
Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

# Independent screening in high-dimensional exponential family predictors' space

Kofi Placid Adragni[a]
[a] Math & Stat, University of Maryland Baltimore County, 1000 Hilltop Circle, 410 MP Building, Baltimore, MD 21250, USA
Published online: 20 Aug 2014.

To cite this article: Kofi Placid Adragni (2014): Independent screening in high-dimensional exponential family predictors' space, Journal of Applied Statistics, DOI: 10.1080/02664763.2014.949640

To link to this article: http://dx.doi.org/10.1080/02664763.2014.949640

*Journal of Applied Statistics*, 2014
http://dx.doi.org/10.1080/02664763.2014.949640



# Independent screening in high-dimensional exponential family predictors' space

Kofi Placid Adragni*

*Math & Stat, University of Maryland Baltimore County, 1000 Hilltop Circle, 410 MP Building, Baltimore, MD 21250, USA*

(*Received 22 November 2013; accepted 25 July 2014*)

We present a methodology for screening predictors that, given the response, follow a one-parameter exponential family distributions. Screening predictors can be an important step in regressions when the number of predictors $p$ is excessively large or larger than $n$ the number of observations. We consider instances where a large number of predictors are suspected irrelevant for having no information about the response. The proposed methodology helps remove these irrelevant predictors while capturing those linearly or nonlinearly related to the response.

**Keywords:** variable screening; inverse regression; sufficient dimension reduction; high dimensionality



## 1. Introduction

We consider a regression context with a response variable $Y$ and a set of $p$ predictors $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$. We propose a new method for variable screening in a high-dimensional predictors' space. It is the first screening methodology for predictors that, given the response, follow one-parameter exponential family distribution. The need for this methodology arises as scientists are routinely formulating regressions for which the number of predictors $p$ is large and often larger than the number of observations $n$. This occurs in research fields including biology, finance, and chemometrics, etc. In genomics, for example, with DNA sequencing technology, information on tens of thousands of single nucleotide polymorphisms (SNPs) (categorical predictors) can be obtained with only a few hundreds of subjects. Dealing with data sets with $p \gg n$ in forward regressions is a challenge often referred to as 'large $p$ small $n$ problems'. It is often observed that among the large set of predictors, a sizable number is irrelevant in explaining the response. These inactive predictors are to be screened out to reduce the excessively large data set with a minimal loss of regression information.

A number of variable screening methods have been developed recently. The sure independent screening (SIS) of Fan and Lv [9] and the forward regression screening of Wang [17] both assume a linear regression model of $Y \mid \mathbf{X}$ to select the most important variables. These are

*Email: kofi@umbc.edu

mostly a correlation screening based on the strength of the marginal linear relationship between individual predictors and the outcome. The idea of sure independence screening was extended to exponential family response $Y$ by Fan and Song [10] where a generalized linear model was used in selecting the most active variables. When the marginal relationship between the predictors and the response is nonlinear, correlation screening often performs poorly. Because of that, Fan *et al.* [8] proposed a nonparametric independence screening in sparse ultra-high dimensional additive models where splines on the predictors were used to help capture the active ones, and Zhu *et al.* [18] developed a model-free approach where both $\mathbf{X}$ and $Y$ are random.

Our approach is different from those mentioned above as we consider an inverse regression of $\mathbf{X} \mid Y$ to determine the active predictors. It has been observed that in most high-dimensional data, the predictors are random and not fixed by design, as well as the response. There is no compelling reason not to consider the regression of $\mathbf{X} \mid Y$ as long as it does not contradict the sampling scheme used to collect the data. With recent development of sufficient dimension methodologies, inverse regression become viable methods to help deal with high-dimensional data. A reduction $\zeta^{\mathrm{T}}\mathbf{X}, \zeta \in \mathbb{R}^{p\times d}, d \leq p$ is sufficient if $Y \perp\!\!\!\perp \mathbf{X}|\zeta^{\mathrm{T}}\mathbf{X}$, that is $\zeta^{\mathrm{T}}\mathbf{X}$ retains all regression information about $Y$ contained in $\mathbf{X}$. The symbol $\perp\!\!\!\perp$ stands for statistical independence. Our screening method is based on a class of likelihood-based inverse regression methods called principal fitted components (PFC) pioneered by Cook [5] and further developed by Cook and Forzani [6] and Cook and Li [7] for sufficient dimension reduction. These methodologies utilize basis functions to help capture any arbitrary relationship between the predictors and the response. The basis functions can be splines obtained using the response observations and they equip the PFC models with versatility to adapt to a range of relationships between the predictors and the response. The relevance of each predictor is determined by a test statistic.

In the remainder of the article, we present the development of the screening procedure for exponential family predictors in Section 2. We show a connection between our method and existing methods in Section 3, and provide a set of simulations in Section 4 where an application to a data set is also featured. We finally provide some existing theoretical results about the consistency of the test, followed by a discussion in Section 6.



