Journal of Multivariate Analysis 135 (2015) 25–42

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

MDR method for nonbinary response variable

Alexander Bulinski *, Alexander Rakitko

Moscow State University, Faculty of Mathematics and Mechanics, Moscow 119991, Russia

ARTICLE INFO

Article history:
Received 20 September 2013
Available online 11 December 2014

AMS 2000 subject classifications:
primary 62H12
secondary 62H25
60F05

Keywords:
Nonbinary response variable
Factors
Error function
Penalty function
i.i.d. observations
Prediction algorithm
K-fold cross-validation
Error estimation
Criterion of a.s. estimators convergence
Dimensionality reduction of factors
Central limit theorem

ABSTRACT

For nonbinary response variable depending on a finite collection of factors with values in a finite subset of $\mathbb{R}$ the problem of the optimal forecast is considered. The quality of prediction is described by the error function involving a penalty function. The criterion of almost sure convergence to unknown error function for proposed estimates constructed by means of a prediction algorithm and K-fold cross-validation procedure is established. It is demonstrated that imposed conditions admit the efficient verification. The developed approach permits to realize the dimensionality reduction of factors under consideration. One can see that the results obtained provide the base to identify the set of significant factors. Such problem arises, e.g., in medicine and biology. The central limit theorem for proposed statistics is proven as well. In this way one can indicate the approximate confidence intervals for employed error function.

1. Introduction

The high dimensional data are widely used in various stochastic models. Such data arise naturally when a response variable Y depends on a number of factors $X_1, \ldots, X_n$. For instance, in medical and biological studies Y can describe the state of the health of a patient, e.g., $Y = 1$ or $Y = -1$ mean that a person is sick or healthy, respectively. The challenging problem is to find in huge number of given factors the collection $X_{k_1}, \ldots, X_{k_r}$ of significant ones which are responsible for certain complex disease provoking (see, e.g., [14]). Note also that in pharmacological studies the values -1 or 1 of a response variable can describe efficient or nonefficient employment of some medicine (see, e.g., [19]). Thus solution of the problem to identify the set of significant factors has important applications even for binary response variable.

Now we assume that Y takes values in a finite subset of $\mathbb{R}$ (with more than two elements in general) and $X_1, \ldots, X_n$ take values in arbitrary finite set. This assumption is quite natural, e.g., for medical applications because we can consider the health state of a patient in more detail. In Section 2 we will describe our general model.

There are many complementary approaches concerning the prediction of response variable and the identification of the significant combinations of factors. Such analysis in medical and biological investigations is included in the special research domain called the *genome-wide association studies* (GWAS). The progress in this domain is discussed in the recent paper [27]. Among powerful statistical tools applied in GWAS one can indicate the principal component analysis [11], logistic and logic

* Corresponding author.
E-mail address: bulinski@yandex.ru (A. Bulinski).

http://dx.doi.org/10.1016/j.jmva.2014.11.008
0047-259X/

regression [20,21,23], LASSO [12,25] and various methods of statistical learning [10]. Mention in passing that there are new modifications of these methods. In the present paper we concentrate on the development of *multifactor dimensionality reduction* (MDR) method. This method was introduced in the paper by M. Ritchie et al. [18] for binary response variable. It goes back to the Michalski algorithm [13]. During the last decade more than 300 publications were devoted to this method. We are also interested in the dimensionality reduction. However instead of consideration of contingency tables (to specify zones of low and high risk) presented in [18] and many subsequent works we choose another way. Note that researchers use different terminology for specified approaches leading to dimensionality reduction of factors. For instance in [7,15,16,22] the authors propose the following methods: MDR-PDT (pedigree disequilibrium test), MDR-SP (structured populations), Gene-based MDR and MDR-FS (feature selection), respectively. We could call our method MDR-EFE (error function estimation). Contributions containing various improvements of the original MDR method are available also, e.g., in [6,9,17,19].

To predict Y we use some function f in factors $X_1, \dots, X_n$. The quality of such f is determined by means of error function $Err(f)$ involving a penalty function ψ. This penalty function allows us to take into account the importance of different values of Y. As the law of Y and $X = (X_1, \dots, X_n)$ is unknown we cannot find $Err(f)$. Thus statistical inference is based on the estimates of error function. Developing [2–4] we propose (in more general setting) statistics constructed by means of a prediction algorithm for response variable and K-fold cross-validation procedure. One of our main results gives the criterion of strong consistency of the mentioned error function when the number of observations tends to infinity. The strong consistency is essential because to identify the "significant collection" of factors we have to compare simultaneously a number of statistics. We demonstrate that this criterion admits the efficient employment even when instead of the penalty function one uses its strongly consistent estimates. In contrast to [2] the situation with the choice of the penalty function is more complicated.

