MATHEMATICAL MODELING AND COMPUTING, Vol. 10, No. 3, pp. 703–711 (2023)

Feature screening algorithm for high dimensional data

Chamlal H., Benzmane A., Ouaderhman T.

Faculty of Sciences Ain Chock, Hassan II University,
km. 8 Route of EL Jadida, B.P. 5366 Maarif 20100, Casablanca 20000, Morocco

(Received 15 February 2023; Revised 3 July 2023; Accepted 5 July 2023)

Currently, feature screening is becoming an important topic in the fields of machine learning and high-dimensional data analysis. Filtering out irrelevant features from a set of variables is considered to be an important preliminary step that should be performed before any data analysis. Many approaches have been proposed to the same topic after the work of Fan and Lv (J. Royal Stat. Soc., Ser. B. **70** (5), 849–911 (2008)), who introduced the sure screening property. However, the performance of these methods differs from one paper to another. In this work, we aim to add to this list a new algorithm performing feature screening inspired by the Kendall interaction filter (J. Appl. Stat. **50** (7), 1496–1514 (2020)) when the response variable is continuous. The good behavior of our algorithm is proved through a comparison with an existing method, proposed in this work under several simulation scenarios.

Keywords: *feature screening; discretization; high dimensional data; regression.*
2010 MSC: 62G08, 62F07, 62G08 **DOI:** 10.23939/mmc2023.03.703

1. Introduction

Variable selection has attracted researcher attention in different domains and has demonstrated its importance and ability to improve the prediction performance of so many statistical learning models. Over time, a large scientific community was interested in the subject of feature screening whose purpose is less challenging than that of feature selection [1]. Feature screening remains a necessity. Especially when the features space is ultra-high dimensional, that is, when the number of features exceeds exponentially the sample size. In this case, the majority of classical prediction methods remain inefficient and this is what motivated the creation of methods filtering main effects and/or interaction effects. Among others, there are [2–5].

However, interaction screening is a topic in full bloom. Few researchers have dealt with it in the literature, and thus only a few methods exist enabling to filter interactive pairs of features that import, together, useful information. Papers that have tackled this subject count: [3, 4, 6–9] and few others.

Despite the low number of results in this area, interaction screening brings important insights into the domain of genetics because a lot of genes interact together and cause diseases which means that identifying these interactions enables researchers and scientists in this domain to understand and take action to combat these diseases [10]. A limitation of most of the methods tackling feature screening is the heredity assumption [11]. That is assuming in advance strong or weak heredity. However, sometimes predictors can act only in pairs without having the main effect on the response [7]. Thus, ignoring their interactive effect on the outcome can distort the results.

Through our procedure, we aim to add a new efficient way of selecting interactive pairs among all candidate features of ultra-high dimensional regression models, and this, without assuming necessarily beforehand neither a strong nor weak heredity. Throughout this paper, we consider the model:

$$y = a_0 + \sum_{j=1}^{p} a_j X_j + \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} b_{kl} \cdot X_k \cdot X_l + \varepsilon. \tag{1}$$

Such that, similar to [4], a_0 is the intercept, a_j is the regression coefficient for marginal effects, b_{kl} is the regression coefficient for interaction effects, and ε is a random error variable independent of the predictors $X_{j\in\{1,\ldots,p\}}$.

In this paper, we present a new algorithm that performs slicing and fusion in 2 steps to enable the application of the Kendall interaction filter proposed by Y. Anzarmo et al. [6] to regression problems with a continuous response variable and extends its benefits. The first step of our proposed method concerns a slicing approach that aims to turn the continuous response variable into a categorical target variable by grouping its values in labeled intervals. Then, similarly to [1], the second step is fusion. Its approach consists of considering different partitions with a precise number of slices. Then, we compute for each partition the KIF measure [6]. And at the end, we sum all the resulting measures. This final measure is our new filter. This last trick allows us to avoid the results being dependent on the scheme used for slicing [1] and thus increase the chances of obtaining high performance and good interaction screening. The procedure used in our approach for the construction of partitions is quantile.

This paper is organized as follows: in Section 2, we present the methodology of our approach in which we give an overview of the Kendall Interaction Filter method, we detail our proposal, and then present our simulation scenarios. Section 3 is devoted to simulation studies results and comparison of our algorithm to another interaction screening method to evaluate its performance. And Section 4 is the conclusion of our work.