## 2. Variable screening with generalized PFC

Let $X_{jy}$, $j = 1, \ldots, p$ denote the random variable distributed as $X_j \mid (Y = y)$. We assume that the conditional predictors $X_{jy}$ follows a one-parameter exponential model of the form

$$g_j(x \mid Y = y, \eta_{jy}) = \exp\{\eta_{jy}x - B(\eta_{jy})\}H(x). \tag{1}$$

We assume that $\eta_{jy}$, the natural parameters, is a function of $y$ and suppose that $\eta_{jy} = \bar{\eta} + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\nu}_y$. Let $\bar{\boldsymbol{\eta}} = \mathrm{E}(\boldsymbol{\eta}_y)$ with $\boldsymbol{\eta}_y = (\eta_{1y}, \ldots, \eta_{py})^{\mathrm{T}}$, and let $\mathcal{S}_{\boldsymbol{\Gamma}} = \mathrm{span}\{\boldsymbol{\eta}_y - \bar{\boldsymbol{\eta}} | y \in S_Y\}$, where $S_Y$ is the sample space of $Y$. The term $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ is a semi-orthogonal matrix whose columns form a basis for the $d$ dimensional subspace $\mathcal{S}_{\boldsymbol{\Gamma}}$. The vector of natural parameters can be written as

$$\boldsymbol{\eta}_y = \bar{\boldsymbol{\eta}} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y, \tag{2}$$

where $\boldsymbol{\nu}_y$ is an unknown function of $y$. However, once $y$ is observed, $\boldsymbol{\nu}_y$ can be modeled and assumed to be $\boldsymbol{\nu}_y = \boldsymbol{\beta}\{\mathbf{f}_y - \bar{\mathbf{f}}\}$, where $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$ is an unrestricted rank $d$ matrix, $\mathbf{f}_y \in \mathbb{R}^r$ is a known user-selected function of $y$ with linearly independent elements, and $\bar{\mathbf{f}}$ is the sample mean of $\mathbf{f}_{y_i}$, $i = 1, \ldots, n$. We will refer to $\mathbf{f}_y$ as a *basis function*. The vector of natural parameters becomes

$$\boldsymbol{\eta}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y, \tag{3}$$

with $\boldsymbol{\mu} = \bar{\boldsymbol{\eta}} + \boldsymbol{\Gamma}\boldsymbol{\beta}\bar{\mathbf{f}}$. This model is referred to as the generalized PFCs model. The dependency between the predictors and the response is captured through $\boldsymbol{\Gamma}$. The following proposition states that $\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{X}$ is a minimal sufficient reduction.

PROPOSITION 2.1 ([7]) *Let $R(\mathbf{X}) = \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{X}$ and let $T(\mathbf{X})$ be any sufficient reduction. Then, under models (2) and (3), $R$ is a sufficient reduction and $R$ is a function of $T$.*

The sufficient reduction is a set of linear combinations of the $p$ predictors. A predictor $X_j$ will contribute to the reduction when the corresponding row $\boldsymbol{\gamma}_j$ of $\boldsymbol{\Gamma}$ is nonzero, thus we can focus on individual predictors for their dependency on the response. The natural parameter of the distribution of $X_{jy}$ is

$$\eta_{jy} = \mu_j + \boldsymbol{\phi}_j^{\mathrm{T}}\mathbf{f}_y, \tag{4}$$

where $\boldsymbol{\phi}_j = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\gamma}_j$. This univariate model (4) is a generalized linear model with predictor vector $\mathbf{f}_y$. The term $\boldsymbol{\nu}_y$ in Equation (2) is the exact function of $y$ that helps in capturing the dependency of a predictor $X_j$ on the response. It is approximated by a basis function. Restating the natural parameter as

$$\eta_{jy} = \bar{\eta}_j + \boldsymbol{\gamma}_j^{\mathrm{T}}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}) + \boldsymbol{\gamma}_j^{\mathrm{T}}[\boldsymbol{\nu}_y - \boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})], \tag{5}$$

it appears that $\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})$ would a reasonable proxy for $\boldsymbol{\nu}_y$ if the following conditions are satisfied.

$$E\{\boldsymbol{\nu}_Y - \boldsymbol{\beta}[\mathbf{f}_Y - E(\mathbf{f}_Y)]\} = 0, \tag{6}$$

$$\boldsymbol{\nu}_Y - \boldsymbol{\beta}[\mathbf{f}_Y - E(\mathbf{f}_Y)] \perp\!\!\!\perp Y. \tag{7}$$

Condition (7) seems the most important since (6) can be obtained by construction. The relationship between $\boldsymbol{\nu}_Y$ and $\boldsymbol{\beta}[\mathbf{f}_Y - E(\mathbf{f}_Y)]$ should be linear. The following proposition sets the framework for screening method.

