Signal Processing 155 (2019) 170–181

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

The matching pursuit algorithm revisited: A variant for big data and new stopping rules

Fangyao Li [a], Christopher M. Triggs [a], Bogdan Dumitrescu [b], Ciprian Doru Giurcăneanu [a,*]

[a] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[b] University Politehnica of Bucharest, 313 Spl. Independenței, Bucharest 060042, Romania

ARTICLE INFO

Article history:
Received 22 March 2018
Revised 19 September 2018
Accepted 22 September 2018
Available online 24 September 2018

Keywords:
Matching pursuit algorithm
Hat matrix
Big data
Information theoretic criteria
Air pollution data

ABSTRACT

The matching pursuit algorithm (MPA) is used in many applications for selecting the best predictors for a vector of measurements of size n from a dictionary that contains p_n atoms, where usually $n \leq p_n$. A major unsolved problem is to determine the optimal stopping rule. In this work, we investigate various stopping rules which are modifications of the information theoretic (IT) criteria derived for Gaussian linear regression. Because all of them involve the degrees of freedom (df) given by the trace of the hat matrix, we provide some theoretical results concerning this matrix. We also propose novel stopping rules. An important contribution of this paper is a method for computing the df efficiently when big data ($n \gg p_n$) are processed. The significance of the auxiliary variables appearing in MPA for big data is clarified via a theoretical analysis. The superiority of the new stopping rules in comparison with the traditional approaches is demonstrated in simulations involving big data ($n \gg p_n$) or overcomplete dictionaries ($n < p_n$) and in experiments with air pollution data.

1. Introduction

1.1. Motivation

An important problem in multivariate signal processing is the prediction of a particular entry of the vector random process $\{\mathbf{y}(t)\}$ by using the past measurements as well as the current measurements available for the other entries of the vector (see, for example, [1] and the references therein). The problem can be easily solved by applying the techniques for the identification of autoregressive models with exogenous input [2]. The most difficult part is the selection of the best possible predictors from the existing set of observations. In many practical applications, a large number of past samples are available and this restrains the use of the full-search approach during the training phase when the predictors are chosen.

The computational effort for selecting the predictors can be reduced significantly by applying greedy algorithms [3]. From this family of algorithms, we are especially interested in the matching pursuit algorithm (MPA), which is extensively used in signal processing [4], statistics [5], and approximation theory [6]. At each iteration, MPA yields a linear model for the response vector $\mathbf{y}$ of size n; each such model is a linear combination of some of the entries of a given set of p_n predictors. Theoretical results on the performance of MPA have been recently proven [7] under the hypotheses that (i) p_n grows very fast when n increases and (ii) the predictors are not independent.

The number of iterations for MPA can be as large as $m_{\text{ub}} = 20,000$ and a different model is created at each iteration. The outcome of the algorithm is the model deemed to be "the best" with respect to the selection rule. Because a selection rule decides the outcome of MPA, it is often called the *stopping rule*. An open problem concerns the stopping rule that should be applied as the use of cross-validation (CV) is computationally intensive when the number of iterations, m_{ub}, is large [3].

In our conference paper [8], we have investigated the performance of eleven stopping rules based on different information theoretic (IT) criteria. All of them have been derived from selection rules previously applied in classical linear regression. Another common feature is the presence of the degrees of freedom (df) in their expressions. According to the definition [9], df is evaluated as the trace of the linear operator mapping $\mathbf{y}$ to $\hat{\mathbf{y}}$, where $\hat{\mathbf{y}}$ is the estimate of $\mathbf{y}$ produced by a certain model. This linear operator is known as the hat matrix. Importantly [10], there is empirical evidence that the trace-based computation may underestimate the value of df.

* corresponding author.
E-mail addresses: lfan523@aucklanduni.ac.nz (F. Li), cm.triggs@auckland.ac.nz (C.M. Triggs), bogdan.dumitrescu@acse.pub.ro (B. Dumitrescu), c.giurcaneanu@auckland.ac.nz (C.D. Giurcăneanu).

https://doi.org/10.1016/j.sigpro.2018.09.033
0165-1684/

1.2. Contributions

After presenting MPA in Section 2, we outline the following results in the rest of the paper.

In Section 3, we briefly discuss the IT criteria which are currently used in conjunction with MPA and introduce new stopping rules. The new formulae are based on the properties of the hat matrix that are presented in Appendix A. We show that, in general, the hat matrix is not a projector and give an upper bound on the increase of df from the mth iteration of the algorithm to the $(m+1)$th iteration. These results were stated without proof in [8].

It has been already pointed out in [7] that, because of the massive amount of data produced nowadays, a formulation of MPA for $n \gg p_n$ is really needed. Re-writing the algorithm for the big data case is straightforward and it was already done in [7], but the most difficult part is the calculation of df at each iteration. In Theorem 1, we demonstrate how this can be done efficiently. There is no result similar to Theorem 1 in the previous literature. In Section 4 we perform a theoretical analysis that clarifies the significance of the auxiliary variables appearing in the formulation of MPA for big data, but not in the classical formulation of the algorithm.

In Section 5 we present the results of an extensive empirical study which shows the superiority of the newly introduced IT criteria. In experiments with air pollution data, the new criteria work better than CV. A more comprehensive discussion of the theoretical and empirical results obtained in this work can be found in Section 6.

