S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru

Том 18, №1, стр. 720–728 (2021) УДК 519.2
DOI 10.33048/semi.2021.18.052 MSC 62G20, 62H12

FEATURE SELECTION BASED ON STATISTICAL ESTIMATION OF MUTUAL INFORMATION

A.A. KOZHEVIN

Abstract. An algorithm to identify significant factors is proposed in the mixed model framework. It employs statistical estimation of mutual information. Consistency of this procedure is established. Numerical experiments demonstrating its accuracy supplement theoretical results.

Keywords: feature selection, mixed model, mutual information, conditional Shannon entropy, logistic regression.

1. Introduction

Feature selection plays important role in various domains, see, e.g., the books [1], [7], [11] and references therein. The paper is devoted to a certain feature selection procedure based on information theory approach (see, e.g., [12]). In the framework of a mixed model ([3], [4], [5]) we employ the estimators of conditional entropy and mutual information studied in [2] and [3] to identify the relevant features.

Mixed model is described in papers [2], [5], [9]. We recall its definition. All the random elements (vectors, variables) are defined on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Consider a random vector X taking values in $\mathbb{R}^d$ endowed with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$. Let $Y : \Omega \to M$, M being a finite set, have $\mathsf{P}(Y = y) > 0$ for each $y \in M$. Components of a vector X are called explanatory variables (features), Y is called a response variable. Introduce σ-algebra $\mathcal{A} := 2^M$. We write P_Z for the distribution of a random element Z. Assume that the distribution of vector (X, Y) is absolutely continuous with respect to measure $\mu \otimes \lambda$, i.e. $\mathsf{P}_{(X,Y)} \ll \mu \otimes \lambda$, where μ is the Lebesgue measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and λ is a counting measure on $(M, \mathcal{A})$. Hence

Kozhevin, A.A., Feature selection based on statistical estimation of mutual information.

Received March, 31, 2021, published June, 23, 2021.

there exists the Radon - Nikodym derivative

$$\frac{d\mathsf{P}_{(X,Y)}}{d(\mu\otimes\lambda)} := f_{X,Y},$$

and one can write $\frac{d\mathsf{P}_X}{d\mu} = f_X$, where $f_X(x) = \sum_{y\in M} f_{X,Y}(x,y)$, $x\in\mathbb{R}^d$, $y\in M$.

Further the following notation will be employed. For a vector $Z = (Z_1,\dots,Z_d)$ with values in $\mathbb{R}^d$ variables $Z_1,\dots,Z_d$ are its components. For a set $L = \{l_1,\dots,l_m\}$, where integers $l_1,\dots,l_m$ are such that $1\le l_1 < \dots < l_m \le d$ and $m\in\{1,\dots,d\}$, set $Z_L := (Z_{l_1},\dots,Z_{l_m})$. In other words, we consider the sub-vector of a vector Z with components $Z_{l_1},\dots,Z_{l_m}$. The density $f_{X_L,Y}$ of a vector (X_L,Y) and the density f_{X_L} of a vector X_L are easily evaluated by means of densities $f_{X,Y}$ and f_X.

The mixed model arises in a number of important problems, including the analysis of medical and biological data. For instance X can describe genetic and non-genetic factors having potential impact on provoking certain disease whereas Y characterizes the health state of a patient ($Y=1$ means the disease occurrence and $Y=0$ corresponds to its absence).

In many situations a response variable does not depend on the whole collection of the explanatory variables but it depends on certain components of X. One can formalize this as follows.

Definition 1. *A set of indices $S = \{s_1,\dots s_m\}$ and a set of features $X_S := (X_{s_1},\dots,X_{s_m})$, where $1\le s_1 < \dots,< s_m \le d$, are called relevant if for each $y\in M$ and μ-almost all $x\in\mathbb{R}^d$ the following relation between conditional densities holds*

$$f_{Y|X}(y|x) = f_{Y|X_S}(y|x_S). \tag{1}$$

The conditional density $f_{Y|X}$ of Y given X is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)},\quad x\in\mathbb{R}^d,\ y\in M.$$

Note that various approaches to define relevance, redundancy and complementarity are discussed, e.g., in [12] (see also the references therein).

