Journal of Multivariate Analysis 133 (2015) 51–60

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

# Nonparametric significance testing and group variable selection

Adriano Zanin Zambom [a,*], Michael G. Akritas [b]

[a] Department of Statistics, State University of Campinas (UNICAMP), Brazil

[b] Department of Statistics, The Pennsylvania State University, USA

ARTICLE INFO

*Article history:*
Received 20 July 2013
Available online 4 October 2014

*AMS subject classification:*
62Gxx

*Keywords:*
Backward elimination
Dimension reduction
Heteroscedasticity
Lack-of-fit tests
Local polynomial regression
Nonparametric regression

ABSTRACT

In the context of a heteroscedastic nonparametric regression model, we develop a test for the null hypothesis that a subset of the predictors has no influence on the regression function. The test uses residuals obtained from local polynomial fitting of the null model and is based on a test statistic inspired from high-dimensional analysis of variance. Using $p$-values from this test, and multiple testing ideas, a group variable selection method is proposed, which can consistently select even groups of variables with diminishing predictive significance. A backward elimination version of this procedure, called GBEAMS for Group Backward Elimination Anova-type Model Selection, is recommended for practical applications. Simulation studies, suggest that the proposed test procedure outperforms the generalized likelihood ratio test when the alternative is non-additive or there is heteroscedasticity. Additional simulation studies reveal that the proposed group variable selection procedure performs competitively against other variable selection methods, and outperforms them in selecting groups having nonlinear or dependent effects. The proposed group variable selection procedure is illustrated on a real data set.

## 1. Introduction

Advances in data collection technologies and data storage devices have enabled the collection of large data sets, involving a large number of observations on many variables, in several disciplines. When the objective is to build a predictive model, the challenges presented by high dimensional data sets have opened new frontiers for statistical research. Including insignificant predictors results in complicated models with less predictive power and reduced ability to discern and interpret the influence of the predictors. The underlying principles of modern model building are parsimony and sparseness. As such, variable selection is a fundamental component of model building, and has received extensive attention over the last 20 years.

Due to readily available software, variable selection is often performed by modeling the expected response at covariate value $\mathbf{x}$ as $m(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$. Classical approaches to variable selection, such as stepwise selection or elimination procedures, and best subset variable selection, can be computationally intensive or ignore stochastic errors. A new class of methodologies addresses variable selection through minimization of a constrained or penalized objective function, such as Tibshirani's [26] Lasso, Fan and Li's [15] SCAD, Efron, Hastie, Johnstone, and Tibshirani's [13] least angle regression, Zou's [36] adaptive Lasso and Candes and Tao's [10] Dantzig selector. A different approach exploits the conceptual connection between model testing and variable selection: dropping variable $j$ from the model is equivalent to not rejecting the null hypothesis $H_0^j : \beta_j = 0$.

[*] Corresponding author.
*E-mail addresses:* adriano.zambom@gmail.com, zambom@ime.unicamp.br (A.Z. Zambom).

http://dx.doi.org/10.1016/j.jmva.2014.08.014
0047-259X/© 2014 Elsevier Inc. All rights reserved.

Abramovich, Benjamini, Donoho and Johnstone [1] showed that application of the false discovery rate (FDR) controlling procedure of Benjamini and Hochberg [6] on $p$-values resulting from testing each $H_0^j$ can be translated into minimizing a model selection criterion similar to that used in [27] and others. Working with orthogonal designs, they also showed that their method is asymptotically minimax for $\ell^r$ loss, $0 < r \leq 2$, simultaneously throughout a range of sparsity classes, provided the level $q$ for the false discovery rate is set to $q < 0.5$. Generalizations of this methodology to non-orthogonal designs differ mainly in the generation of the $p$-values for testing $H_0^j : \beta_j = 0$, and the false discovery rate method employed. Bunea, Wegkamp and Auguste [9] use $p$-values generated from the standardized regression coefficients resulting from fitting the full model and employ the Benjamini and Yekutieli [8] method for controlling false discovery rate under dependency, while Benjamini and Gavrilov [5] use $p$-values from a forward selection procedure where the $i$th stage $p$-to-enter is the $i$th stage constant in the multiple-stage false discovery rate procedure in [7].