2. Method

2.1. Motivation

The Kendal Interaction Filter is a method of interaction screening proposed by Y Anzarmo et al. [6] that uses Kendall's τ to rate the interaction between two features for multi-class classification problems. The KIF measure is defined by:

$$\omega_{j,l} = \sum_{k=1}^{K} \pi_k \cdot \big|\tau_k(X_j, X_l) - \tau(X_j, X_l)\big|.$$

Such that: K is the number of classes, π_k is the prior probability, τ is Kendall's tau, and τ_k is the conditional Kendall's tau [6].

This method works well with categorical response variables but can not be applied directly to regression problems with a continuous response variable. The good behavior of the Kendall Interaction Filter in multi-class classification applications motivated us to test its attitude in the case of high and ultra-high dimensional models with continuous response. The first idea is to apply the slicing procedure used in the literature for transforming a continuous variable into a categorical variable. Then, to increase the performance of the interaction screening, we use a second method which consists in summing all the KIF measures found after the slicing step. The role of fusion [6,12], is demonstrated in practice. Therefore, to create our new algorithm, we apply the two steps that use the KIF approach for interaction screening, and we call it the fused interaction filter.

2.2. The fused filter

The slicing procedure is widely used in literature [1]. It aims to transform a continuous variable into a categorical variable. Using the slicing approach, some of the proposed methods for feature screening can be applied to both classification and regression problems. However, the slicing trick is not a sufficient way to extend a method of classification to a regression setting and expect perfect results. That is why, our proposal, as indicated previously in the introduction, is a two-step method based on slicing and fusion [12]. First, we partition the observed data of the response y into slices according to N different partitions using the quantile method, we calculate the KIF measure for each partition $P_{i\in\{1,\ldots,N\}}$, then we sum all the N measures founded in the previous step (see Figure 1). The resulted measure is the new fused interaction filter.

We define the P_i partitions by

$$\begin{cases} P_1 = \{[a_s, a_{s+1}[, s \in \{0, \ldots, g_1 - 1\}\}, \\ P_2 = \{[a_s, a_{s+1}[, s \in \{0, \ldots, g_2 - 1\}\}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ P_N = \{[a_s, a_{s+1}[, s \in \{0, \ldots, g_N - 1\}\}. \end{cases}$$

With: g_i is the number of intervals $[a_s, a_{s+1}[$ in the ith partition and $a_s < a_{s+1}$.

For each fixed partition, we have:

$$y \in [a_s, a_{s+1}[\Longleftrightarrow y^{new} = s + 1. \tag{E}$$

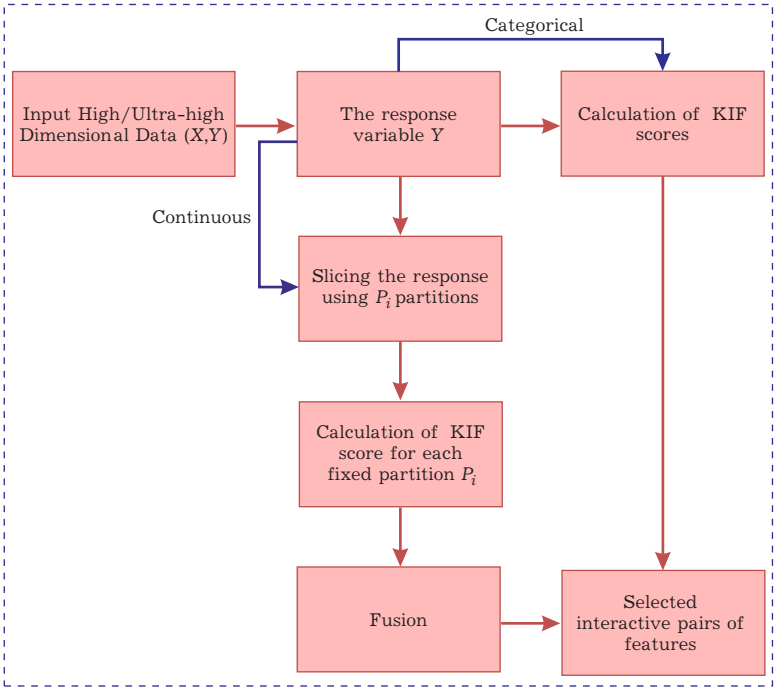

Fig. 1. The interaction screening procedure using the Kendall Interaction Filter.

Similarly to [1], throughout this paper, we use the following fixed values: $g_1 = 3$, $g_n = \lceil \log(n) \rceil$ and $N = \lceil \log(n) \rceil - 2$.