PROPOSITION 2.2 *Let us consider model (4) and assume that $Cov(\boldsymbol{\nu}_Y, \mathbf{f}_Y)$ is of full rank. If $Var(\mathbf{f}_Y)$ is nonsingular with finite elements, then $X_j \perp\!\!\!\perp Y$ if and only if $\boldsymbol{\phi}_j = 0$.*

This proposition ties the screening procedure to the choice of the basis function. The performance of the method depends on the adequate choice of $\mathbf{f}_y$ which elements should be reasonably flexible. We may still expect the method to perform adequately under misspecification of $\mathbf{f}_y$ provided that $\mathrm{Cov}(\boldsymbol{\nu}_Y, \mathbf{f}_Y) \neq 0$.

The dimension $d$ of the sufficient reduction in Proposition 2.1 is to be estimated. When $n$ is large enough, and $p$ relatively small, $d$ could be estimated by a likelihood ratio test [6], or by cross-validation [1]. In this screening process, and henceforth, we assume that $d = r$ where $r$ is determined by the choice of the basis function $\mathbf{f}_y$. The construction and choice of $\mathbf{f}_y$ is provided in Section 2.2.

In the above development, we have assumed that the predictors are conditionally independent. Cook and Li [7] studied the case of quadratic exponential models to allow dependency among the conditional predictors. They demonstrate that the methodology based on the conditional independence to obtain the sufficient reduction can still be useful when the predictors are conditionally dependent.

### 2.1 *Screening via hypothesis test*

The screening procedure falls upon whether $\boldsymbol{\phi}_j = 0$ for each predictor $X_j$. At least two approaches could be used. The first is by testing the hypothesis

$$H_0 : \boldsymbol{\phi}_j = 0 \quad \text{against} \quad H_a : \boldsymbol{\phi}_j \neq 0 \tag{8}$$

at a specified significance level $\alpha$. The predictor $X_j$ will be set as *active* or *relevant* when we reject $H_0$, and *inactive* or *irrelevant* otherwise. In an alternative approach, the magnitude of the estimates $\|\hat{\boldsymbol{\phi}}_j\|$ of $\boldsymbol{\phi}_j$ for $j = 1, \ldots, p$ would be obtained and sorted decreasingly as $\|\hat{\boldsymbol{\phi}}_{(1)}\| \geq \|\hat{\boldsymbol{\phi}}_{(2)}\| \geq \cdots \geq \|\hat{\boldsymbol{\phi}}_{(p)}\|$. The active predictors are those with $\|\hat{\boldsymbol{\phi}}_j\|$ greater than a threshold $\theta$. While the value of $\theta$ may be arbitrary for the screening procedure, it could be estimated by an information criterion or by cross-validation.

We proceed henceforth with the first approach tieing the selection of active predictors to a test statistic. The method is likely to select any predictor having any marginal relationship with the response, whether it is linear or arbitrary. There is no restriction to the number of predictors to be selected once the level of significance $\alpha$ is set. Dimension reduction methods such as PFC [5] may be applied to the screened set for further reduction.

### 2.2 *The basis functions*

The use of basis functions is usually convenient to approximate nonlinear functions. Given a function $\boldsymbol{v}_y$, we want to find the transformations $\mathbf{f}_y = (f_1(y), \ldots, f_r(y))^{\mathrm{T}}$ such that

$$\boldsymbol{v}_y \approx \sum_{i=1}^{r} \beta_i f_i(y).$$

The vector function $\mathbf{f}_y$ constitutes the basis functions to be used. Several basis functions have been suggested in conjunction with PFC models [1,5]. To gain an insight into the choice of the appropriate basis function, inverse response plots (Cook, 1998) of $X_j$ versus $y$, $j = 1, \ldots, p$ could be used. In some regressions there may be a natural choice for $\mathbf{f}_y$. For example, suppose for instance that $Y$ is categorical, taking values in one of $h$ categories $C_k$, $k = 1, \ldots, h$. We can then set the $k$th element of $\mathbf{f}_y$ to be $I(y \in C_k)$, where $I(\cdot)$ is the indicator function.

When graphical guidance is not available, Cook [5] suggested constructing a piecewise constant basis. With a continuous $Y$, an option consists of 'slicing' the observed values into $h$ bins (categories) $C_k$, $k = 1, \ldots, h$, and then specifying the $k$th coordinate of $\mathbf{f}_y$ as for the case of a categorical $Y$. This has the effect of approximating each conditional mean $\mathrm{E}(X_j \mid Y = y)$ as a step function of $Y$ with $h$ steps.

