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 $X_1, \dots, X_n$. 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 $X_{k_1}, \dots, X_{k_r}$ 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 $X_1, \dots, X_n$ 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 $\widehat{\psi}_N(y) \equiv 0$ 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),

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

and $\top$ stands for transposition. All vectors are considered as column-vectors. Further $\sharp A$ means the cardinality of a finite set A.

Let us describe $f : \mathbb{X} \to \mathbb{Y}$ which is a solution of the problem $Err(f) \to \inf$. In other words we search for a partition A_y, $y \in \mathbb{Y}$, of a set $\mathbb{X}$ such that a function

$$f = \sum_{y \in \mathbb{Y}} y\, \mathbb{I}\{A_y\} \tag{2}$$

has the minimal $Err(f)$. We call any solution f of this problem an *optimal function*.

For each nonempty set J such that $J \subset \mathbb{Y}$ ("$\subset$" is used as "$\subseteq$") put

$$B_J = \{x \in \mathbb{X} : w^\top(x)q(y) = w^\top(x)q(z), y, z \in J;\ w^\top(x)q(y) < w^\top(x)q(v), y \in J, v \in \mathbb{Y} \setminus J\}. \tag{3}$$

If $J = \mathbb{Y}$ then $B_{\mathbb{Y}} = \{x \in \mathbb{X} \colon w^\top(x)q(y) = w^\top(x)q(z),\ y, z \in \mathbb{Y}\}$. Note that $B_J \cap B_I = \varnothing$ if $J \neq I$ $(I, J \subset \mathbb{Y})$. Moreover,

$$\cup_{J \subset \mathbb{Y}, J \neq \varnothing} B_J = \mathbb{X}. \tag{4}$$

We write B_y for B_J when $J = \{y\}$, $y \in \mathbb{Y}$. Let $\mathbb{I}\{A\}$ be an indicator of a set A. As usual $\mathbb{I}\{\varnothing\} = 0$.

In view of (1) one can claim that $B_y \subset A_y$ for each $y \in \mathbb{Y}$. If $\cup_{y \in \mathbb{Y}} B_y \neq \mathbb{X}$ then, for any $x \in \mathbb{X} \setminus \cup_{y \in \mathbb{Y}} B_y$, there exists $J = J(x)$ such that $x \in B_J$ where $J \subset \mathbb{Y}$ and $\sharp J > 1$. Then one can include x in any A_y with y belonging to J (i.e. enlarge one of the sets B_y, $y \in J$, by means of element x). In such a way we obtain that $A_y = B_y \cup C_y$ where C_y, $y \in \mathbb{Y}$, form a partition of the set $\mathbb{X} \setminus \cup_{y \in \mathbb{Y}} B_y$. Obviously the proposed construction leads to f with minimal $Err(f)$. Other choice of A_y, $y \in \mathbb{Y}$, would lead to f with greater $Err(f)$ than one for a function proposed above. Thus we come to elementary

Lemma 1. *Any function $f : \mathbb{X} \to \mathbb{Y}$ providing the solution to the problem $Err(f) \to \inf$ has the form (2) with A_y, $y \in \mathbb{Y}$, specified above.*

Remark 2. Clearly, we can specify the unique way to construct the sets C_y, $y \in \mathbb{Y}$. For example, if $\mathbb{X} \setminus \cup_{y \in \mathbb{Y}} B_y \neq \varnothing$ then, for each $x \in \mathbb{X} \setminus \cup_{y \in \mathbb{Y}} B_y$, we find the unique B_J such that $x \in B_J$ $(J = J(x), J \subset \mathbb{Y}, \sharp J > 1)$. If $J = \{y_1, \ldots, y_r\}$ where $y_1 < \cdots < y_r$ we include x in C_{y_1}. Note also that we can consider the optimal f with $A^*_y = A_y \cap M$ for $y \in \mathbb{Y} \setminus \{-m\}$ and $A^*_{-m} = A_{-m} \cup (\mathbb{X} \setminus M)$.

Our next aim is to provide the convenient form for an optimal function f and rewrite $Err(f)$ in appropriate manner. For this purpose we represent B_y in the following way

$$x \in B_y \iff \begin{cases} w^\top(x)q(-m) < w^\top(x)q(-m+1), & y = -m, \\ w^\top(x)q(y) < w^\top(x)q(z), \quad z = y \pm 1, & y \neq \pm m, \\ w^\top(x)q(m-1) > w^\top(x)q(m), & y = m. \end{cases} \tag{5}$$

Note that B_y can be an empty set. To show that (5) is true we define, for $y \in \mathbb{Y}$, $y > -m$, the vector $\Delta(y) := q(y) - q(y-1)$. Clearly,

$$\Delta(y) = (\underbrace{1, \ldots, 1}_{m+y}, \underbrace{-1, \ldots, -1}_{m-y+1})^\top. \tag{6}$$

Inequality $w^\top(x)q(y) < w^\top(x)q(y+1)$ is equivalent to the following one $w^\top(x)\Delta(y+1) > 0$. For all $x \in \mathbb{X}$ the vector $w(x)$ has nonnegative components

$$w_y(x) := \psi(y)\mathsf{P}(Y = y, X = x), \quad y \in \mathbb{Y}. \tag{7}$$

Therefore inequality $w^\top(x)\Delta(y+1) > 0$ and (6) yields $w^\top(x)\Delta(z) > 0$ if $z \geq y+1$ $(z \in \mathbb{Y})$. For $z \geq y+1$ $(z \in \mathbb{Y})$, one has

$$w^\top(x)(q(z) - q(y)) = \sum_{k=y+1}^{z} w^\top(x)\Delta(k). \tag{8}$$

Consequently, $w^\top(x)(q(z) - q(y)) > 0$. In a similar way one can see that inequality $w^\top(x)q(y) < w^\top(x)q(y-1)$ implies, for $t < y$, $t \in \mathbb{Y}$, relation $w^\top(x)q(y) < w^\top(x)q(t)$. Thus (5) is established. Employing (8) we observe that the set $J = \{y_1, \ldots, y_r\}$ appearing in Remark 2 has the form $\{y_1, y_1+1, \ldots, y_1+r-1\}$.

For $x \in \mathbb{X}$, consider the vector $L(x)$ having the following $2m$ components

$$L_y(x) := w^\top(x)\Delta(y) = w_{-m}(x) + \cdots + w_{y-1}(x) - w_y(x) - \cdots - w_m(x), \tag{9}$$

here $y \in \mathbb{Y}$, $y > -m$. Then, due to (5) one has, for each $y \in \mathbb{Y}$,

$$x \in B_y \iff \begin{cases} L_{-m+1}(x) > 0, & & y = -m, \\ L_{y+1}(x) > 0, & L_y(x) < 0, & y \neq \pm m, \\ L_m(x) < 0, & & y = m. \end{cases} \tag{10}$$

Further we will use the property of the vector-function $L(x)$, $x \in \mathbb{X}$, containing in the following statement.

Lemma 2. *Let $L_t(x) = 0$ and $L_z(x) = 0$ for some $x \in \mathbb{X}$, $t, z \in \mathbb{Y}$, $-m < t < z$. Then $L_y(x) = 0$ for any $y \in \mathbb{Y}$ such that $t \le y \le z$.*

Proof. For each $x \in \mathbb{X}$ the vector $w(x)$ has nonnegative components. Formula (9) shows that for any $x \in \mathbb{X}$ the function $L_y(x)$ is nondecreasing function in y ($y \in \mathbb{Y}, y > -m$). This observation leads to the desired statement. □

Using Remark 2 it is convenient to make the following choice of the *optimal function* f_{opt}. Namely, according to (10) we can write

$$f_{opt}(x) = y \iff \begin{cases} L_{-m+1}(x) \ge 0, & & y = -m, \\ L_{y+1}(x) \ge 0, & L_y(x) < 0, & y \ne \pm m, \\ L_m(x) < 0, & & y = m. \end{cases} \tag{11}$$

In fact, according to Remark 1 we have $A_m = B_m$ and therefore we write in (11) the strict inequality $L_m(x) < 0$ when $y = m$.

Now consider random vectors φ and χ with the respective components

$$\varphi_y = \psi(y)\mathbb{I}\{Y = y\}, \qquad \chi_y = \mathbb{I}\{X \in A_y\}, \quad y \in \mathbb{Y}.$$

Then we can rewrite (1) as

$$Err(f) = \mathsf{E}\varphi^\top Q\chi.$$

Note that Q can be represented as the sum of $2m$ symmetric matrices with 0 and 1 entries.

$$Q = \begin{pmatrix} 0 & 1 & 1 & \dots & 1 & 1 & 1 \\ 1 & 0 & 1 & \dots & 1 & 1 & 1 \\ 1 & 1 & 0 & \dots & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 0 & 1 & 1 \\ 1 & 1 & 1 & \dots & 1 & 0 & 1 \\ 1 & 1 & 1 & \dots & 1 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 & \dots & 1 & 1 & 1 \\ 0 & 0 & 0 & \dots & 1 & 1 & 1 \\ 1 & 0 & 0 & \dots & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 0 & 0 & 1 \\ 1 & 1 & 1 & \dots & 0 & 0 & 0 \\ 1 & 1 & 1 & \dots & 1 & 0 & 0 \end{pmatrix} + \dots + \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 & 0 \end{pmatrix}.$$

In other words,

$$Q = \sum_{i=0}^{2m-1} Q^{(i)} \tag{12}$$

where the matrix $Q^{(i)} = (q^{(i)}_{y,z})_{y,z\in\mathbb{Y}}$ has entries $q^{(i)}_{y,z} = 0$ if $|y-z| \le i$ and $q^{(i)}_{y,z} = 1$ otherwise. Formula (12) permits to rewrite $Err(f)$ as follows

$$Err(f) = \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \psi(y)\mathsf{P}(Y = y, |f(X) - y| > i). \tag{13}$$

Here we take into account that in representation of Q as the sum of matrices some of these matrices have rows containing only zero entries. Thus we obtain formula (13) which is the key formula for $Err(f)$ further analysis.

3. Criterion of prediction error estimates strong consistency

The law of (X, Y) is unknown, therefore, for a given $f : \mathbb{X} \to \mathbb{Y}$, we cannot calculate $Err(f)$. Thus it is natural that statistical inference concerning the quality of prediction of the response variable Y by means of $f(X)$ is based on the estimates of $Err(f)$.