Diverse dependence measures between X and Y play an important role in the problem of identification of relevant factors. The method considered in the paper bases on the concept of mutual information between X and Y defined by way of

$$I(X;Y) := D_{KL}(\mathsf{P}_{(X,Y)}||\mathsf{P}_X\otimes\mathsf{P}_Y),$$

here D_{KL} is the Kullback-Leibler divergence between probability measures $\mathsf{P}_{X,Y}$ and $\mathsf{P}_X\otimes\mathsf{P}_Y$ (see [8]).

Evidently, for a mixed model

$$\frac{d\mathsf{P}_{(X,Y)}}{d(\mathsf{P}_X\otimes\mathsf{P}_Y)}(x,y) = \frac{f_{X,Y}(x,y)}{f_X(x)\mathsf{P}(Y=y)},\quad x\in\mathbb{R}^d,\ y\in M,$$

(with usual agreement $0/0 := 0$), thus one can write mutual information as

$$I(X;Y) = \sum_{y\in M}\int_{\mathbb{R}^d}\left(\log\frac{f_{X,Y}(x,y)}{f_X(x)\mathsf{P}(Y=y)}\right) f_{X,Y}(x,y)\mu(dx),$$

where log stands for $\log_e$ and $0\log 0 := 0$. Henceforth we write dx instead of $\mu(dx)$ for simplicity. Note that

$$I(X;Y) = H(Y) - H(Y|X) \tag{2}$$

where

$$H(Y) = -\sum_{y\in M} (\log \mathsf{P}(Y=y))\mathsf{P}(Y=y)$$

is the Shannon entropy of Y and

$$H(Y|X) = -\sum_{y\in M} \int_{\mathbb{R}^d} (\log f_{Y|X}(y|x)) f_{X,Y}(x,y)dx$$

is the conditional entropy of a random variable Y given a random vector X.

In the case of mixed model mutual information $I(X;Y)$ is always finite (see the reasoning after formula (13) in [3]).

2. Identification of relevant factors

Let independent vectors (X^i, Y^i), $i \in \mathbb{N}$, have the same distribution as the vector (X, Y). Consider a sample $\zeta_n = \{(X^i, Y^i)\}_{i=1}^n$. For description of the relevant factors identification we recall some definitions.

Let $|\cdot|$ stand for the cardinality of a finite set and $\|\cdot\|$ be the Euclidean norm in $\mathbb{R}^d$. For $i \in \{1,\dots,n\}$ and $k \in \{1,\dots,n-1\}$ the random vector $X^i_{(k)}$ is the k-th nearest neighbor for X^i among points $\{X^1,\dots,X^n\} \setminus \{X^i\}$ with respect to Euclidean distance (for each $\omega \in \Omega$ we find for a point $X^i(\omega)$ in $\mathbb{R}^d$ its k-th nearest neighbor among points $\{X^1(\omega),\dots,X^n(\omega)\} \setminus \{X^i(\omega)\}$). Then the random variable $\xi_{n,k,i}$ is introduced for $i \in \{1,\dots,n\}$ by formula

$$\xi_{n,k,i} = \xi_{n,k,i}(\zeta_n) := |\{j \in \{1,\dots,n\} \setminus \{i\} : Y^j = Y^i, \|X^i - X^j\| \le \|X^i - X^i_{(k)}\|\}|.$$

In other words, $\xi_{n,k,i}$ indicates a number of observations having the response Y^i among j such that the distance between X^j and X^i is not greater than $X^i_{(k)}$.