For each pair (j, k), the fused interaction filter measure is defined by

$$f_{j,k} = \sum_{i=1}^{N} \omega_{j,k,P_i}. \tag{2}$$

Such that $j, k \in \{1, \ldots, p\} \times \{1, \ldots, p\}$ and $j < k$. Where $\omega_{j,k,P_i} = \sum_{s=1}^{g_i} \pi_s \cdot |\tau_s(X_j, X_k) - \tau(X_j, X_k)|$ which is the Kendall Interaction Filter [6] for a fixed partition P_i. And $\pi_s = \mathbb{P}(Y \in [a_{s-1}, a_s[) = \mathbb{P}(Y^{new} = s)$, $\tau_s(X_j, X_k) = 2\mathbb{P}((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0 | Y^{new} = s, \widetilde{Y}^{new} = s) - 1$, $\tau(X_j, X_k) = 2\mathbb{P}((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0) - 1$. With $(\widetilde{X}, \widetilde{Y})$ a copy of (X, Y) such that they are independent of each other.

The interaction screening filter set, which is the indices set of interactive pairs, is similar to [6] given by:

$$I = \big\{(j, k); F(Y|(X_1, \ldots, X_p)) \text{ functionally depends on } (X_j, X_k) \text{ rank association}\big\}. \tag{3}$$

The empirical version of $f_{j,k}$ measure is defined by

$$\widetilde{f}_{j,k} = \sum_{i=1}^{N} \widetilde{\omega}_{j,k,P_i}. \tag{4}$$

Such that $j, k \in \{1, \ldots, p\} \times \{1, \ldots, p\}$ and $j < k$. And $\widetilde{\omega}_{j,k,P_i} = \sum_{s=1}^{g_i} \widetilde{\pi}_s \cdot |\widetilde{\tau}_s(X_j, X_k) - \widetilde{\tau}(X_j, X_k)|$, where: $\widetilde{\pi}_s = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i^{new} = s\}$, $\widetilde{\tau}(X_j, X_k) = \frac{4}{n(n-1)} \sum_{i<t=1}^{n} \mathbb{I}\{(X_{ij} - X_{tj})(X_{ik} - X_{tk}) > 0\} - 1$, and

$\widetilde{\tau}_s(X_j, X_k) = \frac{4}{n_s \cdot (n_s - 1)} \sum_{i<t=1}^{n} \mathbb{I}\{(X_{ij} - X_{tj})(X_{ik} - X_{tk}) > 0, Y_i^{new} = s, Y_t^{new} = s\} - 1$, n_s corresponds to the number of observations in the slice s [6].

The empirical version of I is given by:

$$\widehat{I} = \big\{(j,k); j < k \text{ and } \widetilde{f}_{j,k} > c \cdot n^{-r}, \text{ such that } c, r \text{ are positive constants}\big\}. \tag{5}$$

In practice, we consider the following estimated set:

$$\widehat{I} = \big\{(j,k); j < k \text{ and } \widetilde{f}_{j,k} \text{ is among} \lceil \tfrac{n}{\log(n)} \rceil \text{ largest of all}\big\}. \tag{6}$$

2.3. Theoretical results

This section concerns the main theoretical results of our proposal. We prove that the sure screening property still holds after the extension of the Kendall interaction filter to the case of a continuous target variable. Furthermore, the conditions presented here are the same as in the work [6], with only some differences, due to the slicing-fusion transformation of the algorithm.

To this end, it is important to recall that: $\widetilde{\pi}_s$ is a consistent estimator of π_s and $\widetilde{\pi}$ is a consistent estimator of π [6].

Before establishing the theorem concerning our main theoretical result, we adapt the conditions assumed for the Kendall Interaction Filter [6] to our Fused Interaction Filter extension of the first:

Condition 1: $\exists c_1, c_2 \geqslant 0$ such that: $\frac{c_1}{g_i} \leqslant \min\limits_{1 \leqslant s \leqslant g_i} \pi_s \leqslant \max\limits_{1 \leqslant s \leqslant g_i} \pi_s \leqslant \frac{c_2}{g_i}$, for all $i \in \{1, \ldots, N\}$.

Condition 2: $\exists (c, r) \in]0, +\infty[\times[0, \frac{1}{2}[$, such that: $\min\limits_{j,k \in I} f_{j,k} > 2 \cdot c \cdot n^{-r}$.