Let $\xi^1, \xi^2, \dots$ be a sequence of independent identically distributed (i.i.d.) random vectors $(X^1, Y^1), (X^2, Y^2), \dots$ having the same law as (X, Y). For $N \in \mathbb{N}$, set $\xi_N = (\xi^1, \dots, \xi^N)$. We will use approximation of $Err(f)$ by means of ξ_N (as $N \to \infty$) and a *prediction algorithm* (PA). This PA employs a function $f_{PA} = f_{PA}(x, \xi_N)$ defined for $x \in \mathbb{X}$ and ξ_N and taking values in $\mathbb{Y}$. More exactly, we operate with a *family of functions* $f_{PA}(x, v_p)$ (with values in $\mathbb{Y}$) defined for $x \in \mathbb{X}$ and $v_p \in (\mathbb{X} \times \mathbb{Y})^p$ where $p \in \mathbb{N}, p \le N$. To simplify the notation we write $f_{PA}(x, v_p)$ instead of $f^p_{PA}(x, v_p)$. For $S \subset \{1, \dots, N\}$ we set $\xi_N(S) = \{\xi^j, j \in S\}$ and $\overline{S} := \{1, \dots, N\} \setminus S$. For $K \in \mathbb{N}$ ($K > 1$), introduce a partition of a set $\{1, \dots, N\}$ into the subsets

$$S_k(N) = \{(k-1)[N/K] + 1, \dots, k[N/K]\mathbb{I}\{k < K\} + N\mathbb{I}\{k = K\}\}, \quad k = 1, \dots, K,$$

here $[a]$ is the integer part of a number $a \in \mathbb{R}$. Following [2] we can construct an estimate of $Err(f)$ involving ξ_N as well as prediction algorithm defined by f_{PA} and K-cross-validation (on cross-validation we refer, e.g., to [1]). Namely, set

$$\widehat{Err}_K(f_{PA}, \xi_N) := \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \frac{1}{K} \sum_{k=1}^{K} \sum_{j\in S_k(N)} \frac{\widehat{\psi}(y, \xi_N(S_k(N)))\mathbb{I}\{A_N(y, i, k, j)\}}{\sharp S_k(N)}. \tag{14}$$

Here $A_N(y,i,k,j) = \{Y^j = y, |f_{PA}(X^j, \xi_N(\overline{S_k(N)})) - y| > i\}$ and, for each $k \in \{1,\dots,K\}$, let $\widehat{\psi}(y, \xi_N(S_k(N)))$ be strongly consistent estimates of $\psi(y)$ (as $N \to \infty$) for all $y \in \mathbb{Y}$, i.e.

$$\widehat{\psi}(y, \xi_N(S_k(N))) \to \psi(y) \quad \text{a.s.}, \ y \in \mathbb{Y}, \ N \to \infty. \tag{15}$$

We want to guarantee that convergence (in a certain sense) of $f_{PA}(\cdot, \xi_N)$ to $f(\cdot)$ as $N \to \infty$ implies the relation

$$\widehat{Err}_K(f_{PA}, \xi_N) \to Err(f) \quad \text{a.s.}, \ N \to \infty. \tag{16}$$

In what follows the sum over empty set is equal to zero as usual.

Theorem 1. *Let $\xi^1, \xi^2, \dots$ be a sequence of i.i.d. random vectors with the same law as (X, Y), ψ be a penalty function, $f : \mathbb{X} \to \mathbb{Y}$ and f_{PA} define the prediction algorithm. Assume that there exists nonempty set $U \subset \mathbb{X}$ such that for each $x \in U$ and every $k = 1, \dots, K$ one has*

$$f_{PA}(x, \xi_N(\overline{S_k(N)})) \to f(x) \quad \text{a.s.}, \ N \to \infty. \tag{17}$$

Then (16) holds if and only if

$$\sum_{k=1}^{K} \sum_{x \in \mathbb{X}\setminus U} w^{\top}(x)\, Q\, \delta(N, x, k) \to 0 \quad \text{a.s.}, \ N \to \infty, \tag{18}$$

where, for $x \in \mathbb{X}$, $N \in \mathbb{N}$ and $k = 1, \dots, K$, the vector $\delta(N, x, k)$ has components

$$\delta_y(N, x, k) = \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) = y\} - \mathbb{I}\{f(x) = y\}, \quad y \in \mathbb{Y}.$$

Proof. Let us show that asymptotic behavior of $\widehat{Err}_K(f_{PA}, \xi_N)$ as $N \to \infty$ will be the same if one replaces $\widehat{\psi}(y, \xi_N(S_k(N)))$ by $\psi(y)$ in (14). In other words relation (16) is equivalent to the following one

$$\sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \frac{1}{K} \sum_{k=1}^{K} \sum_{j \in S_k(N)} \frac{\psi(y)\mathbb{I}\{Y^j = y, |f_{PA}(X^j, \xi_N(\overline{S_k(N)})) - y| > i\}}{\sharp S_k(N)} \to Err(f) \quad \text{a.s.} \tag{19}$$

as $N \to \infty$. Indeed, (15) holds and, for any $\omega \in \Omega$, $i = 0, \dots, 2m-1$ and $k = 1, \dots, K$, one has

$$\frac{1}{\sharp S_k(N)} \sum_{j \in S_k(N)} \mathbb{I}\{Y^j = y, |f_{PA}(X^j, \xi_N(\overline{S_k(N)})) - y| > i\} \le 1.$$

For each y and i, due to the strong law of large numbers for arrays (SLLNA) (see, e.g., [24])

$$\frac{1}{\sharp S_k(N)} \sum_{j \in S_k(N)} \mathbb{I}\{Y^j = y, |f(X^j) - y| > i\} \to \mathsf{P}(Y = y, |f(X) - y| > i) \quad \text{a.s.}, \ N \to \infty.$$

Consequently, for any $k = 1, \dots, K$ we get

$$\sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \sum_{j \in S_k(N)} \frac{\psi(y)\mathbb{I}\{Y^j = y, |f(X^j) - y| > i\}}{\sharp S_k(N)} \to Err(f) \quad \text{a.s.}, \ N \to \infty. \tag{20}$$

For $y \in \mathbb{Y}$, $N \in \mathbb{N}$, $k = 1, \dots, K$ and $i = 0, \dots, 2m-1$, introduce the random variables

$$Q_{N,k}^{(i)}(y) = \frac{1}{\sharp S_k(N)} \sum_{j \in S_k(N)} \mathbb{I}\{Y^j = y\} F_{N,k}^{(i)}(X^j, y)$$

where

$$F_{N,k}^{(i)}(x, y) := \mathbb{I}\{|f_{PA}(x, \xi_N(\overline{S_k(N)})) - y| > i\} - \mathbb{I}\{|f(x) - y| > i\}. \tag{21}$$

In view of (20) relation (19) is equivalent to the following one

$$\sum_{k=1}^{K} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \psi(y) Q_{N,k}^{(i)}(y) \to 0 \quad \text{a.s.}, \ N \to \infty. \tag{22}$$

We can write $Q_{N,k}^{(i)}(y) = Q_{N,k}^{(i),U}(y) + Q_{N,k}^{(i),\mathbb{X}\setminus U}(y)$, $i = 0, \dots, 2m-1$, where

$$Q_{N,k}^{(i),V}(y) = \frac{1}{\sharp S_k(N)} \sum_{j \in S_k(N)} \mathbb{I}\{X^j \in V\}\mathbb{I}\{Y^j = y\} F_{N,k}^{(i)}(X^j, y), \quad V \subset \mathbb{X}.$$

In view of (17) we can examine $Q^{(i),U}_{N,k}(y)$. For $y \in \mathbb{Y}, N \in \mathbb{N}, k = 1, \dots, K$ and $i = 0, \dots, 2m-1$, we come to the inequalities

$$|Q^{(i),U}_{N,k}(y)| \le \sum_{x\in U} \left| \mathbb{I}\{|f_{PA}(x, \xi_N(\overline{S_k(N)})) - y| > i\} - \mathbb{I}\{|f(x) - y| > i\} \right|.$$

Functions f and f_{PA} take values in $\mathbb{Y}$. Thus (17) yields that for each $x \in U$, $k = 1, \dots, K$ and almost every $\omega \in \Omega$ one can find an integer $N_1(x, k, \omega)$ such that $f_{PA}(x, \xi_N(\overline{S_k(N)})) = f(x)$ if $N \ge N_1(x, k, \omega)$. Therefore, $Q^{(i),U}_{N,k}(y) = 0$ for any $i = 0, \dots, 2m-1, y \in \mathbb{Y}, k = 1, \dots, K$ and almost every $\omega \in \Omega$ when $N \ge N_1(\omega) = \max_{x\in U, k=1,\dots,K} N_1(x, k, \omega)$. Obviously, $N_1 < \infty$ a.s. because $\sharp U < \infty$. Thus we have shown that

$$\sum_{k=1}^{K} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \psi(y) Q^{(i),U}_{N,k}(y) \to 0 \quad \text{a.s., } N \to \infty. \tag{23}$$

Now we turn to analysis of $Q^{(i),\mathbb{X}\setminus U}_{N,k}(y)$. If $U = \mathbb{X}$ then $Q^{(i),\mathbb{X}\setminus U}_{N,k}(y) = 0$ for all i, N, k and y under consideration. In this case (23) is equivalent to (22). Thus for $U = \mathbb{X}$ the claim of theorem is verified. Let now $U \ne \mathbb{X}$. Set

$$\tau_N(\mathbb{X} \setminus U) = \sum_{k=1}^{K} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \psi(y) Q^{(i),\mathbb{X}\setminus U}_{N,k}(y).$$

Obviously,

$$\tau_N(\mathbb{X} \setminus U) = \sum_{k=1}^{K} \sum_{x\in\mathbb{X}\setminus U} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \psi(y) \frac{1}{\sharp S_k(N)} \sum_{j\in S_k(N)} \mathbb{I}\{X^j = x\}\mathbb{I}\{Y^j = y\} F^{(i)}_{N,k}(x, y). \tag{24}$$

Due to SLLNA, for each $x \in \mathbb{X}, y \in \mathbb{Y}$ and $k = 1, \dots, K$,

$$\frac{1}{\sharp S_k(N)} \sum_{j\in S_k(N)} \mathbb{I}\{X^j = x\}\mathbb{I}\{Y^j = y\} \to \mathsf{P}(X = x, Y = y) \quad \text{a.s., } N \to \infty. \tag{25}$$

Thus (24) and (25) demonstrate that $\lim_{N\to\infty} \tau_N(\mathbb{X} \setminus U) = 0$ a.s. if and only if

$$\nu_N(\mathbb{X} \setminus U) := \sum_{k=1}^{K} \sum_{x\in\mathbb{X}\setminus U} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} w_y(x) F^{(i)}_{N,k}(x, y) \to 0 \quad \text{a.s., } N \to \infty. \tag{26}$$

Taking into account that $\mathbb{I}\{\cup_{j\in J} D_j\} = \sum_{j=1}^{J} \mathbb{I}\{D_j\}$ for pairwise disjoint sets $D_1, \dots, D_J$, we can write

$$\begin{aligned} \nu_N(\mathbb{X} \setminus U) &= \sum_{k=1}^{K} \sum_{x\in\mathbb{X}\setminus U} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \sum_{|r-y|>i} w_y(x) \delta_r(N, x, k) \\ &= \sum_{k=1}^{K} \sum_{x\in\mathbb{X}\setminus U} \sum_{y\in\mathbb{Y}} w_y(x) \sum_{i=0}^{I(y)} \sum_{|r-y|>i} \delta_r(N, x, k) \end{aligned} \tag{27}$$

where $I(y) = m - 1 + |y|$ for $y \in \mathbb{Y}$. Note that

$$\sum_{i=0}^{I(y)} \sum_{|r-y|>i} \delta_r(N, x, k) = \sum_{r=-m}^{m} \sum_{i<|r-y|} \delta_r(N, x, k) = \sum_{r=-m}^{m} |y - r| \delta_r(N, x, k) = (Q\delta(N, x, k))_y \tag{28}$$

as $|r - y| - 1 \le I(y)$ for all $r, y \in \mathbb{Y}$. Here $(Q\delta(N, x, k))_y, y \in \mathbb{Y}$, are coordinates of the vector $Q\delta(N, x, k)$. Therefore (27) and (28) yield that condition (26) is equivalent to (18). The proof is complete. □

Remark 3. Theorem 1 provides the criterion what one has to assume outside the "good set" U where (17) holds to guarantee the desired relation (16). Further we will see that it is possible to verify conditions (17) and (18) efficiently. Note also that the statement of Theorem 1 will be true if instead of $\xi^1, \dots, \xi^N$ we consider independent random vectors $\xi^{N,1}, \dots, \xi^{N,N}$ such that $\xi^{N,j} := (X^{(N,j)}, Y^{(N,j)})$ has the same law as $(X, Y), j = 1, \dots, N, N \in \mathbb{N}$.