Following [2] and [3], for $n \in \mathbb{N}$, $n > 1$, and $k = k(n) \in \{1,\dots,n-1\}$ introduce the estimates

$$\widehat{H}_{n,k}(Y|X) := \frac{1}{n}\sum_{i=1}^n \widehat{H}_{n,k,i},$$

$$\widehat{H}_n(Y) := -\frac{1}{n}\sum_{y\in M} \widehat{\mathsf{P}}_n(y)\log(\widehat{\mathsf{P}}_n(y)).$$

Here

$$\widehat{H}_{n,k,i} := -\log(\xi_{n,k,i}+1) + \log k,$$

$$\widehat{\mathsf{P}}_n(y) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Y^i = y\}, \quad y \in M,\ n \in \mathbb{N}.$$

By analogy with (2) we come to the following estimate proposed in [3]:

$$\widehat{I}_{n,k}(X,Y) := \widehat{H}_n(Y) - \widehat{H}_{n,k}(Y|X). \tag{3}$$

In [3] instead of $\widehat{I}_{n,k}(X,Y)$ the authors used notation $\widehat{I}^{(1)}_{n,k}(X,Y)$ (see formula (38) in [3]). We simplified notation since we do not compare various estimates of mutual information.

Introduce

$$Q_m = \{L := (l_1,\dots,l_m) : 1 \le l_1 < \dots < l_m \le d\},$$

i.e. Q_m is a collection of subsets of a set $\{1,\dots,d\}$, containing exactly m elements. For any $L \in Q_m$ set $\zeta_{n,L} = \{(X_L^i, Y^i)\}_{i=1}^n$ and estimate the mutual information $I(X_L, Y)$ for each sample $\zeta_{n,L}$. For this purpose we employ an estimate of (3) type, using $\zeta_{n,L}$ instead of ζ_n. We write $\widehat{I}_{n,k,L} := \widehat{I}_{n,k}(X_L; Y)$, where $k = k(n)$ is a specified function, $k(n) \in \{1,\dots,n-1\}$.

Introduce a collection of random sets

$$\widehat{\mathbb{S}}_{n,k}(\omega) = \arg\max_{L\in Q_m} \widehat{I}_{n,k,L}(\omega).$$

Thus $\widehat{\mathbb{S}}_{n,k}$ is a collection of sets $\widehat{S}_{n,k}$ such that

$$\max_{L\in Q_m} \widehat{I}_{n,k,L} = \widehat{I}_{n,k,\widehat{S}_{n,k}}$$

for each $\widehat{S}_{n,k} \in \widehat{\mathbb{S}}_{n,k}$. Note that we can facilitate computations by using estimates of the conditional entropy to find $\widehat{\mathbb{S}}_{n,k}(\omega)$. Indeed,

$$\widehat{\mathbb{S}}_{n,k}(\omega) = \arg\max_{L\in Q_m} \left(\widehat{H}_n(Y) - \frac{1}{n}\sum_{i=1}^n \widehat{H}_{n,k,i}(Y|X_L^i) \right) = \arg\min_{L\in Q_m} \widehat{H}_{n,k}(Y|X_L),$$

since $\widehat{H}_n(Y)$ does not depend on $L \in Q_m$.

3. Main results

We need the following definitions. A function $g : \mathbb{R}^d \to \mathbb{R}$ is called C_0-constricted (see [2]) if there exists $R_0 > 0$ such that, for μ-almost all $x \in \mathbb{R}^d$ and any $R \in (0, R_0)$,

$$\left| g(x) - \frac{1}{\mu(B(x,R))} \int_{B(x,R)} g(u)du \right| \le C_0 R, \tag{4}$$

where $B(x,R)$ is a ball in $\mathbb{R}^d$ with center x and radius R.

According to Remark 2.2 [2], if a real-valued function g satisfies the Lipschitz condition in $\mathbb{R}^d$, i.e. $|g(x) - g(u)| \le C_0\|x-u\|$ for $x, u \in \mathbb{R}^d$, then (4) is true for any $R > 0$. In particular, the density $p_{a,\Sigma}$ of nondegenerate Gaussian distribution $N(a,\Sigma)$ in $\mathbb{R}^d$ satisfies the Lipschitz condition in $\mathbb{R}^d$ with the constant

$$C_0 = \max_{u\in R^d} \|\nabla p_{a,\Sigma}(u)\| < \infty.$$

For sequences of nonnegative numbers $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$ one writes $a_n \propto b_n$, whenever

$$c_1 b_n \le a_n \le c_2 b_n,\ n \in \mathbb{N},$$

where c_1 and c_2 are some positive constants ($c_1 < c_2$).

