Computational Statistics and Data Analysis 81 (2015) 107–120

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A sequential test for variable selection in high dimensional complex data

Kofi P. Adragni *, Moumita Karmakar

Department of Mathematics & Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, United States

ARTICLE INFO

Article history:
Received 16 April 2013
Received in revised form 28 January 2014
Accepted 24 July 2014
Available online 2 August 2014

Keywords:
Inverse regression
Dimension reduction
Principal components
Variable selection
Variable screening
High dimensionality
Data analysis

ABSTRACT

Given a high dimensional p-vector of continuous predictors X and a univariate response Y, principal fitted components (PFC) provide a sufficient reduction of X that retains all regression information about Y in X while reducing the dimensionality. The reduction is a set of linear combinations of all the p predictors, where with the use of a flexible set of basis functions, predictors related to Y via complex, nonlinear relationship can be detected. In the presence of possibly large number of irrelevant predictors, the accuracy of the sufficient reduction is hindered. The proposed method adapts a sequential test to the PFC to obtain a "pruned" sufficient reduction that shed off the irrelevant predictors. The sequential test is based on the likelihood ratio which expression is derived under different covariance structures of $X|Y$. The resulting reduction has an improved accuracy and also allows the identification of the relevant variables.

1. Introduction

Consider the Big-Mac dataset (Enz, 1991), a simple dataset that gives average values in 1991 of several economic indicators for 45 world cities. It has nine continuous predictors and a continuous outcome variable. The outcome is the minimum labor to buy a Big Mac and fries in US dollars. A regression fitting to the raw data without any transformation of the response or predictors yields a multiple R^2, the square of the correlation between the observed and the fitted response to be 0.46. After a graphical exploration and the appropriate transformation of the variables, we obtained $R^2 = 0.87$. The reason of this drastic improvement is that the relationships between the response and the predictors, initially nonlinear (Fig. 1), were transformed into linear through a model fitting procedure guided by diagnostics (see Cook and Weisberg, 1982, Section 1.2). With nine predictors, this procedure is easily doable. However, when p is large, say 50 or more, a regression modeling using this iterative procedure is rather a daunting task, tedious and imponderable. Ubiquitously, a forward linear model is considered, and the relationship between individual predictors and the response is often unexplored because of the high dimensionality of the predictors. Diagnostic methods are seldom used for model checking. Using an ill-fitting model to solve a variable selection problem can result in reduced performance.

Most variable selection methods are constructed around forward linear regression models. Because the ordinary least squares estimation does not yield satisfactory results when p is large, it is often assumed that a large portion of these p predictors is irrelevant in explaining the response Y. The corresponding coefficients of these predictors in a linear regression model are shrunk or even set to zero. This brings the concept of sparsity into regression modeling with two induced consequences: parsimony of the model and accuracy in prediction. A flurry of research on algorithms and theory for variable

[*] Corresponding author.
E-mail addresses: kofi@umbc.edu, kofi.adragni@gmail.com (K.P. Adragni).

http://dx.doi.org/10.1016/j.csda.2014.07.016
0167-9473/© 2014 Elsevier B.V. All rights reserved.

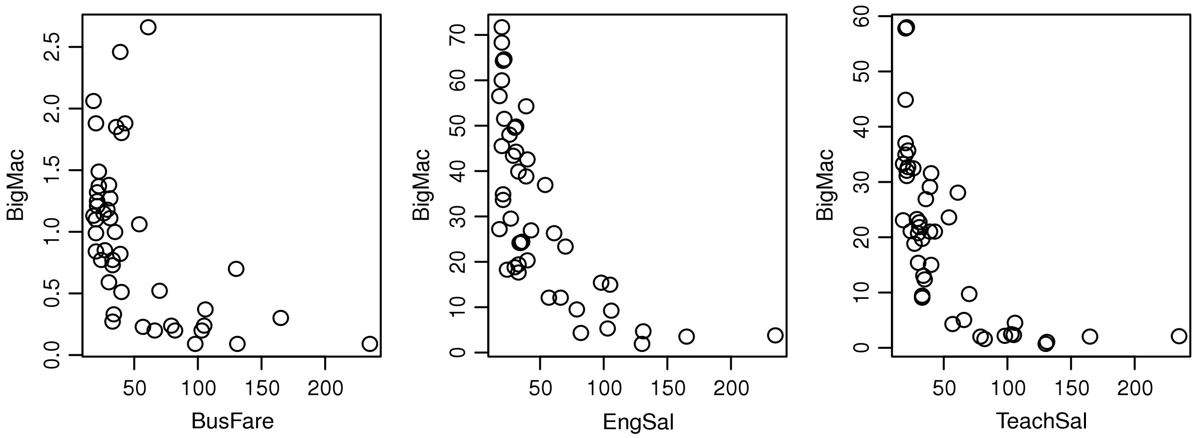

Fig. 1. BigMac partial scatter-matrix plot.

selection involving sparsity constraints have been observed in recent years. These methods include the soft thresholding (Donoho, 1995), the nonnegative garotte (Breiman, 1995), lasso (Tibshirani, 1996), the smoothly clipped absolute deviation penalty (SCAD; Fan and Li, 2001), elastic net (Zou and Hastie, 2005), and Dantzig selector (Candès and Tao, 2007) among many others. These methods work exceptionally well when the model is accurate. However they do not perform adequately when the predictors and the response have an arbitrary non-linear relationship.

