. In general, the variants of gMDL behave similarly to the corresponding variants of SC; (b) The most intriguing alteration is the one which adds the weighted logarithm of the binomial coefficient to the expressions of $\mathrm{BIC}(\mathbf{y}; m)$, $\mathrm{SC}_{\mathrm{alt}}(\mathbf{y}; m)$ and $\mathrm{gMDL}_{\mathrm{alt}}(\mathbf{y}; m)$: In some cases, it deteriorates the results because it does not act as a penalty term when the number of selected predictors is greater than $p_n/2$; (c) Some of the altered variants of SC and gMDL are superior to AIC_C when $n = 20$ (see Table 1 in the supplemental material).

All experiments can be reproduced by using the code from `https://www.stat.auckland.ac.nz/%7Ecgiu216/`.

7. REFERENCES

[1] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[2] P. Bühlmann and S. van de Geer, *Statistics for high-dimensional data. Methods, theory and applications*, Springer-Verlag, 2011.

[3] A.R. Barron, A. Cohen, W. Dahmen, and R.A. DeVore, "Approximation and learning by greedy algorithms," *Annals of Statistics*, vol. 36, pp. 64–94, 2008.

[4] A. Sancetta, "Greedy algorithms for prediction," *Bernoulli*, vol. 22, pp. 1227–1277, 2016.

[5] C.M. Hurvich and J.S. Simonoff, "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion," *Journal of the Royal Statistical Society: Series B*, vol. Part 2, pp. 271–293, 1998.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning. Data mining, inference, and prediction*, Springer Science+Business Media, 2 edition, 2008.

ς^2	ν	SC$_1$	SC$_2$	ESC$_1$	ESC$_2$	gMDL$_1$	gMDL$_2$	EgMDL$_1$	EgMDL$_2$	BIC	EBIC	AIC$_C$
Model 1												
8.0	0.95	0.0139	0.0137	**0.0063**	**0.0062**	0.0146	0.0145	**0.0063**	**0.0063**	0.0509	0.0777	0.0254
8.0	0.10	0.0092	**0.0091**	**0.0087**	**0.0087**	0.0094	0.0092	**0.0088**	**0.0087**	0.0229	0.0282	0.0180
0.2	0.95	1.2932	1.2363	7.9890	7.9418	1.3030	1.2513	7.9890	8.0400	1.6981	2.7463	**0.9629**
0.2	0.10	0.9374	0.9539	1.0601	1.2734	0.9495	0.9634	1.0766	1.3040	0.6764	0.8658	**0.6330**
Model 2												
8.0	0.95	0.0080	0.0080	**0.0073**	**0.0073**	0.0080	0.0080	**0.0073**	**0.0073**	0.0079	0.0079	0.0085
8.0	0.10	**0.0062**	**0.0061**	**0.0061**	**0.0059**	**0.0062**	**0.0061**	**0.0061**	**0.0059**	0.0065	0.0065	**0.0062**
0.2	0.95	0.0473	0.0468	0.2407	0.2433	0.0475	0.0477	0.2407	0.2420	0.0656	0.0795	**0.0354**
0.2	0.10	0.0314	0.0352	0.0384	0.0549	0.0317	0.0353	0.0389	0.0555	0.0274	0.0316	**0.0230**
Model 3												
8.0	0.95	**0.22**	**0.22**	**0.23**	**0.23**	**0.22**	**0.22**	**0.23**	**0.23**	0.49	0.81	**0.23**
8.0	0.10	**0.18**	**0.18**	**0.18**	**0.18**	**0.18**	**0.18**	**0.18**	**0.18**	0.31	0.34	0.19
0.2	0.95	6.69	6.53	37.62	33.71	6.73	6.60	37.36	36.83	7.91	9.45	**4.92**
0.2	0.10	4.54	4.72	5.42	6.89	4.62	4.88	5.73	7.08	4.39	4.69	**3.25**
Model 4												
8.0	0.95	**2.79**	**2.78**	**2.89**	**2.89**	**2.78**	**2.77**	**2.89**	**2.89**	3.25	3.35	**2.79**
8.0	0.10	**1.98**	**1.97**	**1.96**	**1.94**	**1.98**	**1.96**	**1.96**	**1.94**	2.42	2.51	**1.94**
0.2	0.95	22.47	22.89	118.13	119.63	22.92	22.57	118.13	118.47	26.04	29.26	**18.81**
0.2	0.10	16.04	16.85	17.12	19.27	16.18	16.99	17.34	20.27	**10.65**	12.41	12.50

Table 1. MISE computed for various IT criteria by applying the formula in (15) when $n = 100$. For each row of the table, we show in bold the results which are within a range of 5% from the minimum value on that row.

[7] T. Hastie, "Comment: Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, vol. 22, pp. 513–515, 2007.

[8] P. Bühlmann, "Boosting for high-dimensional linear models," *The Annals of Statistics*, pp. 559–583, 2006.

[9] R. Tibshirani, "Regression analysis and selection via the Lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, pp. 267–288, 1996.

[10] C. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, pp. 1135–1151, 1981.

[11] J. Ye, "On measuring and correcting the effects of data mining and model selection," *Journal of the American Statistical Association*, pp. 120–131, 1998.

[12] B. Efron, "The estimation of prediction error: covariance penalties and crossvalidation," *Journal of the American Statistical Association*, pp. 619–632, 2004.

[13] F. Li, C.M. Triggs, B. Dumitrescu, and C.D. Giurcăneanu, "Supplemental material to: "On the number of iterations for the matching pursuit algorithm," `https://www.stat.auckland.ac.nz/%7Ecgiu216/PUBLICATIONS.htm`, 2017.

[14] H. Yanai, K. Takeuchi, and Y. Takane, *Projection matrices, generalized inverse matrices, and singular value decomposition*, Springer Science+Business Media, 2011.

[15] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. AC-19, pp. 716–723, 1974.

[16] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[17] J. Rissanen, *Information and complexity in statistical modeling*, Springer Verlag, 2007.

[18] M Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, pp. 746–774, 2001.

[19] G.V. Rocha and B. Yu, "Greedy and relaxed approximations to model selection: A simulation study," in *Festschrift in Honor of Jorma Rissanen on the occasion of his 75th birthday*, P. Grünwald, P. Myllymäki, I. Tabus, M. Weinberger, and B. Yu, Eds., vol. TICSP series #38, pp. 63–80. Tampere International Center for Signal Processing, 2008.

[20] P. Bühlmann and T. Hothorn, "Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, vol. 22, pp. 477–505, 2007.

[21] M Hansen and B. Yu, "Minimum description length model selection criteria for generalized linear models," in *Science and statistics: A festchrift for Terry Speed*, D. Goldstein, Ed., vol. 40, pp. 145–164. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2002.

[22] C.D. Giurcăneanu and S.A. Razavi, "New insights on stochastic complexity," in *Proc. 17th European Signal Processing Conference (Eusipco 2009)*, Glasgow, Scotland, 2009, pp. 2475–2479.

[23] C.M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, pp. 297–307, 1989.

[24] J. Chen and Z. Chen, "Extended Bayesian information criteria for model with large model spaces," *Biometrika*, vol. 95, pp. 759–771, 2008.

[25] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," *IEEE Transactions on Signal Processing*, vol. 57, pp. 3347–3360, 2009.