Remark 4. In Theorem 1 we did not suppose that nonempty set U consists of all $x \in \mathbb{X}$ satisfying (17). However, if relation (17) holds for some $u \in \mathbb{X} \setminus U$ then $f_{PA}(u, \xi_N(\overline{S_k(N)})) = f(u)$ a.s. for $k = 1, \dots, K$ when N is large enough (i.e. $N > N_1(\omega)$). Therefore relation (26) is equivalent to the analogous one where summation over $x \in \mathbb{X} \setminus U$ is replaced by summation over $x \in \mathbb{X} \setminus (U \cup \{u\})$. Therefore we obtain an equivalent formulation of Theorem 1 if U consists of all $x \in \mathbb{X}$ satisfying (17). Moreover, if there is no nonempty $U \subset \mathbb{X}$ such that (17) holds then relation (16) is equivalent to (18) with $U = \varnothing$. Consequently, in Theorem 1 we need not assume that the set U is nonempty.

Remark 5. Let (16) be satisfied and assume that, for some constant C_0,

$$\widehat{\psi}(y, \xi_N(S_k(N))) \le C_0 \quad \text{a.s. for } N \in \mathbb{N},\ k = 1, \dots, K,\ y \in \mathbb{Y}. \tag{29}$$

Then $\widehat{Err}_K(f_{PA}, \xi_N) \le 2m(2m+1)C_0$, $N \in \mathbb{N}$. Thus the Lebesgue theorem on dominated convergence implies that $\widehat{Err}_K(f_{PA}, \xi_N)$ is asymptotically unbiased estimate of $Err(f)$.

For $N \in \mathbb{N}, x \in \mathbb{X}, k \in \{1, \dots, K\}$ and $t \in \mathbb{Y}$, introduce the random vector $I(N, x, k, t)$ with components

$$I_y(N, x, k, t) = \begin{cases} -\mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) < y\}, & -m < y \le t, \\ \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \ge y\}, & t < y \le m. \end{cases} \tag{30}$$

If $t = -m$ then $\{-m < y \le t\} = \varnothing$ and $I_y(N, x, k, -m) = \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \ge y\}$, if $t = m$ then $\{t < y \le m\} = \varnothing$ and $I_y(N, x, k, m) = -\mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \le y - 1\}$, here $y > -m, y \in \mathbb{Y}$.

Corollary 1. *Condition (18) of Theorem 1 is equivalent to the requirement*

$$\sum_{k=1}^{K} \sum_{t \in \mathbb{Y}} \sum_{x \in \mathbb{X}(t,U)} L^{\top}(x) I(N, x, k, t) \to 0 \quad a.s.,\ N \to \infty, \tag{31}$$

where $L^{\top}(x) := (L_{-m+1}(x), \dots, L_m(x))$, $L_y(x),\ y = -m+1, \dots, m$, are defined in (9) and $\mathbb{X}(t, U) := (\mathbb{X} \setminus U) \cap \{x \in M : f(x) = t\}$.

Proof. It is easily seen that condition (18) can be written in the following manner

$$\sum_{k=1}^{K} \sum_{t \in \mathbb{Y}} \sum_{x \in \mathbb{X}(t,U)} w^{\top}(x)\, Q\, \delta(N, x, k) \to 0 \quad \text{a.s.},\ N \to \infty. \tag{32}$$

Note that, for $x \in \mathbb{X}(t, U)$,

$$w^{\top}(x)\, Q\, \delta(N, x, k) = w^{\top}(x) \sum_{y \in \mathbb{Y}} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) = y\}(q(y) - q(t))$$

because $\sum_{y \in \mathbb{Y}} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) = y\} = 1$ and Q is symmetric matrix. For $y, t \in \mathbb{Y}$, according to (8) and (9) we get

$$w^{\top}(x)(q(y) - q(t)) = \begin{cases} -L_{y+1}(x) - \dots - L_t(x), & y < t, \\ 0, & y = t, \\ L_{t+1}(x) + \dots + L_y(x), & y > t. \end{cases}$$

If $t = m$ then $\sum_{t+1 \le r \le y} L_r(x) = 0$ and if $t = -m$ then $\sum_{y+1 \le r \le t} L_r(x) = 0$ as the sums over empty set. Changing the order of summation we obtain

$$\begin{aligned} \sum_{y<t} \sum_{r=y+1}^{t} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) = y\} L_r(x) &= \sum_{r=-m+1}^{t} \sum_{y=-m}^{r-1} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) = y\} L_r(x) \\ &= \sum_{r=-m+1}^{t} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \le r - 1\} L_r(x). \end{aligned} \tag{33}$$

In a similar way one has

$$\sum_{y>t} \sum_{r=t+1}^{y} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) = y\} L_r(x) = \sum_{r=t+1}^{m} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \ge r\} L_r(x). \tag{34}$$

Thus (33) and (34) entail

$$w^{\top}(x)\, Q\, \delta(N, x, k) = L^{\top}(x) I(N, x, k, t). \tag{35}$$

Combining relations (32) and (35) we come to (31). The proof is complete. □

Corollary 2. *Let ψ be a penalty function, $f : \mathbb{X} \to \mathbb{Y}$ and the prediction algorithm be defined by a function (family) f_{PA}. Suppose that for some set $U \subset \mathbb{X}$ condition (17) is satisfied. Assume that for each $t \in \mathbb{Y}$ and any $x \in \mathbb{X}(t, U)$ there exist $i = i(x), j = j(x)$ belonging to $\mathbb{Y}$, $i < j$, such that*

$$i \le f_{PA}(x, \xi_N(\overline{S_k(N)})) \le j \quad a.s. \text{ for } k = 1, \dots, K \tag{36}$$

when N is large enough. Then the condition

$$L_{\min\{t,i\}+1}(x) = \cdots = L_{\max\{t,j\}}(x) = 0 \tag{37}$$

implies that (31) *holds.*

Proof. Obviously, for each $x \in \mathbb{X}(t, U)$, we have

$$L^{\top}(x)I(N, x, k, t) = \sum_{y=-m+1}^{\min\{t,i\}} L_y^{\top}(x)I_y(N, x, k, t) + \sum_{y=\max\{t,j\}+1}^{m} L_y^{\top}(x)I_y(N, x, k, t). \tag{38}$$

In view of (30) every summand in the right-hand side of (38) will vanish a.s. for all N large enough by virtue of (36). Taking into account that $\sharp X < \infty$ we obtain the desired statement. □

4. Applications

Example 1. Let ψ be a penalty function and $f = f_{opt}$ where f_{opt} is defined in (11). For $x \in \mathbb{X}$ and a set $W_N \subset \{1, \dots, N\}$ introduce the random vector $\widetilde{w}^{W_N}(x, \omega)$ with components

$$\widetilde{w}_y^{W_N}(x, \omega) = \frac{\psi(y)}{\sharp W_N} \sum_{j \in W_N} I\{Y^j = y, X^j = x\}, \quad y \in \mathbb{Y}, \tag{39}$$

here $0/0 := 0$ (if W_N is empty). Put

$$\widetilde{f}_{PA}(x, \xi_N(W_N, \omega)) := \sum_{y \in \mathbb{Y}} y\, \mathbb{I}\{x \in \widetilde{A}_y^{W_N}(\omega)\} \tag{40}$$

where $\xi_N(W_N, \omega) = \{\xi_i(\omega), \omega \in \Omega, i \in W_N\}$,

$$x \in \widetilde{A}_y^{W_N}(\omega) \iff \begin{cases} \widetilde{L}_{-m+1}^{W_N}(x, \omega) \geq 0, & y = -m, \\ \widetilde{L}_{y+1}^{W_N}(x, \omega) \geq 0, \quad \widetilde{L}_y^{W_N}(x, \omega) < 0, & y \neq \pm m, \\ \widetilde{L}_m^{W_N}(x, \omega) < 0, & y = m, \end{cases} \tag{41}$$

and

$$\widetilde{L}_y^{W_N}(x, \omega) := (\widetilde{w}^{W_N}(x, \omega))^{\top} \Delta(y). \tag{42}$$

We write ω in (39)–(42) to emphasize the randomness of variables under consideration. Clearly we can define $\widetilde{f}_{PA}(x, v_p)$ for $x \in \mathbb{X}$, $v_p \in (\mathbb{X} \times \mathbb{Y})^p$ and then obtain $\widetilde{f}_{PA}(x, \xi_N(W_N))$ appearing in (40) by setting $v_{\sharp W_N} = \xi_N(W_N)$.

Let us show that f and f_{PA} satisfy conditions of Corollary 2 if we take

$$U = \{x \in \mathbb{X} : L_y(x) \neq 0 \text{ for all } y = -m+1, \dots, m\}. \tag{43}$$

One can claim that not only (17) is true but

$$\widetilde{f}_{PA}(x, \xi_N(W_N)) \to f(x) \quad \text{a.s., } N \to \infty,$$

for any $W_N \subset \{1, \dots, N\}$ such that $\sharp W_N \to \infty$ $(N \to \infty)$. Indeed, according to SLLNA, for any $x \in \mathbb{X}, y \in \mathbb{Y}, y > -m$, and such sets W_N, one has

$$\widetilde{L}_y^{W_N}(x, \omega) \to L_y(x) \quad \text{a.s., } N \to \infty. \tag{44}$$

Let $x \in U$ then, for each $y > -m, y \in \mathbb{Y}$, we can claim that $L_y(x) < 0$ or $L_y(x) > 0$. Hence for almost every $\omega \in \Omega$ there exists $N_2 = N_2(\omega)$ such that either $\widetilde{L}_y^{W_N}(x, \omega) < 0$ or $\widetilde{L}_y^{W_N}(x, \omega) > 0$ when $N > N_2(\omega)$. Thus (41) yields that condition (17) is true for U defined by (43). Take now $x \in \mathbb{X} \setminus U$. Then $L_v(x) = 0$ for some $v \in \mathbb{Y}$, $v > -m$. In this case according to Remark 2 we find in $\mathbb{Y}$ the subset $J = J(x)$ with $\sharp J(x) > 1$ such that $x \in B_J$ (see (3)). Then $v \in J(x)$. In view of (11) we see that $x \in \mathbb{X}(t, U)$ where $t = \min\{y : y \in J(x)\}$. For any $k \in \{1, \dots, K\}$ and all N large enough one has

$$\widetilde{f}_{PA}(x, \xi_N(\overline{S_k(N)})) \in J(x) \quad \text{a.s.}$$

Indeed, according to Lemma 2 we can state that $L_y(x) \neq 0$ for $y \in (\mathbb{Y} \setminus J(x)) \cup \{t\}$, $y > -m$, and thus the relation $f_{PA}(x, \xi_N(\overline{S_k(N)})) \in \mathbb{Y} \setminus J(x)$, for any $k \in \{1, \dots, K\}$ and all N large enough, is impossible due to (41) and (44). Here we bear on the observation that $L_z(x) < 0$ for $z \leq t$ and $L_z(x) > 0$ for $z > \max\{y : y \in J(x)\}$. Consequently we can apply Corollary 2 with $i = t$ and $j = \max\{y : y \in J(x)\}$ because by Lemma 2 this choice guarantees the validity of (37). □

Example 2. Now we will stipulate that the penalty function ψ is unknown. Assume that, for a sequence of sets $(W_N)_{N\in\mathbb{N}}$ such that $W_N \subset \{1,\dots,N\}$ and $\sharp W_N \to \infty$ as $N \to \infty$, there exists a sequence of random variables $(\widehat{\psi}(y, \xi_N(W_N)))_{N\in\mathbb{N}}$ satisfying for every $y \in \mathbb{Y}$ the relation

$$\widehat{\psi}(y, \xi_N(W_N)) \to \psi(y) \quad \text{a.s.}, \; N \to \infty. \tag{45}$$

For $x \in \mathbb{X}$ and $N \in \mathbb{N}$, introduce the random vectors $\widehat{w}^{W_N}(x,\omega)$ and $\widehat{L}^{W_N}(x,\omega)$ with the components

$$\widehat{w}_y^{W_N}(x,\omega) := \frac{\widehat{\psi}(y, \xi_N(W_N))}{\sharp W_N} \sum_{j\in W_N} \mathbb{I}\{Y^j = y, X^j = x\}, \tag{46}$$

$$\widehat{L}_y^{W_N}(x,\omega) := (\widehat{w}^{W_N}(x,\omega))^\top \Delta(y), \quad y \in \mathbb{Y}, \tag{47}$$

respectively. Now define $\widehat{A}_y^{W_N}(\omega)$ by way of (41) where instead of $\widetilde{L}_y^{W_N}(x,\omega)$ one uses $\widehat{L}_y^{W_N}(x,\omega)$, $y \in \mathbb{Y}$, $x \in \mathbb{X}$, $\omega \in \Omega$ and $N \in \mathbb{N}$. Set

$$\widehat{f}_{PA}(x, \xi_N(W_N)) = \sum_{y\in\mathbb{Y}} y\, \mathbb{I}\{x \in \widehat{A}_y^{W_N}(\omega)\}. \tag{48}$$

Similarly to Example 1 we can show that $f = f_{opt}$ where f_{opt} is defined in (11) and $\widehat{f}_{PA}$ given by (48) satisfies conditions of Corollary 2 with U introduced in (43).