Condition 3: Assume that $\log(p) = O(n^{\eta})$ for some $\eta > 0$.

The following lemma is a necessary preliminary step for understanding and proofing the theorem.

2.3.1. Lemma and Theorem

Lemma 1. *Under the conditions above and for all $\varepsilon > 0$, we have:*

$$\mathbb{P}\Big(\max_{1 \leqslant j,k \leqslant p} |\widetilde{f}_{j,k} - f_{j,k}| > \varepsilon\Big) \leqslant 2p^2 \cdot \left(\lceil \log(n) \rceil \cdot \exp\left(\frac{-n \cdot \varepsilon^2}{18 \cdot \lceil \log(n) \rceil^4}\right) + \exp\left(\frac{-n \cdot \varepsilon^2}{72 \cdot \lceil \log(n) \rceil^2}\right) + \lceil \log(n) \rceil \cdot \exp\left(\frac{-n \cdot \varepsilon^2}{72 \cdot \lceil \log(n) \rceil^4}\right) \right).$$

Proof. Similar to [6], we have:

$$\begin{aligned} \mathbb{P}\left(\max_{1 \leqslant j,k \leqslant p} |\widetilde{f}_{j,k} - f_{j,k}| > \varepsilon\right) &\leqslant \mathbb{P}\left(\sum_{i=1}^{N} \sum_{s=1}^{g_i} \tilde{\pi}_s \cdot |\tilde{\tau}_s(X_j, X_k) - \tau_s(X_j, X_k)| > \frac{\varepsilon}{3}\right) \\ &\quad + \mathbb{P}\left(\sum_{i=1}^{N} g_i \tilde{\pi}_s \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3}\right) \\ &\quad + \mathbb{P}\left(2 \cdot \sum_{i=1}^{N} \sum_{s=1}^{g_i} |\tilde{\pi}_s - \pi_s| > \frac{\varepsilon}{3}\right). \end{aligned}$$

However,

$$\begin{aligned} \mathbb{P}\left(2 \cdot \sum_{i=1}^{N} \sum_{s=1}^{g_i} |\tilde{\pi}_s - \pi_s| > \frac{\varepsilon}{3}\right) &\leqslant \mathbb{P}\left(2 \cdot \sum_{i=1}^{N} g_i \cdot \max_{s \in \{1, \ldots, g_i\}} |\tilde{\pi}_s - \pi_s| > \frac{\varepsilon}{3}\right) \\ &\leqslant \mathbb{P}\left(2 \cdot N \cdot \lceil \log(n) \rceil \cdot \max_{s \in \{1, \ldots, \lceil \log(n) \rceil\}} |\tilde{\pi}_s - \pi_s| > \frac{\varepsilon}{3}\right) \\ &\leqslant \mathbb{P}\left(2 \cdot (\lceil \log(n) \rceil - 2) \cdot \lceil \log(n) \rceil \cdot \max_{s \in \{1, \ldots, \lceil \log(n) \rceil\}} |\tilde{\pi}_s - \pi_s| > \frac{\varepsilon}{3}\right) \\ &\leqslant \mathbb{P}\left(2 \cdot \lceil \log(n) \rceil \cdot \lceil \log(n) \rceil \cdot \max_{s \in \{1, \ldots, \lceil \log(n) \rceil\}} |\tilde{\pi}_s - \pi_s| > \frac{\varepsilon}{3}\right) \end{aligned}$$

$$\leqslant \sum_{s=1}^{\lceil \log(n) \rceil} \mathbb{P}\left(|\tilde{\pi}_s - \pi_s| > \frac{\varepsilon}{6 \cdot \lceil \log(n) \rceil^2} \right).$$

According to Hoeffding inequality, we have:

$$\sum_{s=1}^{\lceil \log(n) \rceil} \mathbb{P}\left(|\tilde{\pi}_s - \pi_s| > \frac{\varepsilon}{6 \cdot \lceil \log(n) \rceil^2} \right) \leqslant 2 \cdot \lceil \log(n) \rceil \cdot \exp\left(\frac{-n \cdot \varepsilon^2}{18 \cdot \lceil \log(n) \rceil^4} \right).$$

And

$$\mathbb{P}\left(\sum_{i=1}^{N} g_i \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3} \right) \leqslant \mathbb{P}\left((\lceil \log(n) \rceil - 2) \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3} \right)$$
$$\leqslant \mathbb{P}\left(|\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3 \cdot \lceil \log(n) \rceil} \right).$$