Theorem 1. *Assume that, for some $m \in \{1,\dots,n-1\}$, there exists a nonempty set $\mathbb{S}_m$ consisting of all collections S of relevant factors with cardinality $|S| = m$. Let a (version of) density $f_{X,Y}$ be strictly positive. Moreover, for any $L \in Q_m$ and $y \in M$, let a density $f_{X_L,Y}(\cdot, y)$ be C_0-constricted and $\mathsf{E}|\log f_{X_L}(X_L)|^{2+\varepsilon} < \infty$, for some $\varepsilon > 0$. Then, for each $\alpha \in (0,1)$ and $k = k(n) \propto n^\alpha$,*

$$\mathsf{P}(\widehat{\mathbb{S}}_{n,k} \subset \mathbb{S}_m) \to 1 \ \text{ as } \ n \to \infty.$$

In particular, if $\mathbb{S}_m$ consists of a single set S_m, then $\mathsf{P}(\widehat{\mathbb{S}}_{n,k} = S_m) \to 1$, $n \to \infty$.

Condition $\mathsf{E}|\log f_{X_L}(X_L)|^{2+\varepsilon}<\infty$ means that one considers $|\log f_{X_L}(u)|^{2+\varepsilon}$, substitutes X_L instead of u, and then takes an expectation. Thus in the present condition the lower index X_L is not connected with averaging.

Proof. First of all we show that if $S \in \mathbb{S}_m$ then

$$I(X_S;Y) = \max_{L\in Q_m} I(X_L;Y). \tag{5}$$

It is enough to verify that

$$H(Y|X_S) = \min_{L\in Q_m} H(Y|X_L).$$

For $x \in \mathbb{R}^d$ and $U = \{j_1,\dots,j_q\}$, where $1 \le j_1 < \dots < j_q \le d$, we write $x_U = (x_{j_1},\dots,x_{j_q})$. Consequently, for $L \in Q_m$ and $S \in \mathbb{S}_m$ one has

$$H(Y|X_L) - H(Y|X_S) = \sum_{y\in M} \int_{\mathbb{R}^d} f_{X,Y}(x,y) \log \frac{f_{Y|X_S}(y|x_S)}{f_{Y|X_L}(y|x_L)}\, dx.$$

The latter integral exists since a function

$$f_{X,Y}(x,y) \log \frac{f_{Y|X_S}(y|x_S)}{f_{Y|X_L}(y|x_L)}$$

is defined for each $x \in \mathbb{R}^d$ and $y \in M$. Indeed, $f_{Y|X_L}(y|x_L) > 0$ for all $x \in \mathbb{R}^d$, $y \in M$ and $L \in Q_m$ because

$$f_{X_L,Y}(x_L,y) = \int_{\mathbb{R}^m} f_{X,Y}(x,y)\, dx_{\overline{L}} = f_{Y|X_L}(y|x_L) f_{X_L}(x_L),$$

where $\overline{L} = \{1,\dots,d\} \setminus L$. For all $y \in M$ a function $f_{X,Y}(\cdot,y)$ is strictly positive. Hence, $f_{X_L,Y}(x_L,y) > 0$ for each x_L. Consequently,

$$f_{Y|X_L}(y|x_L) f_{X_L}(x_L) > 0$$

and

$$f_{X_L}(x_L) = \sum_{y\in M} f_{X_L,Y}(x_L,y) > 0.$$

Thus $f_{Y|X_L}(y|x_L) > 0$ for any $x \in \mathbb{R}^d$, $y \in M$ and $L \in Q_m$.