In [26] the following choice of a penalty function ψ was proposed when a binary response variable Y takes values -1 and 1

$$\psi(y) = c\,(\mathsf{P}(Y = y))^{-1}, \quad y \in \mathbb{Y}, \; c = const > 0. \tag{49}$$

Assuming here that $\mathsf{P}(Y = y) > 0$ for $y \in \mathbb{Y}$ one can take $c = 1$ in (49) without loss of generality. In [2] it was explained that this choice is natural. For general case $\mathbb{Y} = \{-m,\dots,m\}$ ($m \in \mathbb{N}$) we also consider the penalty function given by (49) (with $c = 1$ and $\mathsf{P}(Y = y) > 0$ for any $y \in \mathbb{Y}$). For $y \in \mathbb{Y}$ and $N \in \mathbb{N}$, set $A_{W_N}(y) = \{Y^j \neq y \text{ for all } j \in W_N\}$,

$$\widehat{\mathsf{P}}_{W_N}(Y = y) = \frac{1}{\sharp W_N} \sum_{j\in W_N} \mathbb{I}\{Y^j = y\},$$

$$\widehat{\psi}(y, \xi_N(W_N)) = \frac{1 - \mathbb{I}\{A_{W_N}(y)\}}{\widehat{\mathsf{P}}_{W_N}(Y = y)} \tag{50}$$

where $0/0 := 0$, as usual. Then (45) holds since $\mathbb{I}\{A_{W_N}(y)\} \to 0$ a.s., for each $y \in \mathbb{Y}$, when $N \to \infty$. Therefore, by virtue of (45) we see that, for $k = 1,\dots,K$, relation (15) holds. □

Remark 6. Recall that in medical applications the response function Y often describes the health state of a patient. Namely, for binary variable the values 1 and -1 mean “sick” and “healthy” (“control”), respectively. If Y takes values in the set $\{-1, 0, 1\}$ then values 1 and -1 have the same meaning and the value 0 describes the “intermediate state”, that is one cannot decide whether disease (will) appears or not. Thus for this important case of ternary response variable Corollary 1 provides the criterion of (18) validity involving asymptotic behavior of f_{PA} and properties of functions $L_0(x)$, $L_1(x)$ for $x \in \mathbb{X} \setminus U$.

Now we discuss the problem of a penalty function ψ choice. It was mentioned above that (49) is appropriate for binary response variable Y with values -1 and 1. Namely, in [2] it was shown that if we assume that f_{PA} does not capture the dependence of Y on X outside the “good” set U (i.e. such set that (17) holds) then the independence of events $\{Y = 1\}$ and $\{X = x\}$ for $x \in \mathbb{X} \setminus U$ naturally leads (see Corollary 1 in [2]) to formula (49). However for Y taking values in the set $\mathbb{Y} = \{-m,\dots,m\}$ with $m \in \mathbb{N}$ the situation is more complicated. If we want (in a similar way to the case of binary Y) to have $L_y(x) = 0$ for any $y \in \mathbb{Y}$, $y > -m$, and $x \in \mathbb{X} \setminus U$ then we see that it is equivalent to relations

$$\psi(y)\mathsf{P}(Y = y, X = x) = 0, \quad -m < y < m, \tag{51}$$

$$\psi(-m)\mathsf{P}(Y = -m, X = x) = \psi(m)\mathsf{P}(Y = m, X = x). \tag{52}$$

Thus if we assume that events $\{Y = y\}$ and $\{X = x\}$ are independent for $y \in \mathbb{Y}$ and $x \in \mathbb{X} \setminus U$ then (52) is satisfied when (49) holds for $y \in \{-m, m\}$ whereas (51) means $\psi(y)\mathsf{P}(Y = y) = 0$ if $\mathsf{P}(X \in \mathbb{X} \setminus U) > 0$. Therefore (51) is valid if $\psi(y) = 0$ when $\mathsf{P}(Y = y) \neq 0$. Thus in this way for general $\mathbb{Y}$ one cannot justify the choice of $\psi(y)$ provided by (49). If $\mathbb{Y} = \{-1, 0, 1\}$ then the choice of $\psi(y)$ according to (49) for $y \in \{-1, 1\}$ and $\psi(0) = 0$ can be viewed as possible when one can say that we lose nothing in the “intermediate” case corresponding to $Y = 0$. Note that the choice of ψ by (49) is attractive if we want to take into account the rare values of Y. We also emphasize that in Corollary 2 we did not suppose that $L_y(x) = 0$ for any $y \in \mathbb{Y}$, $y > -m$, and $x \in \mathbb{X} \setminus U$.

5. Dimensionality reduction

For many models it is natural to assume that response variable Y depends only on some factors $X_{k_1},\dots,X_{k_r}$ where $1\le k_1<\cdots<k_r\le n$. In other words, for any $x=(x_1,\dots,x_n)\in M$ and $y\in\mathbb{Y}$,

$$\mathsf{P}(Y=y|X_1=x_1,\dots,X_n=x_n)=\mathsf{P}(Y=y|X_{k_1}=x_{k_1},\dots,X_{k_r}=x_{k_r}). \tag{53}$$

In the framework of medical applications it means that the factors $X_{k_1},\dots,X_{k_r}$ can be viewed as essential for provoking complex disease whereas the impact of others can be neglected. Any collection of such indexes $\{k_1,\dots,k_r\}$ is called *significant*. Clearly, if $\{k_1,\dots,k_r\}$ is a significant collection and if $\{k_1,\dots,k_r\}\subset\{m_1,\dots,m_p\}\subset\{1,\dots,n\}$ then $\{m_1,\dots,m_p\}$ is significant as well.

For $r=1,\dots,n$, set $\mathbb{X}_r=\{0,1\dots,s\}^r$. Thus $\mathbb{X}=\mathbb{X}_n$. Further we write $\alpha=(k_1,\dots,k_r)$, $X_\alpha=(X_{k_1},\dots,X_{k_r})$ and $x_\alpha=(x_{k_1},\dots,x_{k_r})$ where $x_i\in\{0,\dots,s\}$, $i=1,\dots,n$. For $x\in M$ and $y\in\mathbb{Y}$, formula (53) can be written as follows

$$\mathsf{P}(Y=y|X=x)=\mathsf{P}(Y=y|X_\alpha=x_\alpha). \tag{54}$$

Here $\mathsf{P}(X=x_\alpha)\ge\mathsf{P}(X=x)>0$ as $x\in M$. For $x\in\mathbb{X}$ and $y\in\mathbb{Y}$, let us define the vector $w^\alpha(x)$ with the components

$$w_y^\alpha(x)=\begin{cases}\psi(y)\mathsf{P}(Y=y,X_\alpha=x_\alpha), & x\in M,\\ 0, & x\notin M.\end{cases} \tag{55}$$

Note that (7) and (9) imply that, for each $y\in\mathbb{Y}$, $y>-m$, one has $L_y(x)=0$ if $x\notin M$. Introduce the functions $L_y^\alpha(x)$ according to (9) where instead of $w(x)$ we use $w^\alpha(x)$. In other words, for $x\in\mathbb{X}$,

$$L_y^\alpha(x)=(w^\alpha(x))^\top\Delta(y)=w_{-m}^\alpha(x)+\cdots+w_{y-1}^\alpha(x)-w_y^\alpha(x)-\cdots-w_m^\alpha(x)$$

where $y\in\mathbb{Y}$, $y>-m$. Then (54) yields that, for any $x\in M$ and $y=-m+1,\dots,m$, one has

$$L_y(x)=(w^\alpha(x))^\top\Delta(y)\mathsf{P}(X=x)/\mathsf{P}(X_\alpha=x_\alpha).$$

Consequently, $L_y(x)$ and $L_y^\alpha(x)$ for $x\in\mathbb{X}$ and $y\in\mathbb{Y}$ take positive or negative values or vanish simultaneously.

If (54) is valid then according to (11) the optimal function f_{opt} coincides with

$$f^\alpha(x)=\sum_{y\in\mathbb{Y}}y\,\mathbb{I}\{x\in A_y^\alpha\} \tag{56}$$

where

$$x\in A_y^\alpha\iff\begin{cases}L_{-m+1}^\alpha(x)\ge0, & & y=-m,\\ L_{y+1}^\alpha(x)\ge0, & L_y^\alpha(x)<0, & y\ne\pm m,\\ L_m^\alpha(x)<0, & & y=m.\end{cases} \tag{57}$$

for every $x\in\mathbb{X}$. Actually, $f^\alpha(x)$ depends on x_α only.

Now take any $\beta=(m_1,\dots,m_r)$ where $1\le m_1<\cdots<m_r\le n$ and apply (55)–(57) with β instead of α (we do not assume that collection $\{m_1,\dots,m_r\}$ is significant). Thus we obtain the function $f^\beta(x)$. Note that A_y^β, $y\in\mathbb{Y}$, form a partition of $\mathbb{X}$ (see (57) with α replaced by β) and we conclude that $f^\beta(x)$ is defined correctly for $x\in\mathbb{X}$. Moreover, if the collection of indexes α is significant then optimality of f^α implies that for any $\beta=(m_1,\dots,m_r)$ with $1\le m_1<\cdots<m_r\le n$ the following inequality is true

$$Err(f^\alpha)\le Err(f^\beta). \tag{58}$$

Let ψ be a penalty function. For any $\beta=(m_1,\dots,m_r)$ where $1\le m_1<\cdots<m_r\le n$, $x\in\mathbb{X}$ and a set $W_N\subset\{1,\dots,N\}$, introduce the random vector $\widetilde{w}^{\beta,W_N}(x,\omega)$ with the components

$$\widetilde{w}_y^{\beta,W_N}(x,\omega)=\frac{\psi(y)}{\sharp W_N}\sum_{j\in W_N}\mathbb{I}\{Y^j=y,X_\beta^j=x_\beta\},\quad y\in\mathbb{Y}. \tag{59}$$

Let the prediction algorithm be defined by the function $\widetilde{f}_{PA}^\beta$ such that

$$\widetilde{f}_{PA}^\beta(x,\xi_N(W_N,\omega))=\sum_{y\in\mathbb{Y}}y\,\mathbb{I}\{x\in\widetilde{A}_y^{\beta,W_N}(\omega)\} \tag{60}$$

where

$$x\in\widetilde{A}_y^{\beta,W_N}(\omega)\iff\begin{cases}\widetilde{L}_{-m+1}^{\beta,W_N}(x,\omega)\ge0, & & y=-m,\\ \widetilde{L}_{y+1}^{\beta,W_N}(x,\omega)\ge0, & \widetilde{L}_y^{\beta,W_N}(x,\omega)<0, & y\ne\pm m,\\ \widetilde{L}_m^{\beta,W_N}(x,\omega)<0, & & y=m,\end{cases} \tag{61}$$

and

$$\widetilde{L}_y^{\beta,W_N}(x,\omega):=(\widetilde{w}^{\beta,W_N}(x,\omega))^\top\Delta(y). \tag{62}$$

We write ω in (59)–(62) to emphasize the randomness of variables under consideration.