Using the fact that $\tilde{\tau}$ is a consistent estimator of τ [6], we have:

$$\mathbb{P}\left(|\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3 \cdot \lceil \log(n) \rceil} \right) \leqslant 2 \cdot \exp\left(\frac{-n \cdot \varepsilon^2}{72 \cdot \lceil \log(n) \rceil^2} \right).$$

So,

$$\mathbb{P}\left(\sum_{i=1}^{N} g_i \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3} \right) \leqslant 2 \cdot \exp\left(\frac{-n \cdot \varepsilon^2}{72 \cdot \lceil \log(n) \rceil^2} \right).$$

And

$$\mathbb{P}\left(\sum_{i=1}^{N} \sum_{s=1}^{g_i} \tilde{\pi}_s \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3} \right) \leqslant \mathbb{P}\left(\sum_{i=1}^{N} g_i \cdot \max_{s \in \{1, \ldots, g_i\}} \tilde{\pi}_s \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3} \right)$$
$$\leqslant \mathbb{P}\left(N \cdot \lceil \log(n) \rceil \cdot \max_{s \in \{1, \ldots, \lceil \log(n) \rceil\}} \tilde{\pi}_s \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3} \right)$$
$$\leqslant \sum_{s=1}^{\lceil \log(n) \rceil} \mathbb{P}\left(\lceil \log(n) \rceil^2 \cdot \tilde{\pi}_s \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3} \right)$$
$$\leqslant \sum_{s=1}^{\lceil \log(n) \rceil} \mathbb{E}_{Y^{new}}\left(\mathbb{P}\left(\tilde{\pi}_s \cdot |\tilde{\tau}(X_j, X_k) - \tau(X_j, X_k)| > \frac{\varepsilon}{3 \cdot \lceil \log(n) \rceil^2} | Y^{new} \right) \right)$$
$$\leqslant 2 \cdot \lceil \log(n) \rceil \cdot \exp\left(\frac{-n \cdot \varepsilon^2}{72 \cdot \lceil \log(n) \rceil^4} \right).$$

Therefore we have

$$\mathbb{P}\left(\max_{1 \leqslant j,k \leqslant p} |\widetilde{f}_{j,k} - f_{j,k}| > \varepsilon \right) \leqslant 2p^2 \cdot \left(\lceil \log(n) \rceil \cdot \exp\left(\frac{-n \cdot \varepsilon^2}{18 \cdot \lceil \log(n) \rceil^4} \right) \right.$$
$$\left. + \exp\left(\frac{-n \cdot \varepsilon^2}{72 \cdot \lceil \log(n) \rceil^2} \right) + \lceil \log(n) \rceil \cdot \exp\left(\frac{-n \cdot \varepsilon^2}{72 \cdot \lceil \log(n) \rceil^4} \right) \right). \quad \blacksquare$$

Theorem 1 (The sure screening property). *Under Conditions 1–4,* $\exists (c, r) \in]0, +\infty[\times [0, \frac{1}{2}[$, *such that:*

$$\mathbb{P}(I \subseteq \widehat{I}) \to 1 \text{ as } n \to +\infty.$$

Proof. Using Lemma, we took $\varepsilon = c \cdot n^{-r}$. The rest follows as in [6]: If we suppose that $I \nsubseteq \widetilde{I}$, then $\exists (j, k) \in I$ such that $\widetilde{f}_{j,k} \leqslant c \cdot n^{-r}$ and according to condition 2, we have $\min\limits_{j,k \in I} \widetilde{f}_{j,k} > 2 \cdot c \cdot n^{-r}$.

Therefore,

$$\mathbb{P}(I \subseteq \widetilde{I}) \geqslant 1 - 2p^2 \cdot \left(\lceil \log(n) \rceil \cdot \exp\left(\frac{-c^2 \cdot n^{-2r+1}}{18 \cdot \lceil \log(n) \rceil^4} \right) + \exp\left(\frac{-c^2 \cdot n^{-2r+1}}{72 \cdot \lceil \log(n) \rceil^2} \right) \right.$$

$$+\lceil \log(n) \rceil \cdot \exp\left(\frac{-c^2 \cdot n^{-2r+1}}{72 \cdot \lceil \log(n) \rceil^4}\right)\Bigg). \qquad \blacksquare$$

2.3.2. The fused interaction filter algorithm

Our proposed two-step algorithm is applied to regression models in the form of (1) and presented in detail below.

Algorithm 1 The fused interaction filter algorithm.