According to (1)

$$\begin{aligned} H(Y|X_L) - H(Y|X_S) &= \int_{\mathbb{R}^d} f_X(x) \sum_{y\in M} f_{Y|X}(y|x) \log \frac{f_{Y|X_S}(y|x_S)}{f_{Y|X_L}(y|x_L)}\, dx \\ &= \int_{\mathbb{R}^d} f_X(x) \sum_{y\in M} f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{f_{Y|X_L}(y|x_L)}\, dx \\ &= \int_{\mathbb{R}^d} f_X(x) D_{KL}(\mathsf{P}_{Y|X=x} || \mathsf{P}_{Y|X_L=x_L})\, dx, \end{aligned}$$

where the Kullback - Leibler divergence $D_{KL}(\mathsf{P}_{Y|X=x} || \mathsf{P}_{Y|X_L=x_L})$ is determined for probability measures on $(M, 2^M)$ having densities (with respect to counting measure) $f_{Y|X}(\cdot|x)$ and $f_{Y|X_L}(\cdot|x_L)$, respectively. The Kullback - Leibler divergence is nonnegative. Hence for any $S \in \mathbb{S}_m$ and $L \in Q_m$ one has

$$H(Y|X_S) - H(Y|X_L) \ge 0.$$

Thus (5) is established.

Consider a function $h : G \to \mathbb{R}$, where G is a finite set. Introduce

$$T := \arg\max_{t \in G} h(t).$$

Let $h_n : G \to \mathbb{R}$, $n \in \mathbb{N}$, be a sequence of random functions such that $h_n(t) \xrightarrow{\mathsf{P}} h(t)$ for each $t \in G$, as $n \to \infty$. Take a random set

$$T_n := \arg\max_{t \in G} h_n, \quad n \in \mathbb{N}.$$

In other words, for each $n \in \mathbb{N}$ and any $\omega \in \Omega$, the following equality holds

$$\max_{t \in G} h_n(t, \omega) = h_n(u, \omega)$$

for all $u \in T_n$. Then $\mathsf{P}(T_n \subset T) \to 1$, $n \to \infty$. Indeed, for any $\varepsilon > 0$ and all n large enough,

$$\mathsf{P}\left(\bigcap_{t \in G} |h(t) - h_n(t)| < \varepsilon \right) > 1 - \varepsilon.$$

Clearly, one can take $\varepsilon < \frac{1}{2} \max_{t \in G \setminus T} |h(t) - h(u)|$ where $u \in T$. Then, for all n large enough (with probability close to one) $\max_{t \in G \setminus T} h_n(t) < \max_{t \in T} h_n(t)$. Hence $T_n \subset T$. In other words, with probability close to one any element of T_n can be viewed as a point of maximum of a function h when n is large enough. In particular, if $|T| = 1$ (i.e. T is a singleton) then $\mathsf{P}(T_n = T) \to 1$ as $n \to \infty$.

Now we can apply this reasoning to our functions $h(L) := I(X_L; Y)$, $L \in G := Q_m$, and $h_n(L) := \widehat{I}_{n,k(n),L}$. The estimate $\widehat{I}_{n,k,L}$ is L_2-consistent by virtue of Theorem 4 [3] and, hence, a consistent estimate of $I(X_L; Y)$ for each $L \in Q_m$. The proof is complete. □

Consider the logistic regression model (see, e.g., [6], [10]). Recall that in the framework of this model Y and X take values in $M = \{0, 1\}$ and $\mathbb{R}^d$, respectively. Moreover, the following relations hold:

$$\mathsf{P}(Y = 1 | X = x) = \frac{1}{1 + \exp\{-(w, x) - b\}}, \quad x \in \mathbb{R}^d, \quad w \in \mathbb{R}^d, \quad b \in \mathbb{R}, \tag{6}$$

$$\mathsf{P}(Y = 0 | X = x) = 1 - \mathsf{P}(Y = 1 | X = x), \tag{7}$$

where $(\cdot, \cdot) := (\cdot, \cdot)_d$ is the scalar product in $\mathbb{R}^d$. Further we write (u, u) instead of $(u, u)_k$ whenever it is clear that $u \in \mathbb{R}^k$. The statement given below turns useful.