Lemma 3. *Let $f = f^\beta$ be defined by (56) (with β instead of α). Then for any $\beta = (m_1, \dots, m_r)$, $1 \le m_1 < \cdots < m_r \le n$, and $f_{PA} = \widetilde{f}^{\beta}_{PA}$, relation (16) holds when sets $W_N \subset \{1, \dots, N\}$ are such that $\sharp W_N \to \infty$ as $N \to \infty$. If, moreover, condition (29) is satisfied then $\widehat{Err}_K(f_{PA}, \xi_N)$ is an asymptotically unbiased estimate of $Err(f)$ as $N \to \infty$.*

Proof. For $u \in \mathbb{X}_r$ and $v_p \in (\mathbb{X} \times \mathbb{Y})^p$, introduce the functions

$$f^*(u) := f(x), \qquad f^*_{PA}(u, v_p) := f_{PA}(x, v_p)$$

where $u = x_\beta$. Note that, for each $x \in \mathbb{X}$, $w^\beta(x)$ depends only on x_β. Therefore f^* and f^*_{PA} are defined correctly as $f(x) = f(d)$ and $f_{PA}(x, v_p) = f_{PA}(d, v_p)$ for any $v_p \in (\mathbb{X} \times \mathbb{Y})^p$ $(1 \le p \le N)$ and $x, d \in \mathbb{X}$ such that $x_\beta = d_\beta$. Take

$$U = \{x \in \mathbb{X} : L^\beta_y(x) \ne 0 \text{ for all } y \in \mathbb{Y},\ y > -m\}. \tag{63}$$

For $x, d \in \mathbb{X}$ and $y \in \mathbb{Y}$, one has $L^\beta_y(x) = L^\beta_y(d)$ if $x_\beta = d_\beta$. Introduce $U^* = \{x_\beta : x \in U\}$. Thus we can apply reasoning as in Example 1 and Corollary 2 for f^* and f^*_{PA} defined on $\mathbb{X}_r$ with X_β, U^*, $\mathbb{X}_r(t, U^*)$ instead of X, U and $\mathbb{X}(t, U)$, respectively. To get the second statement of this lemma we use Remark 5. The proof is complete. □

Now in a similar way to (46) we define

$$\widehat{w}^{\beta, W_N}_y(x, \omega) := \frac{\widehat{\psi}(y, \xi_N(W_N))}{\sharp W_N} \sum_{j \in W_N} \mathbb{I}\{Y^j = y, X^j_\beta = x_\beta\}, \quad y \in \mathbb{Y}. \tag{64}$$

Set, for $x \in \mathbb{X}$, $y \in \mathbb{Y}$, $y > -m$, and $N \in \mathbb{N}$,

$$\widehat{L}^{\beta, W_N}_y(x, \omega) := (\widehat{w}^{\beta, W_N}(x, \omega))^\top \Delta(y). \tag{65}$$

Introduce

$$\widehat{f}^\beta_{PA}(x, \xi_N(W_N)) = \sum_{y \in \mathbb{Y}} y\, \mathbb{I}\{x \in \widehat{A}^{\beta, W_N}_y(\omega)\} \tag{66}$$

where $\widehat{A}^{\beta, N}_y(\omega)$ is defined by (61) with replacement of $\widetilde{L}^{\beta, W_N}_y(x, \omega)$ by $\widehat{L}^{\beta, W_N}_y(x, \omega)$.

Remark 7. It is easily seen that the assertion of Lemma 3 is valid if we choose $f_{PA} = \widehat{f}^\beta_{PA}$ instead of $f_{PA} = \widetilde{f}^\beta_{PA}$.

Theorem 2. *Let $\alpha = (k_1, \dots, k_r)$ where a significant collection $\{k_1, \dots, k_r\} \subset \{1, \dots, n\}$. Then, for any $\varepsilon > 0$ and each $\beta = (m_1, \dots, m_r)$ with $\{m_1, \dots, m_r\} \subset \{1, \dots, n\}$, the following inequality holds*

$$\widehat{Err}_K(\widehat{f}^\alpha_{PA}, \xi_N) \le \widehat{Err}_K(\widehat{f}^\beta_{PA}, \xi_N) + \varepsilon \quad a.s. \tag{67}$$

for all N large enough.

Proof. In view of Remark 7 this statement follows from Lemma 3 and relation (58). □

Remark 8. Theorem 2 suggests that it is reasonable to select for further analysis each collection $\{k_1, \dots, k_r\} \subset \{1, \dots, n\}$ as significant if $\widehat{Err}_K(\widehat{f}^\alpha_{PA}, \xi_N)$ with $\alpha = (k_1, \dots, k_r)$ has minimal value (or near the minimal value) among all $\widehat{Err}_K(\widehat{f}^\beta_{PA}, \xi_N)$ where $\beta = (m_1, \dots, m_r)$ and $\{m_1, \dots, m_r\} \subset \{1, \dots, n\}$. It is essential that we established relation (67) almost surely as we have to compare $\widehat{Err}_K(\widehat{f}^\beta_{PA}, \xi_N)$ for various $\beta = (m_1, \dots, m_r)$ simultaneously. Usually one considers models with large number of explanatory variables where the collection of significant factors is rather small. To estimate the predictive power of algorithm one uses the permutation tests, see, e.g., [8], possibly along with simulation. The measure of importance of subsets of factors is treated, e.g., in [22].

It is desirable to estimate the difference between $\widehat{Err}_K(\widehat{f}^\beta_{PA}, \xi_N)$ and $Err(f^\beta)$ as $N \to \infty$. This problem is considered in the next section for regularized versions of estimates.

6. Central limit theorem

Let $\beta = (m_1, \dots, m_r)$ where $1 \le m_1 < \cdots < m_r \le n$. We define the functions which can be viewed as the *regularized versions* of the estimates $\widehat{f}^\beta_{PA}$ of f^β (see (66) and (56)). Namely, for $\varepsilon = (\varepsilon_N)_{N \in \mathbb{N}}$ with non-random positive $\varepsilon_N \to 0$ as $N \to \infty$, put

$$\widehat{f}^\beta_{PA,\varepsilon}(x, \xi_N(W_N)) = \sum_{y \in \mathbb{Y}} y\, \mathbb{I}\{x \in \widehat{A}^{\beta, W_N}_{y,\varepsilon}(\omega)\}$$

where

$$W_N \subset \{1, \dots, N\}, \quad \sharp W_N \to \infty,\ N \to \infty, \tag{68}$$

and

$$x\in\widehat{A}_{y,\varepsilon}^{\beta,W_N}(\omega)\Longleftrightarrow\begin{cases}\widehat{L}_{-m+1}^{\beta,W_N}(x,\omega)+\varepsilon_N\geq 0, & & y=-m,\\ \widehat{L}_{y+1}^{\beta,W_N}(x,\omega)+\varepsilon_N\geq 0, & \widehat{L}_{y}^{\beta,W_N}(x,\omega)+\varepsilon_N<0, & y\neq\pm m,\\ \widehat{L}_{m}^{\beta,W_N}(x,\omega)+\varepsilon_N<0, & & y=m.\end{cases}$$

Here $\widehat{L}_y^{\beta,W_N}(x,\omega)$ (for $y\in\mathbb{Y}$, $y>-m$ and $\omega\in\Omega$) is defined in (65). To simplify notation we write $\widehat{A}_{y,\varepsilon}^{\beta,W_N}(\omega)$ instead of $\widehat{A}_{y,\varepsilon_N}^{\beta,W_N}(\omega)$. Note that assertion of Lemma 3 will be valid if we replace $\widetilde{f}_{PA}^{\beta}$ by $\widehat{f}_{PA,\varepsilon}^{\beta}$.

Now we turn to the central limit theorem for $\widehat{Err}_K(\widehat{f}_{PA,\varepsilon}^{\beta},\xi_N)$. In contrast to [3] we will consider not only nonbinary response variable Y but also an arbitrary penalty function ψ. It is quite natural that in this case we make some assumptions concerning the joint asymptotic behavior (as $N\to\infty$) of estimates $\widehat{\psi}(y,\xi_N(W_N))$, $y\in\mathbb{Y}$, and random variables $Y^j, f^{\beta}(X^j)$ with $j\in W_N$.

For $N\in\mathbb{N}$ and W_N set

$$\widehat{\theta}_N:=\left((\widehat{\psi}(N))^{\top},(\widehat{a}(N,-m))^{\top},\dots,(\widehat{a}(N,m))^{\top}\right)^{\top}$$

where $\widehat{\psi}(N)$ is the random vector with components $\widehat{\psi}_y(N)=\widehat{\psi}(y,\xi_N(W_N))$, $y\in\mathbb{Y}$, and $\widehat{a}(N,z)$, $z\in\mathbb{Y}$, are random vectors with components

$$\widehat{a}_y(N,z)=\frac{1}{\sharp W_N}\sum_{j\in W_N}\mathbb{I}\{Y^j=z,f^{\beta}(X^j)=y\},\quad y\in\mathbb{Y}.$$

To simplify notation we often write $\psi(N)$, $\widehat{\theta}_N$ and $\widehat{a}(N,z)$ instead of $\widehat{\psi}(N,W_N)$, $\widehat{\theta}_N(W_N)$ and $\widehat{a}^{\beta}(N,z,W_N)$, respectively. Define also

$$\theta:=\left(\psi^{\top},(a(-m))^{\top},\dots,(a(m))^{\top}\right)^{\top}$$

where ψ is non-random vector with components $\psi_y=\psi(y)$, $y\in\mathbb{Y}$, and $a(z)$, $z\in\mathbb{Y}$, are non-random vectors with components $a_y(z)=\mathsf{P}(Y=z,f^{\beta}(X)=y)$, $y\in\mathbb{Y}$. The same symbol is used here for a penalty function and a vector ψ because we simply arrange all values of a penalty function in a column. Note that $\widehat{\theta}_N$ and θ are vectors of dimension $(2m+1)(2m+2)$. We also introduce vector ν with components $\nu_y=(a(y))^{\top}q(y)$, $y\in\mathbb{Y}$, and vectors $\gamma(z)=\psi(z)q(z)$, $z\in\mathbb{Y}$. So, we can formulate the CLT.

Theorem 3. *Let $\varepsilon_N\to 0$ and $N^{1/2}\varepsilon_N\to\infty$ as $N\to\infty$. Take any vector $\beta=(m_1,\dots,m_r)$ with $1\leq m_1<\dots<m_r\leq n$, the corresponding function $f=f^{\beta}$ and prediction algorithm defined by $f_{PA}=\widehat{f}_{PA,\varepsilon}^{\beta}$. Assume that for any sets W_N satisfying (68) one has*

$$\sqrt{\sharp W_N}(\widehat{\theta}_N(W_N)-\theta)\xrightarrow{law}R\sim\mathcal{N}(0,C),\quad N\to\infty,\tag{69}$$

where $\mathcal{N}(0,C)$ stands for the multidimensional normal law. Then the following relation holds

$$\sqrt{N}(\widehat{Err}_K(f_{PA},\xi_N)-Err(f))\xrightarrow{law}Z\sim\mathcal{N}(0,\sigma^2),\quad N\to\infty.\tag{70}$$

Here $\sigma^2=\lambda^{\top}C\lambda$ and $\lambda:=\left(\nu^{\top},(\gamma(-m))^{\top},\dots,(\gamma(m))^{\top}\right)^{\top}$.