Require: $(X^i, Y^i)_{i=1}^n$;

initialization: $(\widetilde{\omega}_{j,k,P_i})_{i\in\{1,\dots,N\}}$ as an empty vector of size $\lceil \log(n) \rceil - 2$, $\widetilde{f}$ as an empty $p \times p$ matrix, $I = \varnothing$ and $r = 1$;
for $j = 1, \dots, p-1$
  for $k = j+1, \dots, p$
    for $g_i = 3, \dots, \lceil \log(n) \rceil$
      Construction of a number g_i of intervals using the quantile method,
      Compute Y^{new} as in (E),
      Compute $\widetilde{\omega}_{j,k,P_i} \leftarrow KIF(X_j, X_k, Y^{new})$ (the KIF measure for a fixed partition P_i);
    $\widetilde{f}_{j,k} \leftarrow \text{sum}((\widetilde{\omega}_{j,k,P_i})_{i\in\{1,\dots,N\}})$
for all $(j,k) \in \{1,\dots,p\} \times \{1,\dots,p\}$ such that $j < k$
  while $r \leqslant \lceil \frac{n}{\log(n)} \rceil$
    if $(j,k) = \arg\max\limits_{j,k} \widetilde{f}_{j,k}$ **then**
      Update $\widetilde{I}$: $\widetilde{I} \leftarrow \widetilde{I} \cup \{(j,k)\}$,
      Set $\widetilde{f}_{j,k} \leftarrow 0$;
    $r \leftarrow r + 1$
Output: the selected indices set $\widetilde{I}$.

We performed our algorithm on R.

2.4. Simulation studies

Throughout this section, we compare the results of our proposed algorithm to those of JCIS [7], which is a method of interaction screening. JCIS stands for Joint Cumulant Interaction Screening, accepts continuous response variables, and has demonstrated its success in selecting interactive pairs of features [7] for ultra-high dimensional data. We repeat each scenario of the two first simulations 100 times, the third simulation scenario 47 times, the fourth 50 times and the last one is repeated 100 times. Concerning the analysis of the results, we compute for simulations 1, 3, 4 and 5 the selection frequency of the supposedly interactive pairs, and we plot a violin graph, for the results of simulation 2, to compare the FIF scores of the interactive couples with the maximum FIF scores of the other pairs of features.

2.4.1. Simulation 1

In this scenario, we consider the regression model:

$$Y = X_1 \cdot X_2 + \varepsilon_1.$$

Where X_1, X_2 are, respectively, the first and second columns of the matrix X composed of $n = 100$ rows and $p = 500$ columns. The n corresponds to the sample size and p to the number of predictors. It is easy to remark that $p = 5 \times n$ which means that p is much larger than n.

X is generated such that: $X_i \sim N(0_p, \Sigma)$ for all $i \in \{1,\dots,n\}$, where $\Sigma = (\Sigma_{jk}) = (0.2^{|j-k|})$ and $j,k \in \{1,\dots,n\} \times \{1,\dots,n\}$.

$\varepsilon_1 = (\varepsilon_{1_1}, \dots, \varepsilon_{1_n})$ is independent of X and each $\varepsilon_{1_{i\in\{1,\dots,n\}}} \sim N(0, 0.1)$.

From the above model, we can easily say that the only interaction that influences the target variable is the interaction between the features X_1 and X_2.

2.4.2. Simulation 2

In this second example of simulation, we consider the following regression model:

$$Y = X_1 \cdot X_2 + X_3 \cdot X_4 + \varepsilon_2.$$

Where X is a matrix with $n = 100$ rows and $p = 600$ columns, such that each row is generated as follows: $X_i \sim N(0_p, \Sigma)$ for all $i \in \{1, \ldots, n\}$, where $\Sigma = (\Sigma_{jk}) = (0.2^{|j-k|})$ and $(j, k) \in \{1, \ldots, p\} \times \{1, \ldots, p\}$.

$\varepsilon_2 = (\varepsilon_{2_1}, \ldots, \varepsilon_{2_n})$ is independent of X and each $\varepsilon_{2_{i\in\{1,\ldots,n\}}} \sim N(0, 0.04)$.

Here, the interactive couples are (X_1, X_2) and (X_3, X_4).

