Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 136 (2006) 4349–4364
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Consistent variable selection in high dimensional regression via multiple testing

Florentina Bunea*, Marten H. Wegkamp, Anna Auguste

Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA

Received 29 May 2004; accepted 29 March 2005
Available online 3 August 2005

Abstract

This paper connects consistent variable selection with multiple hypotheses testing procedures in the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, where the dimension p of the parameter $\boldsymbol{\beta}$ is allowed to grow with the sample size n. We view the variable selection problem as one of estimating the index set $I_0 \subseteq \{1, \ldots, p\}$ of the non-zero components of $\boldsymbol{\beta} \in \mathbb{R}^p$. Estimation of I_0 can be further reformulated in terms of testing the hypotheses $\beta_1 = 0, \ldots, \beta_p = 0$. We study here testing via the false discovery rate (FDR) and Bonferroni methods. We show that the set $\widehat{I} \subseteq \{1, \ldots, p\}$ consisting of the indices of rejected hypotheses $\beta_i = 0$ is a consistent estimator of I_0, under appropriate conditions on the design matrix $\mathbf{X}$ and the control values used in either procedure. This technique can handle situations where p is large at a very low computational cost, as no exhaustive search over the space of the 2^p submodels is required.

Keywords: Bonferroni correction; False discovery rate; Multiple hypothesis testing; Consistent variable selection

1. Introduction

The false discovery rate (FDR) procedure has been developed in the context of multiple hypotheses testing by Benjamini and Hochberg (1995). Given a set of p hypotheses, out of which an unknown number p_0 are true, the FDR method identifies the hypotheses to be rejected, while keeping the expected value of the ratio of the number of false rejections

* Corresponding author.
E-mail addresses: bunea@stat.fsu.edu (F. Bunea), wegkamp@stat.fsu.edu (M.H. Wegkamp), auguste@stat.fsu.edu (A. Auguste).

0378-3758/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.jspi.2005.03.011

to the total number of rejections below q, a user specified control value. In addition, this technique can handle problems in which p is very large at a very low computational cost. The span of its applications ranges from denoising in signal processing problems, see for instance Abramovich et al. (2000), to genetics and medicine, see for instance Storey (2002), Benjamini and Yekutieli (2001). Genovese and Wasserman (2004) discuss theoretical aspects of the procedure using a stochastic process approach.

In this paper we indicate how the FDR procedure can be used for variable selection in linear regression models and establish the consistency of selection. We assume that the data are generated from the model

$$\begin{aligned} &\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ &\beta_j \neq 0, \;\; j \in I_0; \quad \beta_j = 0, \;\; j \in \{1, \ldots, p\} \backslash I_0, \end{aligned} \tag{1.1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\boldsymbol{X}$ is a $n \times p$ design matrix with deterministic entries x_{ij}, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the unknown vector of regression coefficients. The number of predictors $x_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ considered, p, is allowed to grow with the sample size n. This means that as n increases, the model is allowed to become more complex. In addition, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is a vector of independent, identically distributed errors ε_i with

$$\mathbb{E}\varepsilon_i = 0, \quad \mathbb{E}\varepsilon_i^2 = \sigma^2, \quad \mathbb{E}|\varepsilon_i|^{4+\delta} < \infty \quad \text{for some } \delta > 0. \tag{1.2}$$

The consistent variable selection problem is equivalent with the problem of estimating consistently the unknown index set $I_0 \subseteq \{1, \ldots, p\} \stackrel{\text{def}}{=} I_p$ of the non-zero components of $\boldsymbol{\beta}$. This problem received considerable attention in the statistical literature. In particular, the bayesian information criterion (BIC) has been shown to lead to consistent estimators of I_0, see Hannan and Quinn (1979), Hannan (1980), Geweke and Meese (1981) for early references. Woodroofe (1982) and Haughton (1988) establish consistency in the context of exponential families and we refer to Bunea (2004) for a recent contribution in semiparametric regression.

The serious drawback of any model selection method based on a penalized criterion is of a computational nature, as a search through the space of all possible 2^p models may be needed. Cross-validation (Shao, 1993) provides an alternative, but again the leave m out of n strategy requires intensive computation. Zheng and Loh (1995), in the context of linear regression, suggested a two-stage procedure, where the first stage consists of ranking test statistics, and the second stage computing a penalized least squares estimator based on p models only, which is a marked improvement over the other strategies, but may still be suboptimal computationally for p large, which is the case of interest in this paper. Finally, Jiang and Liu (2004) study model selection based on parameter estimation in a more general setting. For instance, they allow for Poisson regression with random effects, Cox regression and graphical models. In the linear regression case, their method is intimately related to Zheng and Loh (1995). However p, the number of predictors, is not allowed to depend on the sample size n. This therefore creates the need for a computationally fast method that consistently estimates I_0 in this case. The approach we take here is based on multiple hypotheses testing. Note that the problem of estimating I_0 can be viewed as testing