Theorem 2. *Let the conditional distribution of Y given $X = (X_1, \ldots, X_d)$ be described by formulas (6) and (7). Assume that $X_1, \ldots, X_d$ are independent random variables. Then the set*

$$S := \{i : w_i \neq 0, 1 \leq i \leq d\}$$

is the minimal set of relevant factors, i.e. S is relevant and if L is another relevant set its cardinality $|L| \geq |S|$.

Proof. A set $L \subset \{1, \ldots, d\}$ is relevant if and only if, for each $x_L \in \mathbb{R}^k$,

$$f_{Y|X_L}(1|x_L) = f_{Y|X}(1|x). \tag{8}$$

The left side of this equality can be rewritten in the following way

$$f_{Y|X_L}(1|x_L) = \frac{f_{X_L,Y}(x_L, 1)}{f_{X_L}(x_L)}$$

$$= \frac{\int_{\mathbb{R}^{d-|L|}} (1+\exp\{-(w_S,x_S)-b\})^{-1} f_{X_L}(x_L) f_{X_{\overline{L}}}(x_{\overline{L}})\, dx_{\overline{L}}}{f_{X_L}(x_L)}$$

$$= \int_{\mathbb{R}^{d-|L|}} \frac{f_{X_{\overline{L}}}(x_{\overline{L}})}{1+\exp\{-(w_S,x_S)-b\}}\, dx_{\overline{L}}. \tag{9}$$

Note that $(w_S,x_S) = (w_{S\cap L},x_{S\cap L}) + (w_{S\cap\overline{L}},x_{S\cap\overline{L}})$, so the latter expression in (9) depends only on $x_{S\cap L}$. Moreover, it is easily seen that if $S\cap L \neq \varnothing$ then this function in $x_{S\cap L}$ is not constant on $\mathbb{R}^{|S\cap L|}$. In fact we can compare the values of function under the sign of integral at $x_{S\cap L}$ and $z_{S\cap L}$ $(x,z\in\mathbb{R}^d)$. Next observe that if a function g is integrable on a measurable set $A\subset\mathbb{R}^m$ w.r.t. the Lebesgue measure μ and if also $g(u)>0$ for μ-almost all $u\in A$ and $\mu(A)>0$, then $\int_A g(u)\mu(du)>0$. Clearly, whenever $S\setminus L\neq\varnothing$ the function

$$f_{Y|X}(1|x) = \frac{1}{1+\exp\{-(w,x)-b\}} = \frac{1}{1+\exp\{-(w_S,x_S)-b\}}$$

can not take the same constant value for all $x_{S\setminus L}$. Consequently, $S\subset L$ provided that (8) holds. Evidently, if $L=S$ then by virtue of (9)

$$f_{Y|X_S}(1|x_S) = \frac{1}{1+\exp\{-(w_S,x_S)-b\}} \int_{\mathbb{R}^{d-|S|}} f_{X_{\overline{S}}} dx_{\overline{S}} = \frac{1}{1+\exp\{-(w_S,x_S)-b\}}.$$

Hence (1) is valid. The proof is complete. □

4. Simulations

In the framework of model (6) – (7) we assume that $X\sim N(0,I_d)$ where I_d is a unit matrix of size d. Then in view of [2] the conditions of Theorem 1 are satisfied. For statistical estimation of $\mathsf{P}(\widehat{\mathbb{S}}_{n,k}\subset\mathbb{S}_m)$ we perform the following procedure.