2.4.3. Simulation 3

In this example, we consider the following model:

$$Y = X_1 + 4X_6 + 3X_1 \cdot X_6 + \varepsilon_3.$$

Where X is a matrix of $n = 400$ rows and $p = 1000$ columns. Each row is generated as follows: $X_i \sim N(0_p, \Sigma)$ for all $i \in \{1, \ldots, n\}$, where $\Sigma = (\Sigma_{jk}) = (0.1^{|j-k|})$ and $(j, k) \in \{1, \ldots, p\} \times \{1, \ldots, p\}$.

$\varepsilon_3 = (\varepsilon_{3_1}, \ldots, \varepsilon_{3_n})$ is independent of X and each $\varepsilon_{3_{i\in\{1,\ldots,n\}}} \sim N(0, 0.1)$.

The interactive couple is (X_1, X_6).

2.4.4. Simulation 4

In this scenario of simulation, we consider:

$$Y = X_{10} \cdot X_{51} + \varepsilon_4.$$

Where X is a matrix of $n \in \{50, 100, 150\}$ rows and $p \in \{100, 200, 300\}$ columns. Each row is generated as follows: $X_i \sim N(0_p, \Sigma)$ for all $i \in \{1, \ldots, n\}$, where $\Sigma = (\Sigma_{jk}) = (0.1^{|j-k|})$ and $(j, k) \in \{1, \ldots, p\} \times \{1, \ldots, p\}$.

$\varepsilon_4 = (\varepsilon_{4_1}, \ldots, \varepsilon_{4_n})$ is independent of X and each $\varepsilon_{4_{i\in\{1,\ldots,n\}}} \sim N(0, 0.3)$.

The interactive couple is (X_1, X_6).

2.4.5. Simulation 5

We consider a simulation scenario used in [7], precisely the fourth one:

$$Y = X_1 + X_3 + X_6 + X_{10} + 3 \cdot (X_1 \times X_3) + 3 \cdot (X_6 \times X_{10}).$$

Such that $(n, p) = (100, 500)$ and the number of replicates is 100. Also each row of the matrix X is generated as follows: $X_i \sim N(0_p, \Sigma)$ for all $i \in \{1, \ldots, n\}$, where $\Sigma = (\Sigma_{jk}) = (0.1^{|j-k|})$ and $(j, k) \in \{1, \ldots, p\} \times \{1, \ldots, p\}$.

And here, similar to [7], we will compare the performance of our algorithm with that of iForm [8].

3. Results and discussion

3.1. Results of simulation 1

The analysis results of this example are given in Table 1 which contains the selection frequency of the supposedly interactive pair (X_1,X_2) by our algorithm FIF and by the JCIS method.

After applying 100 times our Fused Interaction Filter algorithm and the JCIS algorithm, we computed the probabilities of selecting interaction effects by the two methods, knowing from the expression of the model for this simulation scenario that the important couple index is $\{(1, 2)\}$. The results show a good performance of the two approaches with these simulated data (see Table 1 above).

Table 1. Frequencies of replicates selecting interactive couples using FIF and JCIS methods in Simulation 1.

Interactive pair	FIF	JCIS
(X_1, X_2)	0.90	1

3.2. Results of simulation 2

Fig. 2. Shows the interaction screening procedure using the Fused Interaction Filter.

In this example, we verify the ability of our approach FIF to detect interactive pairs among the $\frac{p(p-1)}{2}$ pairs of features by comparing the FIF scores of pairs (X_1, X_2) and (X_3, X_4) to the maximum FIF score of other pairs of features using a violin plot [6] (see Figure 2).

3.3. Results of simulation 3

Table 2. Frequencies of replicates selecting interactive couples using FIF and JCIS methods in simulation 3.

Interactive pair	FIF	JCIS
(X_1, X_6)	1	1

We compute the frequencies of replicates selecting interactive couples (X_1, X_6) using FIF and JCIS methods. Table 2 summarizes our findings.

The results show a good performance of our algorithm and its competitiveness (see Table 2).

3.4. Results of simulation 4

For this example, we compute the selection frequency of the interactive pair (X_{10}, X_{51}) by our algorithm and analyze the results found when increasing the dimension of data from $(n, p) = (50, 100)$ then $(n, p) = (100, 200)$ to $(n, p) = (150, 300)$.

Table 3. Frequencies of replicates selecting interactive couples using our algorithm for different values of (n, p) in simulation 4.

(n, p)	Frequency of selecting (X_{10}, X_{51})
$(n, p) = (50, 100)$	0.76
$(n, p) = (100, 200)$	1
$(n, p) = (150, 300)$	1

The results are given in Table 3 below.