Proof. For a fixed $K\in\mathbb{N}$ and any $N\in\mathbb{N}$ set

$$T_N(f):=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\sharp S_k(N)}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\leq m}\psi(y)\sum_{j\in S_k(N)}\mathbb{I}\{Y^j=y,|f(X^j)-y|>i\},$$

$$\widehat{T}_N(f):=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\sharp S_k(N)}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\leq m}\widehat{\psi}_{N,k}(y)\sum_{j\in S_k(N)}\mathbb{I}\{Y^j=y,|f(X^j)-y|>i\}$$

where $\widehat{\psi}_{N,k}(y)=\widehat{\psi}(y,\xi_N(S_k(N)))$. One has

$$\widehat{Err}_K(f_{PA},\xi_N)-Err(f)=(\widehat{Err}_K(f_{PA},\xi_N)-\widehat{T}_N(f))+(\widehat{T}_N(f)-T_N(f))+(T_N(f)-Err(f)).$$

First of all we show that

$$\sqrt{N}(\widehat{Err}_K(f_{PA},\xi_N)-\widehat{T}_N(f))\xrightarrow{\mathsf{P}}0,\quad N\to\infty.\tag{71}$$

Using (21) one can write

$$\widehat{Err}_K(f_{PA},\xi_N)-\widehat{T}_N(f)=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\sharp S_k(N)}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\leq m}\widehat{\psi}_{N,k}(y)\sum_{j\in S_k(N)}\mathbb{I}\{Y^j=y\}F_{N,k}^{(i)}(X^j,y).\tag{72}$$

We define the random variables

$$G^{(i)}_{N,k}(y) := \frac{1}{\sqrt{\sharp S_k(N)}} \sum_{j\in S_k(N)} \mathbb{I}\{Y^j = y\} F^{(i)}_{N,k}(X^j, y)$$

and verify that for each $k = 1, \ldots, K$

$$\sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \widehat{\psi}_{N,k}(y) G^{(i)}_{N,k}(y) \xrightarrow{\mathsf{P}} 0, \quad N \to \infty. \tag{73}$$

Clearly (73) implies (71) in view of (72) as $\sharp S_k(N) = [N/K]$ for $k = 1, \ldots, K-1$ and $[N/K] \le \sharp S_K(N) < [N/K] + K$. For each i, N, k under consideration and U defined by (63) write $G^{(i)}_{N,k}(y) = G^{(i),U}_{N,k}(y) + G^{(i),\mathbb{X}\setminus U}_{N,k}(y)$ where

$$G^{(i),V}_{N,k}(y) = \frac{1}{\sqrt{\sharp S_k(N)}} \sum_{j\in S_k(N)} \mathbb{I}\{X^j \in V\}\mathbb{I}\{Y^j = y\} F^{(i)}_{N,k}(X^j, y), \quad V \subset \mathbb{X}.$$

Obviously,

$$|G^{(i),U}_{N,k}(y)| \le \sqrt{\sharp\, S_k(N)} \sum_{x\in U} \left| \mathbb{I}\{|f_{PA}(x, \xi_N(\overline{S_k(N)})) - y| > i\} - \mathbb{I}\{|f(x) - y| > i\} \right|.$$

Functions f_{PA} and f take values in the set $\mathbb{Y}$. Thus, for any $x \in U$, $k = 1, \ldots, K$ and almost all $\omega \in \Omega$ relation (17) with U given by (63) ensures the existence of an integer $N_1(x, k, \omega)$ such that $f_{PA}(x, \xi_N(\overline{S_k(N)})) = f(x)$ for $N \ge N_1(x, k, \omega)$. Hence $G^{(i),U}_{N,k}(y) = 0$ for any y belonging to $\mathbb{Y}$, each $i = 0, \ldots, 2m-1$, $k = 1, \ldots, K$ and almost all $\omega \in \Omega$ when $N \ge N_1(\omega) = \max_{x\in U, k=1,\ldots,K} N_1(x, k, \omega)$. Evidently, $N_1 < \infty$ a.s. because $\sharp\mathbb{X} < \infty$. We obtain that

$$\sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \widehat{\psi}_{N,k}(y) G^{(i),U}_{N,k}(y) \to 0 \quad \text{a.s.}, \ N \to \infty. \tag{74}$$

If $U = \mathbb{X}$ then $G^{(i),\mathbb{X}\setminus U}_{N,k}(y) = 0$ for all i, N, k and y under consideration. Consequently, (73) is valid and thus, for $U = \mathbb{X}$, relation (71) holds. Let now $U \ne \mathbb{X}$. In view of (4) we can claim that

$$\mathbb{X} \setminus U = \cup_{J\subset\mathbb{Y}, \sharp J>1} B_J. \tag{75}$$

We have seen in Section 2 that each B_J appearing in (75) can be represented by way of $B_J = D_{t,z}$, for some $t, z \in \mathbb{Y}$ such that $t < z$, with $J = \{y \in \mathbb{Y} : t \le y \le z\}$ and

$$D_{t,z} := \{x \in \mathbb{X} \setminus U :\ L^{\beta}_y(x) < 0, -m < y \le t;\ L^{\beta}_y(x) = 0, t < y \le z;\ L^{\beta}_y(x) > 0, y > z\}. \tag{76}$$

Then for $k = 1, \ldots, K$ and $N \in \mathbb{N}$ one has

$$\sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \widehat{\psi}_{N,k}(y) G^{(i),\mathbb{X}\setminus U}_{N,k}(y) = \sum_{z=-m+1}^{m} \sum_{t=-m}^{z-1} \sum_{x\in D_{t,z}} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \Phi^{(i)}_{N,k}(x, y).$$

Here

$$\Phi^{(i)}_{N,k}(x, y) := \frac{\widehat{\psi}_{N,k}(y)}{\sqrt{\sharp S_k(N)}} \sum_{j\in S_k(N)} \mathbb{I}\{X^j = x, Y^j = y\} F^{(i)}_{N,k}(x, y).$$

Using (27), (28) and (35) we come to the formula

$$\sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \Phi^{(i)}_{N,k}(x, y) = \sqrt{\sharp S_k(N)}\, (\widehat{L}^{S_k(N)}(x))^{\top} I(N, x, k, t)$$

where the components of random vector $I(N, x, k, t)$ are given in (30).

If $x \in (\mathbb{X} \setminus U) \cap (\mathbb{X} \setminus M)$ then $\Phi^{(i)}_{N,k}(x, y) = 0$ a.s. for all i, N, k and y under consideration. Let now $x \in (\mathbb{X} \setminus U) \cap M$, i.e. $x \in M \cap D_{t,z}$ for some $t, z \in \mathbb{Y}$ such that $t < z$. Then $f(x) = t$ according to (76). We will show that

$$\sqrt{\sharp S_k(N)}\, (\widehat{L}^{S_k(N)}(x))^{\top} I(N, x, k, t) \xrightarrow{\mathsf{P}} 0, \quad N \to \infty. \tag{77}$$

One has $(\widehat{L}^{S_k(N)}(x))^{\top} I(N, x, k, t) = \widehat{R}^{(1)}_{N,k}(x, t) + \widehat{R}^{(2)}_{N,k}(x, t)$ where

$$\widehat{R}^{(l)}_{N,k}(x, t) = \sum\nolimits^{(l)} \widehat{L}^{S_k(N)}_y(x) I_y(N, x, k, t), \quad l = 1, 2.$$

Here $\sum^{(1)}$ and $\sum^{(2)}$ are taken over $y \in (\mathbb{Y} \setminus \{-m\}) \setminus (t, z]$ and $y \in (\mathbb{Y} \setminus \{-m\}) \cap (t, z]$, respectively. Clearly,

$$|\widehat{R}^{(1)}_{N,k}(x, t)| \le \sum\nolimits^{(1)} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \notin [t, z]\}\, |\widehat{L}^{S_k(N)}_y(x)|.$$

For any $x \in D_{t,z}$, $k = 1, \dots, K$ and almost all $\omega \in \Omega$ relations (65) and (66) ensure the existence of an integer $N_3(x, k, \omega)$ such that $f_{PA}(x, \xi_N(\overline{S_k(N)}, \omega)) \in [t, z]$ for $N \ge N_3(x, k, \omega)$. Hence $\widehat{R}^{(1)}_{N,k}(x, \omega) = 0$ for any $x \in D_{t,z}$, each $k = 1, \dots, K$ and almost all $\omega \in \Omega$ when $N \ge N_3(\omega) = \max_{x \in D_{t,z}, k=1,\dots,K} N_3(x, k, \omega)$. Thus

$$\sqrt{\sharp\, S_k(N)}\, \widehat{R}^{(1)}_{N,k}(x, t) \to 0 \quad \text{a.s.,}\;\; N \to \infty. \tag{78}$$

Further on

$$|\widehat{R}^{(2)}_{N,k}(x, t)| \le \sum\nolimits^{(2)} \mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \ne t\}\, |\widehat{L}^{S_k(N)}_y(x)|. \tag{79}$$

Let us prove that, for any $x \in M \cap D_{t,z}$ and $k = 1, \dots, K$,

$$\mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \ne t\} \xrightarrow{\mathsf{P}} 0, \quad N \to \infty. \tag{80}$$

For any $\varkappa > 0$ we have

$$\mathsf{P}(\mathbb{I}\{f_{PA}(x, \xi_N(\overline{S_k(N)})) \ne t\} > \varkappa) = \mathsf{P}\left(\{\widehat{L}^{\beta,\overline{S_k(N)}}_t(x) + \varepsilon_N \ge 0\} \cup \{\widehat{L}^{\beta,\overline{S_k(N)}}_{t+1}(x) + \varepsilon_N < 0\}\right).$$

For almost every $\omega \in \Omega$ there exists a positive integer $N_4 = N_4(x, k, \omega)$ such that the inequality $\widehat{L}^{\beta,\overline{S_k(N)}}_t(x, \omega) + \varepsilon_N < 0$ holds for $x \in D_{t,z}$ and $N \ge N_4$. Taking into account that, for any events F and H, one has $\mathbb{I}\{F \cup H\} = \mathbb{I}\{F\} + \mathbb{I}\{H\} - \mathbb{I}\{F \cap H\}$ we see that validity of (80) is equivalent to

$$\mathsf{P}(\widehat{L}^{\beta,\overline{S_k(N)}}_{t+1}(x) + \varepsilon_N < 0) \to 0, \quad N \to \infty,$$

or

$$\mathsf{P}\Big(\sqrt{\sharp\overline{S_k(N)}}\,\widehat{L}^{\beta,\overline{S_k(N)}}_{t+1}(x) < -\varepsilon_N\sqrt{\sharp\overline{S_k(N)}}\Big) \to 0, \quad N \to \infty. \tag{81}$$

Consider sets $W_N \subset \{1, \dots, N\}$ such that $\sharp\, W_N \to \infty$ as $N \to \infty$. Obviously, for any $x \in \mathbb{X}$ and $y \in \mathbb{Y}$,

$$\widehat{L}^{\beta,W_N}_y(x) = \widetilde{L}^{\beta,W_N}_y(x) + \left(\widehat{L}^{\beta,W_N}_y(x) - \widetilde{L}^{\beta,W_N}_y(x)\right) \tag{82}$$

and

$$\widetilde{L}^{\beta,W_N}_y(x) = \frac{1}{\sharp W_N}\sum_{j \in W_N} \widetilde{L}^{\beta,j}_y(x).$$