(1) Fix integers n, m, d and generate N $(N\in\mathbb{N})$ independent random samples $\zeta^{(N)}_{j,n,d} = \{(X^{i,j},Y^{i,j})\}_{i=1}^n$, $j=1,\dots,N$, consisting of i.i.d. random pairs of observations with the same law as (X,Y). Set $b=0$ and

$$w=(w_1,\dots,w_d),\quad w_i = \begin{cases} 1, & i\le m,\\ 0, & i>m.\end{cases}$$

In other words, we consider the case when w has the first m nonzero components and others are equal to zero. According to Theorem 2 we can state that $S_m=\{1,\dots,m\}$ is the unique relevant set of features having cardinality m.

(2) For each sample $\zeta^{(N)}_{j,n,d}$ and each subset of indices $L\in Q_m$ we compute $\widehat{\mathbb{S}}_{n,k}(\zeta^{(N)}_{j,n,d})$, i.e. estimate the set of relevant factors for each generated sample.

(3) Compute

$$ACC^{(N)}_{n,d,k} = \widehat{\mathsf{P}}^{(N)}_n(\widehat{\mathbb{S}}_{n,k}\subset\mathbb{S}_m) = \frac{1}{N}\sum_{j=1}^N \mathbb{I}(\widehat{\mathbb{S}}_{n,k}(\zeta^{(N)}_{j,n,d}) = S_m).$$

Due to the SLLN one can claim that

$$\widehat{\mathsf{P}}^{(N)}_n(\widehat{\mathbb{S}}_{n,k}=S_m)\to\mathsf{P}(\widehat{\mathbb{S}}_{n,k}=S_m)\ \text{ a.s.},\quad N\to\infty.$$

For each configuration of parameters we take $N = 100$. Figures 1-3 show $ACC^{(N)}_{n,d,k}$ as a function of n for different combinations of $d \in \{10, 50\}$, $m \in \{1, 2, 3\}$ and $k \in \{5, 10, 20, 30\}$. Experiments demonstrate that in all considered cases $ACC^{(N)}_{n,d,k}$ can be equal to 1 for finite values of n. Thus we have demonstrated that for rather small values of k, and for n moderately large the proposed procedure permits to identify the relevant features accurately.