A recent methodology proposed by Cook (2007) brings significant openings to address the shortcomings of linear models in capturing information about high dimensional predictors non-linearly related to the response. Cook (2007) proposed the concept of sufficient dimension reduction in regression and set up a new paradigm of dimension reduction through a likelihood-based approach called principal fitted components (PFC). A reduction $R : \mathbb{R}^p \to \mathbb{R}^d$, $d \leq p$, was defined to be sufficient if it satisfies one of the following three statements: (i) $Y|X \sim Y|R(X)$, (ii) $X|(Y, R(X)) \sim X|R(X)$, and (iii) $X \perp\!\!\!\perp Y|R(X)$. The symbol $\perp\!\!\!\perp$ stands for statistical independence, and $U \sim V$ stands for U and V having identical distribution. Statement (i) holds in a forward regression while statement (ii) holds in an inverse regression setup. Under a joint distribution of (Y, X) the three statements are equivalent.

Principal fitted components are a class of inverse regression models that yield a sufficient reduction of the predictors. Let X_y denote the random vector $X|(Y = y)$ and assume that there is a vector-valued function $\nu(Y) \in \mathbb{R}^d$, with $d \ll \min(n, p)$ and $\mathrm{E}[\nu(Y)] = 0$, so that X_y can be represented by the model $X_y \sim N(\mu + \Gamma\nu(y), \Delta)$. The term $\Gamma \in \mathbb{R}^{p\times d}$ is a semi-orthogonal matrix, and $\mu = \mathrm{E}(X)$. The covariance Δ is assumed to be independent of Y. Under this model the translated conditional means $\mathrm{E}(X_y) - \mu$ fall in the d-dimensional subspace span(Γ), and thus Γ captures the dependency between X and Y. Once the response is observed, the term $\nu(y)$ which is unobserved can be approximated using a flexible set of basis functions as $\nu(y) \approx \beta \mathrm{f}(y)$. The subsequent model

$$X_y = \mu + \Gamma\beta\mathrm{f}(y) + \Delta^{1/2}\varepsilon \tag{1}$$

is called a PFC model where ε is assumed to be normally distributed with mean 0 and variance I_p. Under this model, Cook (2007) showed that $\Gamma^T\Delta^{-1}X$ is a sufficient reduction of X. The choice of the basis function allows to capture predictors that are linearly and nonlinearly related to the response. The maximum likelihood estimators of the parameters in the model have been obtained (Cook, 2007; Cook and Forzani, 2008).

In high dimensional settings, irrelevant predictors, which often abound, can hinder the accuracy of the estimated sufficient reduction. Our goal is to obtain a "pruned" estimator of the sufficient reduction, which not only helps achieve accuracy, but also allows the identification of the relevant variables. By "pruning", we mean removing inactive predictors that do not contain any regression information about the response. This is often called a sparse estimator.

An estimation of the sparse reduction kernel $\Delta^{-1}\Gamma$ has been proposed by Li (2007) who established a framework to obtain the sparse sufficient reduction using a regression-type formulation with the lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) penalties. Chen et al. (2010) proposed the coordinate independent sparse sufficient dimension reduction that shrinks row elements of $\Delta^{-1}\Gamma$ while preserving the orthogonality constraint of Γ. Both methodologies are apt when $n \gg p$. We herein construct a sequential likelihood ratio test that is reminiscent of the idea of testing predictor contribution in sufficient dimension reduction of Cook (2004). It helps obtain the sparse reduction under structures of Δ that allow $p > n$. We show the performance of the procedure through simulations.

2. A sequential test for sparse PFC

We assume that the p-vector predictor X can be partitioned as $(X_1^T, X_2^T)^T$, with $X_2 \in \mathbb{R}^{p_2}$, and let $(\Gamma_1^T, \Gamma_2^T)^T$, $\Delta = (\Delta_{ij})_{i,j=1,2}$ and $\Delta^{-1} = (\Delta^{ij})_{i,j=1,2}$ be the corresponding partitions of Γ, Δ and Δ^{-1} following the partition of X. Under