Following the piecewise constant basis, we investigated piecewise polynomial bases, including piecewise linear, piecewise quadratic and piecewise cubic polynomial when the response is continuous. We construct the piecewise polynomial bases also by slicing the observed values of $Y$ into bins $C_k$, $k = 1, \ldots, h$. Let $\tau_0, \tau_1, \ldots \tau_h$ denote the end-points or knots of the slices. For example, $(\tau_0, \tau_1)$ are the end-points of $C_1$; $(\tau_1, \tau_2)$ are the end-points of $C_2$, and so on. We give here the description of the piecewise linear discontinuous basis $\mathbf{f}_y \in \mathbb{R}^{2h-1}$ by its coordinates

$$\begin{aligned} f_{2i-1}(y) &= I(y \in C_i), \quad i = 1, 2, \ldots, h-1 \\ f_{2i}(y) &= I(y \in C_i)(y - \tau_{i-1}), \quad i = 1, 2, \ldots, h-1 \\ f_{2h-1}(y) &= I(y \in C_h)(y - \tau_{h-1}). \end{aligned} \tag{9}$$

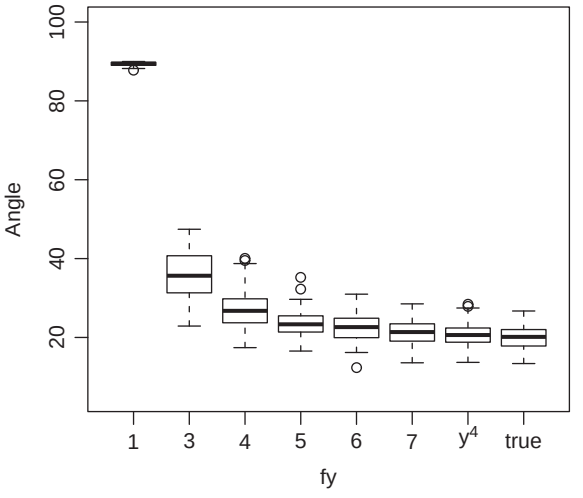


Figure 1. Boxplots of the angle between span{$\boldsymbol{\Gamma}$} and its estimator. Boxplots 1–6 are for the PFC estimators under piecewise linear basis; the numbers 1–7 represent the number of slices. Boxplot on $y^4$ is for $\mathbf{f}_y = (y, y^2, y^3, y^4)$ and the last is the true basis used to generate the data.



The number $r$ of linearly independent components of $\mathbf{f}_y$ is to be small enough compared to $n$ to avoid modeling random variation rather than the overall shape between the response and individual predictor.

We performed simulation studies comparing the use of piecewise linear discontinuous basis functions to polynomial bases in PFC models. We generated $n = 200$ observations with the response $\boldsymbol{y} \in \mathbb{R}^n$ drawn independently from $U(0, 4)$. We obtained $\mathbf{X} = \mathbf{F}\boldsymbol{\beta}\boldsymbol{\Gamma}^{\mathrm{T}} + \mathbf{e}$ where $\mathbf{F}$ is a $n \times 2$ matrix with the $i$th row being $(y_i \cos(y_i), \exp(y_i))$, $\boldsymbol{\beta} = \mathrm{Diag}(1, 0.1)$, $\mathbf{e} \in \mathbb{R}^{n\times p}$ is a matrix of independent $N(0, 1)$ variates. We used $\boldsymbol{\Gamma} = (\Gamma_1, \Gamma_2)$ where $\Gamma_1 = (\mathbf{1}_{p/2}^{\mathrm{T}}, \mathbf{0}_{p/2}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{p/2}$ and $\Gamma_2 = (\mathbf{0}_{p/2}^{\mathrm{T}}, \mathbf{1}_{p/2}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{p/2}$ with $p = 20$ predictors. Here, $\mathbf{1}_p$ and $\mathbf{0}_p$ stand for vectors of length $p$ of, respectively, 1's and 0's. We computed the angle between $\boldsymbol{\Gamma}$ and its estimator. The results shown in Figure 1 are for 100 replications. For reference, the expected angle between $\boldsymbol{\Gamma}$ and a randomly chosen vector in $\mathbb{R}^{20}$ is about 80°. Piecewise linear discontinuous basis with seven slices is not statistically different from the fourth degree polynomial basis or the basis used to generate the data. We were able to obtain similar results with piecewise quadratic and cubic discontinuous bases.

We should note that for a categorical response, there is only one basis function, that is a piecewise constant basis. However for a continuous response, there is in theory an infinite number of basis functions. In practice, a cubic polynomial basis is often adequate to capture nonlinear dependencies. Several basis functions are implemented with more details in the R package `ldr` [2] under the function `bf`.

## 3. Connection with other methods

We consider the normal inverse regression model $\mathbf{X}_y \sim N(\boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y, \boldsymbol{\Delta})$, where $\boldsymbol{\Delta} > 0$ is assumed independent of $Y$, and $\boldsymbol{\mu} = E(\mathbf{X})$. Let $\mathbf{X}_1$ be the vector of active predictors and $\mathbf{X}_2$ the vector of inactive predictors to form a partition of $\mathbf{X}$. The following proposition presents the condition for screening when $\mathbf{X}_1$ and $\mathbf{X}_2$ are conditionally independent.