1.3. Notation

Bold letters denote both vectors and matrices; $\mathbf{I}$ denotes the identity matrix of appropriate size, while $\mathbf{0}$ denotes the vector/matrix whose entries are all equal to zero. The symbol x_a stands for the ath entry of a vector $\mathbf{x}$. If $\mathbf{X}$ is a matrix, then $\mathbf{X}_{a:}$ is the ath row of $\mathbf{X}$, $\mathbf{X}_{:b}$ is the bth column of $\mathbf{X}$, and x_{ab} denotes the entry of $\mathbf{X}$ located in the ath row and the bth column. The operator for transposition is $(\cdot)^\top$; the Euclidean norm of a vector $\mathbf{x}$ is $\|\mathbf{x}\|$; the operator $\odot$ is employed for the element-wise product of vectors. For an arbitrary matrix $\mathbf{X}$, $\mathrm{Sp}(\mathbf{X})$ denotes the linear subspace spanned by the columns of $\mathbf{X}$ and $\mathrm{Ker}(\mathbf{X})$ is the null space of $\mathbf{X}$.

2. The matching pursuit algorithm

2.1. Description

Assume that the response vector $\mathbf{y} = [y_1, \ldots, y_n]^\top$ is given, as well as the matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_{p_n}]$ of potential predictors, which is called dictionary. 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 by $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. In the initialization phase of the algorithm, the vector $\mathbf{y}$ and the columns of $\mathbf{X}$ are centred, and $\hat{\boldsymbol{\beta}}$ 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 jth step of the algorithm, the column of $\mathbf{X}$ indexed by $s(j)$ is selected, where $1 \le s(j) \le p_n$. Then, only the $s(j)$th entry of $\hat{\boldsymbol{\beta}}$ is updated by using the formula $\hat{\beta}_{s(j)} \leftarrow \hat{\beta}_{s(j)} + \nu(\mathbf{x}_{s(j)}^\top \mathbf{x}_{s(j)})^{-1}\mathbf{x}_{s(j)}^\top \mathbf{e}$. MPA can be seen as a coordinate descent on the objective $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, the chosen coordinate corresponding to the largest element of the gradient.

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) which 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 [3, 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 [3, Sec. 12.6.2.1] by emphasizing the relationship between MPA and the well-known Lasso algorithm [11]. 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.

2.2. Hat matrix

Let $\hat{\mathbf{y}}_m = \mathbf{X}\hat{\boldsymbol{\beta}}_m$ be the estimate of $\mathbf{y}$ obtained after the mth step of the algorithm. We denote by $\mathbf{B}_m$ the linear operator, named the hat-matrix, which maps $\mathbf{y}$ to $\hat{\mathbf{y}}_m$:

$$\hat{\mathbf{y}}_m = \mathbf{B}_m \mathbf{y}. \tag{1}$$

Recalling that $\mathbf{x}_{s(j)}$ denotes the predictor selected at the jth iteration of MPA, $\mathbf{B}_m$ is expressed as [12] (see also the discussion in [5, Sec. 5.3]):

$$\mathbf{B}_m = \mathbf{I} - \mathbf{A}_m, \text{ where} \tag{2}$$

$$\mathbf{A}_m = \left(\mathbf{I} - \nu\mathbf{P}_{s(m)}\right)\cdots\left(\mathbf{I} - \nu\mathbf{P}_{s(1)}\right), \tag{3}$$

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

$$\mathbf{A}_m = \sum_{k=0}^{m} \mathbf{S}_{m,k}, \text{ where } \mathbf{S}_{m,0} = \mathbf{I} \tag{4}$$

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

$$\mathbf{S}_{m,k} = (-\nu)^k \sum_{m \ge j_k > j_{k-1} > \cdots > j_1 \ge 1} \mathbf{P}_{s(j_k)}\mathbf{P}_{s(j_{k-1})}\cdots\mathbf{P}_{s(j_1)}. \tag{5}$$

The matrix $\mathbf{B}_m$ is important in evaluating the complexity of the linear model produced at the mth step. More precisely, the degrees of freedom for the fitted model are estimated by

$$\mathrm{df}_m = \mathrm{tr}(\mathbf{B}_m). \tag{6}$$

This formula has been used, for example, in [9]. It follows from Stein's theory on unbiased risk estimation [13] that for the case when the design matrix is fixed and the residuals are i.i.d. Gaussian, with zero-mean and known variance σ^2, $\mathrm{df} = \sum_{j=1}^{n} \mathrm{Cov}(\hat{y}_j, y_j)/\sigma^2$ [14,15]. It is a simple exercise to demonstrate that this expression equals the trace of the hat matrix (see [3, 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 produced during these iterations. Because of the particularities of MPA, the IT criteria that have been previously derived for the classical linear model cannot be applied in their original form [12]. The modifications of the criteria are discussed in Section 3. They are based on the properties of the hat matrix outlined in Appendix A.

3. Modified IT criteria

We consider the classical linear regression problem for which the additive noise is i.i.d. zero-mean Gaussian, 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. The vector of residuals is $\mathbf{e}_\gamma = \mathbf{y} - \hat{\mathbf{y}}_\gamma$, where $\hat{\mathbf{y}}_\gamma$ is the estimate calculated by using $\hat{\boldsymbol{\beta}}_\gamma$. 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. An