Here we write $\widetilde{L}^{\beta,j}_y(x)$ instead of $\widetilde{L}^{\beta,\{j\}}_y(x)$. One has $\mathsf{E}\widetilde{L}^{\beta,j}_y(x) = L^{\beta}_y(x) = 0$ for all $j \in \mathbb{N}$, $x \in D_{t,z}$ and $y \in (t, z]$. The CLT holds for an array of bounded centered i.i.d. random variables $\{\widetilde{L}^{\beta,j}_y(x), j \in W_N, N \in \mathbb{N}\}$. Namely, for a given β and all considered x and y,

$$Z_{N,1}(x, y; \beta) := \sqrt{\sharp W_N}\, \widetilde{L}^{\beta,W_N}_y(x) \xrightarrow{law} Z_1(x, y; \beta) \sim \mathcal{N}(0, \sigma_1^2(x, y; \beta)), \quad N \to \infty, \tag{83}$$

where $\sigma_1^2(x, y; \beta) = var\, \widetilde{L}^{\beta,j}_y(x), j \in W_N$. Further on

$$\begin{aligned}\sqrt{\sharp\, W_N}(\widehat{w}^{\beta,W_N}_y(x) - \widetilde{w}^{\beta,W_N}_y(x)) &= (\widehat{\psi}(y, \xi_N(W_N)) - \psi(y))\frac{1}{\sqrt{\sharp W_N}}\sum_{j \in W_N} \mathbb{I}\{X^j_\beta = x_\beta, Y^j = y\}.\\ &= H_{N,2}(x, y; \beta) + H_{N,3}(x, y; \beta)\end{aligned}$$

where

$$H_{N,2}(x, y; \beta) = (\widehat{\psi}(y, \xi_N(W_N)) - \psi(y))\sqrt{\sharp W_N}\,\mathsf{P}(X_\beta = x_\beta, Y = y),$$
$$H_{N,3}(x, y; \beta) = (\widehat{\psi}(y, \xi_N(W_N)) - \psi(y))\frac{1}{\sqrt{\sharp W_N}}\sum_{j \in W_N}(\mathbb{I}\{X^j_\beta = x_\beta, Y^j = y\} - \mathsf{E}\mathbb{I}\{X^j_\beta = x_\beta, Y^j = y\}).$$

For $i = 2, 3$, we consider the random vector $H_{N,i}(x; \beta)$ with components $H_{N,i}(x, y; \beta)$, $y \in \mathbb{Y}$. In view of (69) we know that

$$H_{N,2}(x; \beta) \xrightarrow{law} H_2(x; \beta) \sim \mathcal{N}(0, C_2(x; \beta)), \quad N \to \infty.$$

Here the matrix $C_2(x;\beta)$ has entries $d_{u,v}\mu_u\mu_v$ and $D=(d_{u,v})$ is the matrix coinciding with $(2m+1)\times(2m+1)$ left upper corner of the initial matrix C introduced in (69), the vector $\mu=\mu(x;\beta)$ has components $\mu_u=\mathsf{P}(X_\beta=x_\beta, Y=u)$ and $u,v\in\mathbb{Y}$.

Due to SLLNA one has

$$\frac{1}{\sharp W_N}\sum_{j\in W_N}(\mathbb{I}\{X_\beta^j=x_\beta, Y^j=y\}-\mathsf{E}\mathbb{I}\{X_\beta^j=x_\beta, Y^j=y\})\to 0\quad \text{a.s.},\ N\to\infty.$$

Hence, for any $x,y\in\mathbb{Y}$,

$$H_{N,3}(x,y;\beta)\xrightarrow{\mathsf{P}}0,\quad N\to\infty. \tag{84}$$

We can write

$$\sqrt{\sharp W_N}\left(\widehat{L}_y^{\beta,W_N}(x)-\widetilde{L}_y^{\beta,W_N}(x)\right)=Z_{N,2}(x,y;\beta)+Z_{N,3}(x,y;\beta) \tag{85}$$

where according to (42) and (47)

$$Z_{N,i}(x,y;\beta)=(H_{N,i}(x;\beta))^\top\Delta(y),\quad i=2,3.$$

Consequently,

$$Z_{N,2}(x,y;\beta)\xrightarrow{law}Z_2(x,y;\beta)\sim\mathcal{N}(0,\sigma_2^2(x,y;\beta)),\quad N\to\infty,$$

with $\sigma_2^2(x,y;\beta)=(\Delta(y))^\top C_2(x;\beta)\Delta(y)$. By virtue of (84)

$$Z_{N,3}(x,y;\beta)\xrightarrow{\mathsf{P}}0,\quad N\to\infty. \tag{86}$$

Thus in view of (82), (83) and (85), for each $y\in(t,z]$ and $x\in D_{t,z}$, we have

$$\mathsf{P}\Big(\sqrt{\sharp W_N}\,\widehat{L}_y^{\beta,W_N}(x)<-\varepsilon_N\sqrt{\sharp W_N}\Big)\le\sum_{i=1}^{3}\mathsf{P}\left(Z_{N,i}(x,y;\beta)<-\frac{\varepsilon_N}{3}\sqrt{\sharp W_N}\right). \tag{87}$$

Let $(Z_N)_{N\in\mathbb{N}}$ be a sequence of random variables such that $Z_N\xrightarrow{law}Z$ as $N\to\infty$ where Z has a (possibly degenerate) Gaussian distribution. Then, for any sequence of real numbers $(c_N)_{N\in\mathbb{N}}$ satisfying relation $c_N\to-\infty$, one has $\mathsf{P}(Z_N<c_N)\to 0$, $N\to\infty$. The latter statement becomes obvious if Z is degenerate. For Z having continuous distribution function we take into account that the distribution functions of Z_N, $N\in\mathbb{N}$, converge to the distribution function of Z uniformly on $\mathbb{R}$. Hence, taking $(\varepsilon_N)_{N\in\mathbb{N}}$ such that $\varepsilon_N\sqrt{\sharp W_N}\to\infty$, as $N\to\infty$, we establish that the right-hand side of (87) tends to 0.

Put $y=t+1$ (it is possible because $y\in(t,z]\cap\mathbb{Y}$ where $t,z\in\mathbb{Y}$ and $t<z$), $W_N=\overline{S_k(N)}$, $k=1,\dots,K$. Note that $\sharp\overline{S_k(N)}\ge(K-1)[N/K]$ for each $k=1,\dots,K$. We conclude that (81) is satisfied when $\varepsilon_N N^{1/2}\to\infty$ as $N\to\infty$. Thus, we come to (80).

Note that, $\widehat{L}_y^{W_N}(x)=\widehat{L}_y^{\beta,W_N}(x)$ if $\beta=(1,\dots,n)$, for $x\in\mathbb{X}$ and $y\in\mathbb{Y}$ such that $y>-m$. Therefore using (82)–(86) we obtain, for $x\in D_{t,z}$ and $y\in(t,z]\cap\mathbb{Y}$, the estimate

$$\sqrt{\sharp W_N}\,|\widehat{L}_y^{W_N}(x)|\le\sum_{i=1}^{3}|\overline{Z}_{N,i}(x,y)| \tag{88}$$

where $\overline{Z}_{N,i}(x,y)$ is the same as $Z_{N,i}(x,y;\beta)$ evaluated for $\beta=(1,\dots,n)$ and $i=1,2,3$. Hence setting now $W_N=S_k(N)$ we see that (78)–(88) lead to (77). Thus we have

$$\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\widehat{\psi}_{N,k}(y)G_{N,k}^{(i),\mathbb{X}\setminus U}(y)\xrightarrow{\mathsf{P}}0,\quad N\to\infty. \tag{89}$$

Taking into account (74) and (89) we come to (73). Consequently, (71) is verified.

Now we turn to the study of $\widehat{T}_N(f)-T_N(f)$. One has

$$\begin{aligned}&\sqrt{N}(\widehat{T}_N(f)-T_N(f))\\&=\frac{\sqrt{N}}{K}\sum_{k=1}^{K}\frac{1}{\sharp S_k(N)}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}(\widehat{\psi}_{N,k}(y)-\psi(y))\sum_{j\in S_k(N)}\mathbb{I}\{Y^j=y, |f(X^j)-y|>i\}.\end{aligned}$$

Put $Z_i^j(y) = \mathbb{I}\{Y^j = y, |f(X^j) - y| > i\}, i = 0, \dots, 2m-1, j = 1, \dots, N, y \in \mathbb{Y}$. For each $k = 1, \dots, K$

$$\begin{aligned}
&\sqrt{N} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} (\widehat{\psi}_{N,k}(y) - \psi(y)) \frac{1}{\sharp\, S_k(N)} \sum_{j \in S_k(N)} \mathbb{I}\{Y^j = y, |f(X^j) - y| > i\} \\
&= \sqrt{N} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} (\widehat{\psi}_{N,k}(y) - \psi(y)) \frac{1}{\sharp\, S_k(N)} \sum_{j \in S_k(N)} (Z_i^j(y) - \mathsf{E} Z_i^j(y)) \\
&\quad + \sqrt{N} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} (\widehat{\psi}_{N,k}(y) - \psi(y)) \mathsf{P}(Y = y, |f(X) - y| > i).
\end{aligned}$$

Since $N/\sharp\, S_k(N) \to K$ for each $k = 1, \dots, K$, as $N \to \infty$, we conclude, taking into account SLLNA and relation (69), that

$$\sqrt{N} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} (\widehat{\psi}_{N,k}(y) - \psi(y)) \frac{1}{\sharp\, S_k(N)} \sum_{j \in S_k(N)} (Z_i^j(y) - \mathsf{E} Z_i^j(y)) \xrightarrow{\mathsf{P}} 0, \quad N \to \infty.$$

We can write

$$\begin{aligned}
&\sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} (\widehat{\psi}_{N,k}(y) - \psi(y)) \mathsf{P}(Y = y, |f(X) - y| > i) \\
&= \sum_{y \in \mathbb{Y}} (\widehat{\psi}_{N,k}(y) - \psi(y))(a(y))^\top q(y) = (\widehat{\psi}(N, S_k(N)) - \psi)^\top \nu.
\end{aligned}$$

Consequently the limit distribution of $\sqrt{N}[(\widehat{T}_N(f) - T_N(f)) + (T_N(f) - Err(f))]$ will be the same as that for random variables

$$\sqrt{N}[(T_N(f) - Err(f)) + \frac{1}{K} \sum_{k=1}^{K} (\widehat{\psi}(N, S_k(N)) - \psi)^\top \nu].$$

Note that

$$\begin{aligned}
(T_N(f) - Err(f)) &= \frac{1}{K} \sum_{k=1}^{K} \sum_{z \in \mathbb{Y}} \psi(z)(a(N, z, S_k(N)) - a(z))^\top q(z) \\
&= \frac{1}{K} \sum_{k=1}^{K} \sum_{z \in \mathbb{Y}} (a(N, z, S_k(N)) - a(z))^\top \gamma(z).
\end{aligned}$$

Therefore

$$\sqrt{N}[(T_N(f) - Err(f)) + \frac{1}{K} \sum_{k=1}^{K} (\widehat{\psi}(N, S_k(N)) - \psi)^\top \nu] = \frac{\sqrt{N}}{K} \sum_{k=1}^{K} (\widehat{\theta}_N(S_k(N)) - \theta)^\top \lambda.$$

According to (69), for each $k = 1, \dots, K$, one has

$$\sqrt{\sharp S_k(N)} (\widehat{\theta}_N(S_k(N)) - \theta)^\top \lambda \xrightarrow{law} R_k \sim N(0, \lambda^\top C \lambda).$$

Since, for each $N \in \mathbb{N}$, collections of random variables $\xi_N(S_1(N)), \dots, \xi_N(S_K(N))$ are independent we can claim that $R_1, \dots, R_K$ are independent. Again recalling that $N/\sharp S_k(N) \to K$ for $k = 1, \dots, K$, as $N \to \infty$, we come to (70). The proof is complete. □

Remark 9. In Theorem 3 we can relax condition (69) by employing only $W_N = S_k(N)$ and $W_N = \overline{S_k(N)}$ for $k = 1, \dots, K$ and $N \in \mathbb{N}$.