PROPOSITION 3.1 *We assume that* $\mathbf{X}_y \sim N(\boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y, \boldsymbol{\Delta})$ *and that* $\mathbf{X}_2 \perp\!\!\!\perp \mathbf{X}_1 \mid Y$. *Let us partition* $\boldsymbol{\Gamma}$ *and* $\boldsymbol{\Delta}$ *according to the partition of* $\mathbf{X}$ *as* $(\boldsymbol{\Gamma}_1^{\mathrm{T}}, \boldsymbol{\Gamma}_2^{\mathrm{T}})^{\mathrm{T}}$, *and* $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{ij})$, $i = 1, 2$; $j = 1, 2$. *Then* $\mathbf{X}_2 \perp\!\!\!\perp Y$ *if and only if* $\boldsymbol{\Gamma}_2 = 0$.

Screening the predictors by testing whether $\boldsymbol{\Gamma}_2 = 0$ can be done by considering individual rows of $\boldsymbol{\Gamma}$ and testing whether $\boldsymbol{\gamma}_j = 0, j = 1, \ldots, p$, or equivalently, the hypotheses (8) for $j = 1, \ldots, p$. The model could be written for individual predictors as

$$X_{jy} = \mu_j + \boldsymbol{\phi}_j^{\mathrm{T}}\mathbf{f}_y + \delta_j\epsilon, \quad \epsilon \sim N(0, 1), \tag{10}$$

The relevance of a predictor $X_j$ is assessed by determining whether the mean function $E(X_{jy}) = \mu_j + \boldsymbol{\phi}_j^{\mathrm{T}}\mathbf{f}_y$ depends on the outcome $y$. Model (10) is a forward linear regression model where the 'predictor' vector is $\mathbf{f}_y$ and the 'response' is $X_j$. An $F$-test statistic can be obtained for the hypotheses (8). The predictor $X_j$ is relevant if the model yields an $F$-statistic larger than a user-specified cutoff value. The cutoff may correspond to a significance level $\alpha$ such as 0.1 or 0.05 for example.



### 3.1 *Correlation screening*

Now, let assume that $Y$ and $\mathbf{X}$ are centered around their means and restrict $\boldsymbol{\Gamma}$ to be a $p \times 1$ matrix. Furthermore, we set $\mathbf{f}_y = y$, $\boldsymbol{\Delta} = \delta^2\mathbf{I}$ and absorb $\boldsymbol{\beta}$ into $\boldsymbol{\Gamma}$. These restrictions yield the simplest expression of the PFC model as $\mathbf{X}_y \sim N(\boldsymbol{\Gamma}y, \delta^2\mathbf{I})$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\mathbb{X}$ the $p \times n$ centered data-matrix of the predictors. Under this model, let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^{\mathrm{T}}$ be the $p$-vector obtain by the componentwise regression of $\mathbf{X}$ on $y$ as $\boldsymbol{\omega} = \boldsymbol{\Gamma}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$. This vector $\boldsymbol{\omega}$ is proportional to the sample correlation of $\mathbb{X}$ and $\boldsymbol{y}$. For a given response vector, it appears that $\boldsymbol{\omega}$ and $\boldsymbol{\Gamma}$ are proportional. The SIS of Fan and Lv [9] sorts the $p$ componentwise magnitudes of the vector $\boldsymbol{\omega}$ in a decreasing order and select the first $p^*$, $(p^* < n)$ corresponding predictors as the active ones. It is clear that the componentwise magnitudes of $\boldsymbol{\Gamma}$ retain the same ordering as $\boldsymbol{\omega}$. Thus, SIS is a particular case of our method. Furthermore, rather than selecting an arbitrary number of active predictors as with SIS, our approach ties the selection to a test statistic.

Screening predictors based on $\boldsymbol{\omega}$ is called a correlation screening since $\boldsymbol{\omega}$ is proportional to the correlation vector of the predictors data-matrix and the response vector. The correlation screening was also used in the Supervised Principal Components [3], where a cross-validation method is used to determine $p^*$.

### 3.2 *T-test based methods*

We now turn to scenarios where the predictors are continuous and the response is discrete with $k$ categories. The discrete response yields a natural basis function that is $\mathbf{f}_y = (I(y = 1), \ldots, I(y = k)^{\mathrm{T}})^{\mathrm{T}}$. The componentwise PFC model is

$$X_{jy} = \mu_j + \sum_{i=1}^{k-1} \phi_{ij} I(y = i) + \delta_j\epsilon \quad j = 1, \ldots, p, \tag{11}$$

where $\phi_{ij}$ is the $i$th component of $\boldsymbol{\phi}_j = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\gamma}_j$. This model is a multiple linear regression with categorical predictors. The $k$th category is set as the baseline. In the particular case of binary outcome ($k = 2$), a $t$-test can be used for the hypothesis $H_0 : \phi_1 = 0$. This has been seen in the literature, especially with microarray data sets. For example, Roberts and Mukherjee [14] used the difference of means, Guyon and Elisseef [11] proposed a $t$-statistic with a pooled variance; Lai *et al.* [13] used the signal-to-noise ratio.