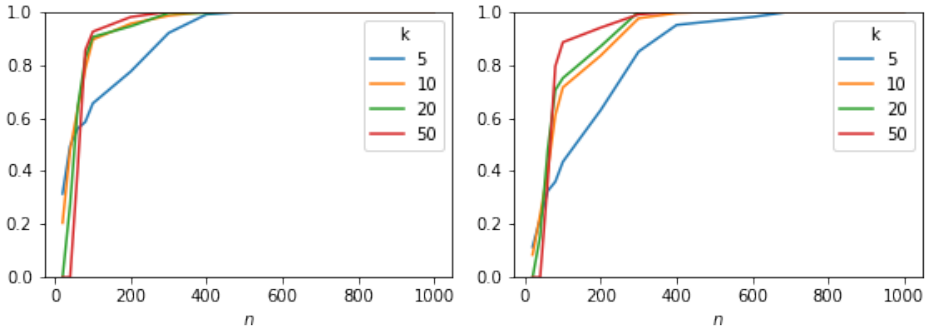

FIGURE 1. $m = 1$. Left: $d = 10$. Right: $d = 50$

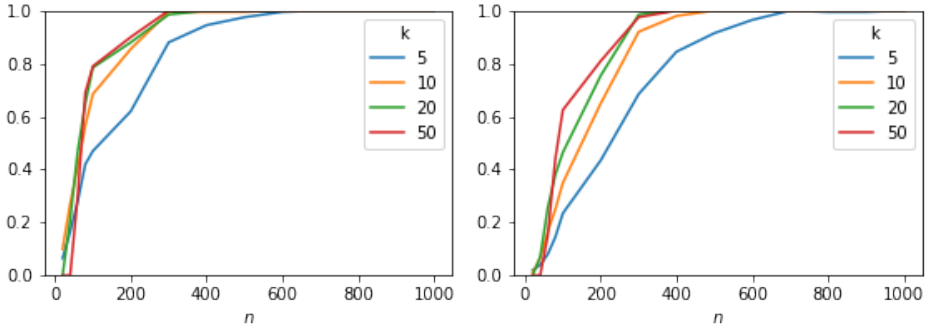

FIGURE 2. $m = 2$. Left: $d = 10$. Right: $d = 50$

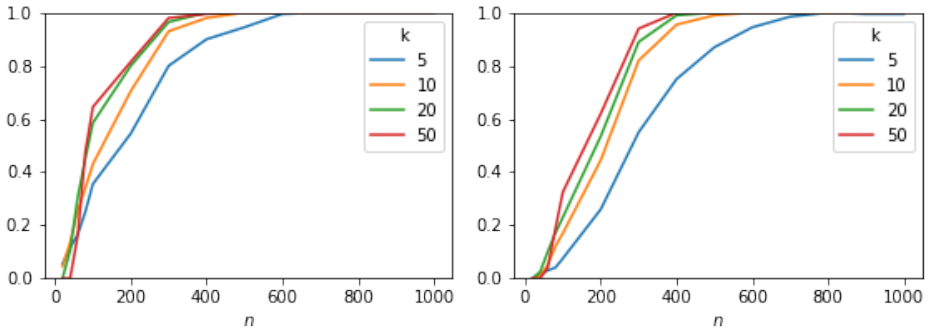

FIGURE 3. $m = 3$. Left: $d = 10$. Right: $d = 50$

Acknowledgements. The author is grateful to Professor A.V.Bulinski for useful discussions and valuable remarks and to Reviewer for detailed recommendations aimed to improve the exposition of the established results.

REFERENCES

[1] V. Bolón-Candedo, A. Alonso-Betanzos, *Recent advances in ensembles for feature selection*, Springer, 2018.
[2] A. Bulinski, A. Kozhevin, *Statistical estimation of conditional Shannon entropy*, ESAIM, Probab. Stat., **23** (2019), 350–386. Zbl 1418.60026
[3] A. Bulinski, A. Kozhevin, *Statistical Estimation of Mutual Information for Mixed Model*, Methodol. Comput. Appl. Probab., **23**:2 (2021) 123–142.
[4] F. Coelho, A.P. Braga, M.A. Verleysen, *Mutual Information estimator for continuous and discrete variables applied to Feature Selection and Classification problems*, International Journal of Computational Intelligence Systems, **9**:4 (2016), 726–733.
[5] W. Gao, S. Kannan, S. Oh, P. Viswanath, *Estimating mutual information for discrete-continuous mixtures*, 31-st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA (2017), 1–12.
[6] D.G. Kleinbaum, M. Klein, *Logistic regression. A self-learning text. With contributions by Erica Rihl Pryor. 3rd ed.*, Springer, New York, 2010. Zbl 1194.62090
[7] M. Kuhn, K. Johnson, *Feature engineering and selection: A practical approach for predictive models*, CRC Press, Boca Raton, 2020.
[8] S. Kullback, R.A. Leibler, *On information and sufficiency*, Ann. Math. Stat., **22**:1 (1951), 79–86. Zbl 0042.38403
[9] F. Macedo, R. Oliveira, A. Pacheco, R. Valadas *Theoretical foundations of forward feature selection methods based on mutual information*, Neurocomputing, **325** (2019), 67–89.
[10] L. Massaron, A. Boschetti, *Regression analysis with Python*, Packt Publishing Ltd., Birmingham, 2016.
[11] U. Stańczyk, B. Zeilosko, L.C. Jain, (Eds.) *Advances in Feature Selection for Data and Pattern Recognition*, Springer, 2018.
[12] J.R. Vergara, P.A. Estévez, *A review of feature selection methods based on mutual information*, Neural Comput. and Applic. **24** (2014), 175–186.

ALEXEY ALEXANDROVICH KOZHEVIN
LOMONOSOV MOSCOW STATE UNIVERSITY,
GSP-1, LENINSKIE GORY, MOSCOW,
MOSCOW, 119991, RUSSIA
Email address: kozhevin.alexey@gmail.com