We demonstrate the stability of proposed statistics. Namely, the central limit theorem (CLT) is proven for error function estimates in the framework of prediction algorithm, the penalty function and K-fold cross-validation for nonbinary response variable (for binary response variable such CLT was established in [3]). Also we pay attention to specification of the optimal forecast of Y and identification of the significant collection of factors.

The paper is organized as follows. Section 2 contains notation and auxiliary results. Here we discuss the problem of optimal (in a sense) prediction of nonbinary response variable with values in a finite set $\mathbb{Y} \subset \mathbb{R}$ by means of a collection of factors taking values in arbitrary finite set. For this purpose we define the prediction error involving a penalty function. In Section 3 we introduce prediction algorithm and for i.i.d. vectors of observations construct the estimator of unknown prediction error. The main result here (Theorem 1) provides the criterion of almost sure convergence of these estimators to prediction error. We also prove two corollaries containing conditions which are easy to handle. Section 4 is devoted to applications. Namely, we consider two important examples of prediction algorithm and verify conditions of the mentioned corollaries. Section 5 can be viewed as the foundation for dimensionality reduction of factors (see Theorem 2). In Section 6 we prove the central limit theorem (Theorem 3) for appropriately normalized and regularized estimators of error function. We complete the paper by multidimensional version of the CLT and some remarks permitting to find approximate confident intervals for unknown error function. The applications of developed method to simulated data are considered in [5].

2. Notation and auxiliary results

Let $X = (X_1, \dots, X_n)$ be a random vector with components $X_k : \Omega \to \{0, 1, \dots, s\}$ where $k = 1, \dots, n$ and $s, n \in \mathbb{N}$. All random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Set $\mathbb{X} = \{0, \dots, s\}^n$, $\mathbb{Y} = \{-m, \dots, 0, \dots, m\}$, here $m \in \mathbb{N}$. We assume that $Y : \Omega \to \mathbb{Y}, f : \mathbb{X} \to \mathbb{Y}$ and a penalty function $\psi : \mathbb{Y} \to \mathbb{R}_+$. The trivial case $\psi \equiv 0$ is excluded.

Remark 1. For instance in medicine one has the response variable which characterizes the state of the health of a patient by means of predetermined scale reflecting the progress of the disease. If such values constitute the set $\{0, 1, \dots, m\}$ then our model comprises this situation since Y can take the values $\{-m, \dots, -1\}$ with probability 0. Moreover, we can assume that Y takes arbitrary rational values $0 \le x_1 \le \dots \le x_m$ where $x_k = s_k/M$ ($s_k \in \mathbb{N}$, $M \in \mathbb{N}$, $k = 1, \dots, m$). Then we use the correspondence $x_k \mapsto s_k$, $k = 1, \dots, m$, and consider $\mathbb{Y} = \{-s_m, \dots, 0, \dots, s_m\}$. We employ the strongly consistent estimates of a penalty function (involving data) and if we know that $\mathsf{P}(Y = y) = 0$ for some $y \in \mathbb{Y}$ then we can take $\psi(y) = 0$ and $\psi_N(y) \equiv 0$ for such y ($N \in \mathbb{N}$) and our results will hold true as we will see further on. Note also that we can specify the importance of deviation of $f(X)$ from Y involving the penalty function ψ.

For $y \in \mathbb{Y}$, consider the set $A_y = \{x \in \mathbb{X} : f(x) = y\}$ and put $M = \{x \in \mathbb{X} : \mathsf{P}(X = x) > 0\}$. Introduce the *error function*

$$Err(f) := \mathsf{E}|Y - f(X)|\psi(Y).$$

It is easily seen that one can write $Err(f)$ as follows

$$Err(f) = \sum_{y,z \in \mathbb{Y}} |y - z|\psi(y)\mathsf{P}(Y = y, f(X) = z) = \sum_{z \in \mathbb{Y}} \sum_{x \in A_z} w^\top(x) q(z). \tag{1}$$

Here $q(z)$ is the zth column of $(2m + 1) \times (2m + 1)$ matrix Q with entries $q_{y,z} = |y - z|$, $y, z \in \mathbb{Y}$ (the entry $q_{-m,-m}$ is located at the left upper corner of Q),

$$w(x) = (\psi(-m)\mathsf{P}(Y = -m, X = x), \dots, \psi(m)\mathsf{P}(Y = m, X = x))^\top$$