### 3.3 *Multinomial predictors*

The multinomial case may be of interest in genetics for example for screening the excessively large data made of SNPs. SNPs are found to be of great importance in biomedical research and are assumed to play an important role in the development of complex diseases, or believed to alter the risk for developing particular diseases. SNPs data are categorical with three levels, or genotypes. Biological considerations may collapse the three levels into two. It is observed that although the SNPs have random occurrence, they are often treated as fixed predictors of a continuous type in forward regressions. It is also observed that the sampling scheme often used in collecting the genomic data, including in the Genome-wise Association Study suggests modeling $\mathbf{X}|Y$ for a number of response variables. Modeling the SNPs as a function of the response could be helpful in screening out inactive SNPs while treating these SNPs as random variables and making use of their intrinsic distribution. These SNPs could be assumed to follow a multinomial distribution with three levels, or a binomial distribution if a dominant or recessive model is assumed.

Let suppose that the predictors $X_j$, $j = 1, \ldots, p$, are independent multinomial random variables with $K$ categories. The model of $\mathbf{X}_y$ can be expressed in terms of a multivariate logit defined coordinate-wise as $\text{multlogit}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y$. The sufficient reduction is still $\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{X}$ [7]. For the univariate screening, the dependency between a predictor $X_j$ and the response is determined by whether $\boldsymbol{\varphi}_j = \boldsymbol{\gamma}_j^{\mathrm{T}}\boldsymbol{\beta}$ is nonnull. Using the set of $K - 1$ equations

$$\log\left(\frac{P(X_j = i \mid Y = y)}{P(X_j = K \mid Y = y)}\right) = \mu_{ij} + \boldsymbol{\varphi}_{ij}^{\mathrm{T}}\mathbf{f}_y, \quad i = 1, \ldots, (K-1), \tag{12}$$

with $\sum_{i=1}^{K} P(X_j = i \mid Y = y) = 1$. A predictor $X_j$ in this context is independent of $Y$ if and only if $\boldsymbol{\phi} = 0$ where $\boldsymbol{\phi} = (\boldsymbol{\varphi}_{1j}, \ldots, \boldsymbol{\varphi}_{(K-1)j})$. Following the development for normal predictors, we could also tie the screening to tests statistics, testing the hypothesis (8).

## 4. Numerical studies and application

We provide three set of simulations. The first simulation set exhibits a nonadditive regression model, and the second uses an inverse regression. Both consider continuous response and predictors. The third is with continuous response and binary predictors. In all three cases, a number of predictors were generated to be related to the response and the goal is to use the screening procedure to capture them. We provide the results as the proportion of times an active predictor is selected with 100 generated data. The significance level uses is $\alpha = 0.05$.

### 4.1 *Nonlinear forward regression simulations*

We generated $n = 70$ observations on $p = 500$ independent predictors $(X_1, \ldots, X_p)^{\mathrm{T}}$ with $X_1 \sim U(1, 10)$ and $X_i \sim N(0, 4)$, $i = 2, \ldots, p$. The response was obtained as $y = (5X_1)\epsilon$ where $\epsilon \sim N(0, 1)$.

The active predictor is $X_1$; it is selected by the correlation screening about 5% of the time. This result was as expected under random selection. On the other hand, with a piecewise linear discontinuous basis using five slices, the new method captured $X_1$ 94% of the time.

### 4.2 *Normal inverse regression simulations*

The outcome $\boldsymbol{y} \in \mathbb{R}^n$ were $n = 100$ independent draws from the uniform $U(-3, 3)$ and the predictors were $\mathbb{X} = \mathbf{F}\boldsymbol{\beta}\boldsymbol{\Gamma}^{\mathrm{T}} + 2\mathbf{e}$ with $\mathbf{F} \in \mathbb{R}^{n\times 4}$ with the $i$th row $(y_i, y_i^2, y_i \sin(y_i), |y_i|^{1/2})$, and $\boldsymbol{\Gamma} = (\Gamma_1, \ldots, \Gamma_4)$ where $\Gamma_i$ is a column vector with all entries equal to 0 except the $i$th that

Table 1. Proportion of selection of active predictors.

| Basis | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| Correlation screening | 1.00 | 0.06 | 0.05 | 0.05 |
| Quadratic polynomial basis | 1.00 | 1.00 | 0.69 | 1.00 |
| Piecewise linear basis | 1.00 | 1.00 | 1.00 | 1.00 |

is 1, and $\boldsymbol{\beta} = \text{Diag}(1, 0.5, 2, 8)$. The error $\mathbf{e} \in \mathbb{R}^{n\times p}$ is a matrix of independent $N(0, 1)$ variates. Specifically, among the $p = 500$ predictors, only the first four were active and effectively related to the response. The first predictor $X_1$ was linearly related to the response while the remaining active three were nonlinearly related to $y$ as

$$X_1 = y + 2\epsilon; \quad X_2 = 0.5y^2 + 2\epsilon; \quad X_3 = 2y\sin(y) + 2\epsilon; \quad X_4 = 8\sqrt{|y|} + 2\epsilon.$$

In screening the predictors, we considered the following three basis function setups: (1) the correlation screening (first degree polynomial basis), (2) a quadratic polynomial basis, and (3) a piecewise linear discontinuous with five slices.