Corollary 3. *Let ψ be the penalty function defined in (49) and $\widehat{\psi}_{N,k}(y)$ be the estimate introduced in (50) with $W_N = S_k(N)$, $k = 1, \dots, K$ where $K \in \mathbb{N}$. Let $\varepsilon_N \to 0$ and $N^{1/2}\varepsilon_N \to \infty$ as $N \to \infty$. Then, for any vector $\beta = (m_1, \dots, m_r)$ with $1 \le m_1 < \dots < m_r \le n$, the corresponding function $f = f^\beta$ and prediction algorithm defined by $f_{PA} = \widehat{f}^{\beta}_{PA,\varepsilon}$, the following relation holds*

$$\sqrt{N}(\widehat{Err}_K(f_{PA}, \xi_N) - Err(f)) \xrightarrow{law} Z \sim \mathcal{N}(0, \sigma^2), \quad N \to \infty. \tag{90}$$

Here σ^2 is variance of the random variable

$$V = \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \frac{\mathbb{I}\{Y = y\}}{\mathsf{P}(Y = y)} \left(\mathbb{I}\{|f(X) - y| > i\} - \mathsf{P}(|f(X) - y| > i \big| Y = y) \right). \tag{91}$$

It is easily seen that conditions of Theorem 3 are satisfied for this particular but important choice of ψ and its estimates. However, it seems more simple to note that, for each $y \in \mathbb{Y}$ and $k = 1, \ldots, K$,

$$\widehat{\mathsf{P}}_{S_k(N)}(Y = y) \xrightarrow{\mathsf{P}} \mathsf{P}(Y = y),$$
$$\sqrt{\sharp S_k(N)}(\widehat{\mathsf{P}}_{S_k(N)}(Y = y) - \mathsf{P}(Y = y)) \xrightarrow{law} Z_4(y) \sim \mathcal{N}(0, \sigma_4^2(y)),$$

as $N \to \infty$, where $\sigma_4^2(y) = \mathsf{P}(Y \neq y)\mathsf{P}(Y = y)$. Then one can employ the Slutsky lemma to show that the limit behavior of $\sqrt{N}[(\widehat{T}_N(f) - T_N(f)) + (T_N(f) - Err(f))]$ will be the same as for

$$\frac{\sqrt{N}}{K} \sum_{k=1}^{K} \frac{1}{\sharp S_k(N)} \sum_{j \in S_k(N)} (V^j - \mathsf{E}V^j)$$

where i.i.d. random variables $V^j, j \in \mathbb{N}$ are defined by way of

$$V^j = \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \frac{\mathbb{I}\{Y^j = y\}}{\mathsf{P}(Y = y)} \left(\mathbb{I}\{|f(X^j) - y| > i\} - \frac{\mathsf{P}(Y = y, |f(X) - y| > i)}{\mathsf{P}(Y = y)} \right).$$

Thus we come to the statement of Corollary 3.

Recall that for a sequence of random variables $(\eta_N)_{N\in\mathbb{N}}$ and a sequence of positive numbers $(c_N)_{N\in\mathbb{N}}$ one writes $\eta_N = o_P(c_N)$ if $\eta_N/c_N \xrightarrow{\mathsf{P}} 0, N \to \infty$.

Remark 10. As usual one can view the CLT as a result describing the exact rate of approximation for random variables under consideration. Theorem 3 implies that

$$\widehat{Err}_K(f_{PA}, \xi_N) - Err(f) = o_P(c_N), \quad N \to \infty, \tag{92}$$

where $c_N = o(N^{-1/2})$. The last relation is optimal in a sense whenever $\sigma^2 > 0$, in other words it is impossible to take $c_N = O(N^{-1/2})$ in (92). One can verify that the same asymptotic result as in Corollary 3 is true if $\widehat{\psi}_{N,k}(y)$ is defined according to (50) with $W_N = \overline{S_k(N)}, k = 1, \ldots, K, N \in \mathbb{N}$.

Remark 11. Using (91) or Theorem 3 one can show that

$$\sigma^2 = \sum_{y\in\mathbb{Y}} \frac{1}{(\mathsf{P}(Y = y))^2} \left[(a(y))^\top (q(y) \circ q(y)) - \frac{1}{\mathsf{P}(Y = y)} \left((a(y))^\top q(y)\right)^2 \right]$$

where $\circ$ stands for the Hadamard product of two vectors, i.e. $q(y) \circ q(y)$ has components $q(y)_z^2, z \in \mathbb{Y}$. Therefore it is not difficult to construct the consistent estimates $\widehat{\sigma}_N$ of unknown σ appearing in (90) and (if $\sigma^2 \neq 0$) we can claim that under conditions of Corollary 3

$$\frac{\sqrt{N}}{\widehat{\sigma}_N}(\widehat{Err}_K(f_{PA}, \xi_N) - Err(f)) \xrightarrow{law} \frac{Z}{\sigma} \sim \mathcal{N}(0, 1), \quad N \to \infty.$$

Now we consider the multidimensional version of Corollary 3. Employing the Cramér–Wold device and the proof of Theorem 3 we come to the following statement.

Corollary 4. *Let conditions of Corollary 3 be satisfied. Then, for any $\alpha(l) = (m_1^{(l)}, \ldots, m_r^{(l)})$ such that $1 \le m_1^{(l)} < \cdots < m_r^{(l)} \le n$ where $l = 1, \ldots, j, j \in \mathbb{N}$, and each $K \in \mathbb{N}$, one has*

$$\sqrt{N}(Z_N^{(1)}, \ldots, Z_N^{(j)})^\top \xrightarrow{law} \mathsf{Z} \sim \mathcal{N}(0, B), \quad N \to \infty.$$

Here $Z_N^{(l)} = \widehat{Err}_K(\widehat{f}_{PA,\varepsilon}^{\alpha(l)}, \xi_N) - Err(f^{\alpha(l)}), l = 1, \ldots, j$, and the elements of covariance matrix $B = (b_{l,p})$ have the form

$$b_{l,p} = cov(V(\alpha(l)), V(\alpha(p))), \quad l, p = 1, \ldots, j,$$

the random variables $V(\alpha(l))$ being defined in the same way as V in (91) with f^β replaced by $f^{\alpha(l)}$.

To conclude we note (see also Remark 11) that one can construct the consistent estimates $\widehat{B}_N$ of the unknown (nondegenerate) covariance matrix B to obtain the statistical version of the last theorem. Namely, under conditions of Corollary 4 the following relation is valid

$$(\widehat{B}_N)^{-1/2}(Z_N^{(1)}, \ldots, Z_N^{(j)})^\top \xrightarrow{law} B^{-1/2}\mathsf{Z} \sim \mathcal{N}(0, I), \quad N \to \infty,$$

where I stands for the unit matrix of order j.

It is worth mentioning that in [5] we demonstrate by simulation that our method leads to correct identification of significant factors even for samples having rather modest size.

Acknowledgments

The authors are grateful to two reviewers for their valuable remarks and suggestions aimed to improve the exposition of the paper. This work is partially supported by RFBR grant 13-01-00612.

References

[1] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79.

[2] A.V. Bulinski, On foundation of the dimensionality reduction method for explanatory variables, J. Math. Sci. (2014). http://dx.doi.org/10.1007/s10958-014-1838-7.

[3] A.V. Bulinski, Central limit theorem related to MDR method. Proceedings of the Fields Institute International Symposium on Asymptotic Methods in Stochastics, in Honour of Miklós Csörgő's Work on the occasion of his anniversary, (2015) in press. arXiv:1301.6609 [math.PR].

[4] A. Bulinski, O. Butkovsky, V. Sadovnichy, A. Shashkin, P. Yaskov, A. Balatskiy, L. Samokhodskaya, V. Tkachuk, Statistical methods of SNP data analysis and applications, Open J. Stat. 2 (1) (2012) 73–87.

[5] A. Bulinski, A. Rakitko, Simulation and analytical approach to the identifiaction of significant factors, arXiv:1406.1138 [math.ST].

[6] W.S. Bush, S.M. Dudek, M.D. Ritchie, Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene–gene interactions, Bioinformatics 22 (2006) 2173–2174.

[7] T.L. Edwards, E.S. Torstenson, E.M. Martin, M.D. Ritchie, A cross-validation procedure for general pedigrees and matched odds ratio fitness metric implemented for the multifactor dimensionality reduction pedigree disequilibrium test MDR-PDT and cross-validation: power studies, Genet. Epidemiol. 34 (2) (2010) 194–199.

[8] P. Golland, F. Liang, S. Mukherjee, D. Panchenko, Permutation tests for classification, Lect. Notes Comput. Sci. 3559 (2005) 501–515.

[9] J. Gui, A.S. Andrew, P. Andrews, H.M. Nelson, K.T. Kelsey, M.R. Karagas, J.H. Moore, A robust multifactor dimensionality reduction method for detecting gene–gene interactions with application to the genetic analysis of bladder cancer susceptibility, Ann. Hum. Genet. 75 (1) (2011) 20–28.

[10] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning; Data Mining, Inference and Prediction, second ed., Springer, New York, 2008.

[11] S. Lee, M.P. Epstein, R. Duncan, X. Lin, Sparse principal component analysis for identifying ancestry-informative markers in genome wide association studies, Genet. Epidemiol. 36 (2012) 293–302.

[12] R. Lockhart, J. Taylor, R.J. Tibshirani, R. Tibshirani, A significance test for the lasso. arXiv:1301.7161 [math.ST].

[13] R.S. Michalski, A theory and methodology of inductive learning, Artif. Intel. 20 (1983) 111–161.

[14] J.B. Moore, F.W. Asselbergs, S.M. Williams, Bioinformatics challenges for genome-wide association studies, Bioinformatics 26 (2010) 445–455.

[15] A. Niu, S. Zhang, Q. Sha, A novel method to detect gene–gene interactions in structured populations: MDR-SP, Ann. Hum. Genet. 75 (6) (2011) 742–754.

[16] S. Oh, J. Lee, M.-S. Kwon, B. Weir, K. Ha, T. Park, A novel method to identify high order gene–gene interactions in genome-wide association studies: Gene-based MDR, BMC Bioinformatics 13 ((Suppl. 9):S5) (2012).

[17] J. Park, Independent rule in classification of multivariate binary data, J. Multivar. Anal. 100 (2009) 2270–2286.

[18] M.D. Ritchie, L.W. Hahn, N. Roodi, R.L. Bailey, W.D. Dupont, F.F. Parl, J.H. Moore, Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, Am. J. Hum. Genet. 69 (1) (2001) 138–147.

[19] M.D. Ritchie, A.A. Motsinger, Multifactor dimensionality reduction for detecting gene–gene and gene-environment interactions in pharmacogenomics studies, Pharmacogenomics 6 (8) (2005) 823–834.

[20] I. Ruczinski, C. Kooperberg, M. Leblanc, Logic regression, J. Comput. Graph. Statist. 12 (3) (2003) 475–511.

[21] H. Schwender, I. Ruczinski, Logic regression and its extensions, Adv. Genet. 72 (2010) 25–45.

[22] H. Schwender, I. Ruczinski, K. Ickstadt, Testing SNPs and sets of SNPs for importance in association studies, Biostatistics 12 (1) (2011) 18–32.

[23] K. Sikorska, E. Lesaffre, P.F.G. Groenen, P.H.C. Eilers, GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies, BMC Bioinformatics 14 (2013) 166.

[24] R.L. Taylor, T.-C. Hu, Strong laws of large numbers for arrays of row-wise independent random elements, Int. J. Math. Math. Sci. 10 (4) (1987) 805–814.

[25] R.J. Tibshirani, J. Taylor, Degrees of freedom in lasso problems, Ann. Statist. 40 (2012) 1198–1232.

[26] D.R. Velez, B.C. White, A.A. Motsinger, W.S. Bush, M.D. Ritchie, S.M. Williams, J.H. Moore, A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, Genet. Epidemiol. 31 (4) (2007) 306–315.

[27] P.M. Visscher, M.A. Brown, M.I. McCarthy, J. Yang, Five years of GWAS discovery, Am. J. Hum. Genet. 90 (2012) 7–24.