After applying 50 times our Fused Interaction Filter algorithm, we computed the probabilities of selecting interaction effects. The results show a good performance of our approach, especially when the dimension is high.

3.5. Results of simulation 5

Table 4. Frequencies of replicates selecting interactive couples using our algorithm and iform method in simulation 5.

Interactive pair	FIF	iform
(X_1, X_3)	0.42	0.21
(X_6, X_{10})	0.36	0.19

In this example, we calculate, over 100 replicates, the selection frequency of the two interactive pairs: (X_1, X_3) and (X_6, X_{10}) using our algorithm and iform method [8]. Then we compare the performance of the two methods (see Table 4 below).

The results assert the good behavior of our algorithm that performs better than the other method.

4. Conclusion

In this paper, we have presented a new method in 2 steps for interaction screening in the case of a continuous response variable for ultra-high dimensional regression models. Our algorithm has been implemented on R with data generated in 5 different scenarios, and the process has been repeated 100 times in simulations 1, 2 and 5, 47 times in simulation 3 and 50 times in simulation 4 to ensure a good simulation study. The performance evaluation phase has demonstrated the advantages of our proposal. In addition to its good behavior in practice, the proposed fused interaction filter enjoys the theoretical property of sure screening. Thus in ultra-high dimensional spaces, our approach estimates the right set of interactive pairs with a probability tending to 1.

[1] Mai Q., Zou H. The fused Kolmogorov filter: A nonparametric model-free screening method. The Annals of Statistics. **43** (4), 1471–1497 (2015).

[2] Fan J., Song R. Sure Independence Screening in Generalized Linear Models With NPDimensionality. The Annals of Statistics. **38** (6), 3567–3604 (2010).

[3] Huang D., Li R., Wang H. Feature Screening for Ultrahigh Dimensional Categorical Data with Applications. Journal of Business & Economic Statistics. **32** (2), 237–244 (2014).

[4] Fan Y., Kong Y., Li D., Lv J. Interaction pursuit with feature screening and selection. Preprint arXiv:1605.08933 (2016).

[5] Fan J., Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society, Series B: Statistical Methodology. **70** (5), 849–911 (2008).

[6] Anzarmou Y., Mkhadri A., Oualkacha K. The Kendall interaction filter for variable interaction screening in ultra high dimensional classification problems. Journal of Applied Statistics. **50** (7), 1496–1514 (2020).

[7] Reese R., Dai X., Fu G. Strong Sure Screening of Ultra-high Dimensional Data with Interaction Effects. Preprint arXiv:1801.07785 (2018).

[8] Hao N., Zhang H. H. Interaction Screening for Ultrahigh-Dimensional Data. Journal of the American Statistical Association. **109** (507), 1285–1301 (2014).

[9] Niu Y. S., Hao N., Zhang H. H. Interaction screening by partial correlation. Statistics and Its Interface. **11** (2), 317–325 (2018).

[10] Moore J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity. **56** (1–3), 73–82 (2003).

[11] Cordell H. J. Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics. **10** (6), 392–404 (2009).

[12] Cook R. D., Zhang X. Fused estimators of the central subspace in sufficient dimension reduction. Journal of the American Statistical Association. **109** (506), 815–827 (2014).

Алгоритм скринінгу ознак для багатовимірних даних

Чамлал Х., Бенцмане А., Уадерман Т.

Факультет наук Айн Чок, Університет Хасана II,
8 км Дороги Ель Джадіда, пошт. скр. 5366 Маарiф 20100, Касабланка 20000, Марокко

На даний час скринінг ознак стає важливою темою в галузі машинного навчання й аналізу багатовимірних даних. Відфільтрування нерелевантних ознак із набору змінних вважається важливим попереднім кроком, який слід виконувати перед будь-яким аналізом даних. Багато дослідників запропонували нові підходи до цієї теми після того, як Фан та Лв (J. Royal Stat. Soc. **70** (5), 849–911 (2008)) ввели властивість надійного скринінгу. Однак продуктивність цих підходів відрізняється від методу до методу. У запропонованій роботі є намагання додати до цього списку новий алгоритм, який виконує скринінг ознак на основі фільтра взаємодії Кендалла (J. Appl. Stat. **50** (7), 1496–1514 (2020)), коли змінна відповідь є неперервною. Добра поведінка нашого алгоритму доводиться за декількома сценаріями моделювання через порівняння з існуючим методом.

Ключові слова: *скринінг ознак; дискретизація; багатовимірні дані; регресія.*