The results (Table 1) show that predictor $X_1$ was selected with the correlation screening 100% of the time. But correlation screening failed drastically to select the other three relevant predictors nonlinearly related to the response variable. Piecewise linear basis showed a better performance compared to quadratic polynomial.

### 4.3 *Binary inverse regression simulations*

We generated $p = 500$ binary predictors where only three of them were active and related to the response as follows. Two hundreds observations were used. The $i$th observation of $X_j, j = 1, \ldots, 3$, was obtained as a binary outcome with probability $\pi_{ij} = (1 + \exp(-\beta_j f_j(y_i)))^{-1}$, where $f_1(y) = y, f_2(y) = |y|$, and $f_3(y) = y\sin(y)$, and $\boldsymbol{\beta} = (1, 3, 3)$. For the remaining $(p - 3)$ predictors, the $i$th observation was generated as a binary outcome with probability 0.5. We used three different basis functions in the screening process: the correlation screening, the cubic polynomial basis, and the piecewise linear discontinuous with two slices. The results in Table 2 here again showed the poor performance of the correlation screening on active predictors nonlinearly related to the response. For comparison, the proportion of selection of an inactive predictors at 5% confidence level is 0.05.

### 4.4 *An application*

We applied the screening method to a chemometrics data set. The data set is about predicting the functional hydroxyl group *OH* activity of compounds from molecular descriptors. The response (act) is the activity level of the compound, and the predictors are the nearly 300 descriptors.

Table 2. Screening binary predictors.

| Basis | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| Correlation screening | 1.00 | 0.00 | 0.00 |
| Cubic polynomial basis | 1.00 | 0.99 | 1.00 |
| Piecewise linear basis | 1.00 | 0.99 | 1.00 |

Table 3. Screening *p*-values under the correlation (CS) and the piecewise linear discontinuous basis (PL).

| Descriptor | *p*-Value CS | *p*-Value PL |
|---|---|---|
| SHsOH | 0.051 | 0.047 |
| sc3 | 0.056 | 0.000 |
| spc5 | 0.058 | 0.008 |
| tets2 | 0.059 | 0.000 |
| asn5 | 0.072 | 0.000 |
| dn2n1 | 0.079 | 0.000 |
| SssCH2 | 0.079 | 0.000 |
| dn2n5 | 0.079 | 0.000 |
| dn213 | 0.080 | 0.000 |

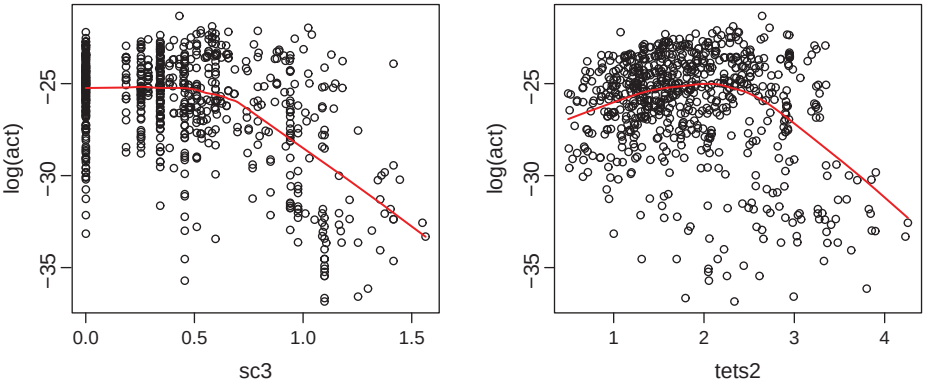


Figure 2. Scatter-plots of sc3 and tets2 against the log-transformed response act showing nonlinear relationships.



The response and predictors are continuous variables and 719 observations were recorded. We screened the descriptors at a 5% significance level.

The application of the correlation screening captured a set of 204 descriptors to be related to the response. However using the piecewise linear discontinuous basis function with two slices, the screening yielded a set of 236 descriptors, including 41 descriptors not detected by the correlation screening. The following table is a short list of some of the 41 descriptors captured by the piecewise linear discontinuous basis, that were not selected by the correlation screening at the 5% significance level.

This example showed that, as expected, when a relevant predictor is not linearly related to the response, a correlation screening may fail to capture it. However, with the appropriate choice of the basis function, all relevant predictors would be detected. Graphical investigations showed nonlinear relationships between the response and some predictors listed in Table 3. As examples, the two plots in Figure 2 are of compounds sc3 and tets2 against log(act) that were not selected by the correlation screening. A clear nonlinear relationship between the predictors and the log-transformed response is displayed. We used the log-transformation of the response to help see the graphical evidence of nonlinear relationship between the response and the mentioned predictors.

## 5. Estimation, consistency, and power

We provide existing results on the maximum likelihood estimation of the parameter $\boldsymbol{\phi}$, and on the consistency of the hypothesis test of (8). We also discuss the power of the test for large sample sizes.