Model checking and variable selection procedures based on the assumption of a linear model may fail to discern the relevance of covariates whose effect on $m(\mathbf{x})$ is nonlinear. Because of this, procedures for both model checking and variable selection have been developed under more general/flexible models. See, for example [20,29,17,25], and references therein. However, the methodological approaches in this literature have been distinct from those of model checking. Zambom [33] and Zambom and Akritas [34] develop a consistent variable selection procedure using $p$-values from testing the residual significance of each variable in a heteroscedastic nonparametric model.

In many applications, covariates come in groups (see Section 4 for an example), and the issue of selecting the groups of variables with predictive significance arises. The most common group selection procedures are the group Lasso [32], and the adaptive group Lasso [29]. Alternative methods with concave penalties have also been used, as for example the group SCAD. Zhu and Li [35] develop an alternative method based on nonlinear dimension reduction which is shown to perform well in nonparametric additive models. See also [23] who, using averages of the genes within each group, perform selection based on a procedure combining hierarchical clustering and Lasso. To our knowledge, test-based variable selection procedures have not been extended to group selection. In principle, this extension is straightforward: the $p$-values for the significance of each individual covariate are replaced by the $p$-values for testing the significance of each group of variables. For the Bunea, Wegkamp and Auguste [9] procedure, such an extension is automatic since $p$-values for groups of variables are readily available in the context of a linear model. Moreover, if the groups of variables are orthogonal, or nearly orthogonal, such an extension of the Bunea et al. [9] procedure is expected to satisfy optimality properties similar to those of Abramovich, Benjamini, Donoho and Johnstone [1] (under sparseness of the groups of variables; see Section 3). In this paper we consider a heteroscedastic nonparametric model, and extend the test-based variable selection method of Zambom and Akritas [34].

Specifically, the present paper achieves two objectives. First, it develops a procedure for testing the residual predictive significance of a group of variables, in a nonparametric heteroscedastic model. The term "predictive significance" is used to highlight the fact that the test procedure is designed to detect the effect of the variables on the regression function, while maintaining its level if the variables only affect other aspects of the conditional distribution of the response variable, such as the variance function. As with the test in [34], the present test is modeled after the high-dimensional one-way ANOVA test of Akritas and Papadatos [2]. The main innovation in the present test lies in the construction of the "augmented cells", which is the most critical part for the construction of a successful test. Simulations suggest that the proposed test achieves competitive power, and accurate type I error rates under heteroscedasticity, in accordance with its asymptotic properties. Secondly, the paper introduces GBEAMS (Group Backward Elimination Anova-type Model Selection), a backward elimination procedure for group variable selection using the Benjamini and Yekutieli [8] method applied on the $p$-values resulting from testing the predictive significance of each group. A group selection version of the consistency result in [34] can be established along the same lines. Additional simulations suggest that the proposed group variable selection procedure performs competitively, and outperforms group Lasso and group SCAD in selecting groups having nonlinear effects.

The paper is organized as follows. Section 2 describes the proposed methodology for testing the predictive significance of a group of variables, derives the asymptotic distribution of the test statistic, under the null hypothesis and local alternatives, and presents results of simulation studies comparing its performance to that of the generalized likelihood ratio test of Fan and Jiang [14]. Section 3 describes the test-based group variable selection procedure, and presents results of simulation studies comparing its performance to that of group Lasso and the procedure introduced in [35]. The analysis of a real data set involving gene expression levels of healthy and cancerous colon tissues is presented in Section 4.

## 2. Nonparametric model checking

### 2.1. The hypothesis and the test statistic

Assume we have $n$ observations, $(Y_i, \mathbf{U}_i)$, $i = 1, \ldots, n$, of the response variable $Y$ and covariates $\mathbf{U} = (\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ and $\mathbf{Z}$ have dimensions $r$ and $s$ respectively ($r + s = d$), which remain fixed as $n \to \infty$. Let $m(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ denote the regression function. The heteroscedastic nonparametric regression model is

$$Y = m(\mathbf{X}, \mathbf{Z}) + \sigma(\mathbf{X}, \mathbf{Z})\epsilon, \tag{1}$$

where $\epsilon$ has zero mean and constant variance and is uncorrelated from $\mathbf{X}$ and $\mathbf{Z}$. The goal is to test the null hypothesis that $\mathbf{Z}$ does not contribute to the regression function, i.e.

$$H_0 : m(\mathbf{x}, \mathbf{z}) = m_1(\mathbf{x}). \tag{2}$$