Journal of the Korean Statistical Society 43 (2014) 567–583

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

# Independent feature screening for ultrahigh-dimensional models with interactions

Yunquan Song [a,b], Xuehu Zhu [b], Lu Lin [b,*]

[a] College of Science, China University of Petroleum, Qingdao, China
[b] School of Mathematics, Shandong University, Jinan, China

ARTICLE INFO

*Article history:*
Received 12 March 2013
Accepted 10 March 2014
Available online 28 March 2014

*AMS 2000 subject classifications:*
62F05
62P10

*Keywords:*
Feature ranking
Variable selection
Interaction term
Model-free

ABSTRACT

Feature selection is an important technique for ultrahigh-dimensional data analysis. Most feature selection methods such as SIS and its relevant versions heavily depend on the specified model structures. Furthermore, feature interactions are usually not taken into account in the existing literature. In this paper, we present a novel feature selection method for the model with variable interactions, without the use of structure assumption. Thus, the new ranking criterion is flexible and can deal with the models that contain interactions. Moreover, the new screening procedures are not complex, consequently, they are computationally efficient and the theoretical properties such as the ranking consistency and sure screening properties can be easily obtained. Several real and simulation examples are presented to illustrate the methodology.



## 1. Introduction

At the beginning of a scientific investigation, we often meet the case where the dimensionality $p$ of data grows at an exponential rate as the sample size $n$ increases (say, $\log p = O(n^b)$ for some $b > 0$). This case is called ultrahigh-dimensional data. Feature selection is an important technique for ultrahigh-dimensional data analysis. Fan and Lv (2008) proposed a feature selection method, sure independence screening (SIS), for a common linear model. Hall and Miller (2009) extended Pearson correlation learning by considering polynomial transformations of predictors. Fan, Samworth, and Wu (2009) and Fan and Song (2010) proposed more general versions of independent learning via ranking the maximum marginal likelihood estimators. Zhu, Li, Li, and Zhu (2011) proposed a sure independent ranking and screening (SIRS) procedure to screen significant predictors in multi-index models. Fan, Feng, and Song (2011) further extended the correlation learning to marginal nonparametric learning. They considered nonparametric independence screening for sparse ultrahigh dimensional additive models. However, the SIS procedure and its extended versions depend on the rank of the Pearson correlation between the response variable and predicator variables, which is not robust against the outliers and strong influence points. Based on the more robust Kendall's $\tau$ rank correlation, Li, Peng, Zhang, and Zhu (2012) developed the robust rank correlation screening (RRCS) and studied the sure screening properties of the RRCS for ultrahigh-dimensional linear regression models and transformation regression models. Lin, Sun, and Zhu (2013) proposed a nonparametric function-correlative feature screening (NRS). The NRS does not need any assumption on structural relationships between response and predictors, and among

* Corresponding author.
*E-mail address:* linlu@sdu.edu.cn (L. Lin).

http://dx.doi.org/10.1016/j.jkss.2014.03.001
1226-3192/

predictors. By using local information flows of model variables, the function-correlation between response and predictors is captured successfully.

Most of the foregoing feature selection methods involve ranking and selecting feature individually and do not require many computational resources. However, feature interactions in the model under study are usually not taken into account in the existing literatures. Interaction effects represent the combined effects of variables on the criterion or dependent measure. When an interaction effect is present, the impact of one variable depends on the level of the other variable. In practice, we also often meet the interaction features. For instance, gene–gene interaction has gained increasing attention in studies of complex diseases. Recent biological studies successfully identified thousands of risk factors associated with common human diseases. However, because most of these studies used a single-variable method and each variable was analyzed individually, the risk factors identification only can account for a small portion of disease heritability. Recently, a growing body of evidence including Carlborg and Haley (2004), Khan et al. (2011), Moore and Williams (2009), Shao et al. (2008) and Zuk et al. (2011), suggests gene–gene interactions as a possible reason for the missing heritability. So gene–gene interaction as an ubiquitous component of genetic architecture of common human diseases has been contemplated in the existing studies.

As can be seen from the above description, feature interactions in the model under study should be taken into account. When interaction effects are present, unintentional removal of these interaction features can result in the loss of useful information and thus may cause poor classification performance (see the simulation studies in Section 5). Therefore, as Pedhazur and Schmelkin (1991) noted, interaction effects should be studied in the research procedure. However, feature screening for the models with the interactions is not well developed.

In this paper, we develop a novel feature screening procedure for the model that contains variable interaction. It is worth mentioning that our proposed feature ranking criterion is simple, but the models under study are complex because they contain the interaction terms. The proposed method extracts different types of information from the data in several steps. We identify the active single variable by the feature screening method in the first step. In the second step, we need to observe all the possible two-way interaction terms that are constituted by the variables selected in the first step and select the active two-way interactions. The treatment procedure for three-way and higher-way interactions is the same as two-way interactions. We will stop the screening procedure until no higher-way interaction term is active. The final step combines these active sets to form the eventually selected active set. Thus we can ensure that all of the active single variables and interactions can be selected by these simple and programmed steps. The feature ranking criteria defined in this paper have more wide application because they are free of any model structure assumption on the relation between the predictors and the response. That is, our proposal can handle not only the models with the interactions but also the classical linear model, the nonlinear model and the additive model. Moreover, the theoretical properties such as the ranking consistency property and the sure screening property can be easily obtained and the numerical computation is not complex as well.

The rest of the paper is organized as follows. In Section 2, we propose the models that may contain interaction terms, and make explanations about the relevant issues of the models. Then, motivated by the idea of the RRCS, we develop a new model-free feature screening method for the ultrahigh dimensional regression models that may contain the interaction terms among the predictors. The theoretical properties for the new proposal are presented in Section 3. In Section 4, we introduce the thresholding rule for model selection. Numerical simulations in Section 5 demonstrate the effectiveness of the new proposed method. We relegate the proofs to Appendix.

## 2. Methodology

### 2.1. Model and problems

Screening features in ultrahigh-dimensional data analysis has become increasingly important in diverse scientific fields. Many feature-selection methods have been proposed for ultrahigh-dimensional data. They do not require many computational resources. However, most of the existing methods ignore feature interactions. Those interactions are of didactic interest, in terms of the information that they convey about how features work together in a particular population. Therefore, it is very necessary to put forward a new approach. Here, a simple and practical approach is investigated, also requiring relatively few computational resources but nevertheless allowing identification of interactions. The new approach also enables remarkable improvements in the performance of simple classical classifiers. For example, the model which contains $p = 2000$ predictor variables is given by

$$Y = \beta_0 + \sum_{j=1}^{3} \frac{4-j}{3} \left(X_j + X_{j+3} + X_j X_{j+3}\right) + \sum_{j=7}^{p} \beta_j X_j + \varepsilon. \tag{2.1}$$

Here $(X_j, X_{j+3})$ for $1 \le j \le 3$ denote the independent pairs of normal random variables with zero means, unit variances and correlation equal to 0.85. Other single features $X_7, \ldots, X_p$ follow the standard $N(0, 1)$ distribution with coefficients $\beta_j = 0,\ j = 7, \ldots, p$, and the intercept $\beta_0 = -2.5$. In this paper we refer to $X_i$ as the constitutive term (single feature) and the product of constitutive terms, $X_j X_{j+3}$, as the two-way interaction term (cross features). We can also call $X_j$ the main effect and $X_j X_{j+3}$ the interaction effect. Assume that we get a sample with the size $n = 100$. From the simulation studies