### 5.1 *Estimation*

The estimation of the parameters in exponential family distribution has been well studied in both univariate and multivariate scenarios. We herein provide the maximum likelihood estimation procedure for $\boldsymbol{\phi}$; for simplicity we drop the index $j$. Let $\boldsymbol{\theta} = (\mu, \boldsymbol{\phi}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{1+r}$, and set $Z = (1, \mathbf{f}_y)^{\mathrm{T}} \in \mathbb{R}^{1+r}$. The density function (1) can be rewritten as

$$g(x \mid Y = y, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^{\mathrm{T}} Z x - B(\boldsymbol{\theta}^{\mathrm{T}} Z)\} h(x). \tag{13}$$

The following proposition gives the MLE of $\boldsymbol{\theta}$.

PROPOSITION 5.1 *Suppose that the distribution of $X$ is from the exponential family with the probability density function (13) and let $x_1, x_2, \ldots, x_n$ be a random sample from the distribution of $X$ and $y_1, \ldots, y_n$ be the observed response values. Let $B'(t)$ represents the derivative of $B$ at $t$. Assuming that the equation*

$$\sum_{i=1}^{n} Z_i [x_i - B'(\boldsymbol{\theta}^{\mathrm{T}} Z_i)] = 0 \tag{14}$$

*has a solution $\hat{\boldsymbol{\theta}}$ in the parameter space of $\boldsymbol{\theta}$, then the solution of (14) is unique, and is the MLE of $\boldsymbol{\theta}$.*

The estimation of $\boldsymbol{\theta}$ using Equation (14) can be obtained by iterated weighted least squares algorithm. Once $\hat{\boldsymbol{\theta}}$ is obtained, $\hat{\boldsymbol{\phi}}$ is made of the last $r$ entries of $\hat{\boldsymbol{\theta}}$. Details about the theorem and estimation of the maximum likelihood for exponential family can be found in [4, p. 412].

### 5.2 *Consistency*

The screening consistency is not the usual consistency of point estimators. It rather focuses on asymptotic hypothesis tests. En route to discuss this consistency, we present some definitions (see [16, p. 140]).

DEFINITION 5.2 *Let $T_n$ be a test statistic for $H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0$ against $H_a : \boldsymbol{\phi} \neq \boldsymbol{\phi}_0$ with a rejection region defined by $R(T_n) = 1$, where $R(T_n)$ can take values $\{0, 1\}$. Let $\alpha_{T_n}(\boldsymbol{\phi}_0) = P(R(T_n) = 1 \mid \boldsymbol{\phi} = \boldsymbol{\phi}_0)$ be the probability of type* I *error and $1 - \alpha_{T_n}(\boldsymbol{\phi}) = P(R(T_n) = 0 \mid \boldsymbol{\phi} \neq \boldsymbol{\phi}_0)$ be the probability of type* II *error.*

(i) *$T_n$ is consistent if and only if $\lim_{n\to\infty} \alpha_{T_n}(\boldsymbol{\phi}) = 1$ for any $\boldsymbol{\phi} \neq \boldsymbol{\phi}_0$.*
(ii) *$T_n$ is Chernoff-consistent if and only if $T_n$ is consistent and $\lim_{n\to\infty} \alpha_{T_n}(\boldsymbol{\phi}_0) = 0$.*

While the consistency in Definition 5.2(i) only requires type II error to converge to 0, Chernoff-consistency requires both type I and II to converge to 0. Let $\Lambda(\mathbf{X})$ be the likelihood ratio statistic. In the following, we consider the conditional predictors to have the density

function of the form (1), but assume that $\mu = 0$ for simplicity. The log-likelihood function $l(\boldsymbol{\phi}) = \sum_{i=1}^{n} \log g(x_i \mid \boldsymbol{\phi})$ can be written as

$$l(\boldsymbol{\phi}) = \sum_{i=1}^{n} [\boldsymbol{\phi}^{\mathrm{T}} \mathbf{f}_{y_i} x_i - B(\mathbf{f}_{y_i}^{\mathrm{T}} \boldsymbol{\phi})] + C, \tag{15}$$

where $C$ is a constant independent of $\boldsymbol{\phi}$. Let assume that $\hat{\boldsymbol{\phi}}$ the maximum-likelihood estimate of $\boldsymbol{\phi}$ exists. The test statistic to be used in testing the hypotheses (8) is $T_n = 2[l(\hat{\boldsymbol{\phi}}) - l(\boldsymbol{\phi})]$. While the exact distribution of $T_n$ can be derived for some exponential family members (normal distribution for instance), in general, the distribution of the test under the null and alternative hypotheses can be difficult to derive, and often recourse is made to asymptotic theory. It is a classical result that $T_n$ converges to a central $\chi_r^2$ under the null hypothesis. With $\mathbf{f}_y$ being the basis function used in conjunction with the model in Equation (4), using a Taylor series, we obtain an approximate expression of the test as

$$T_n = (\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})^{\mathrm{T}} \hat{\mathbf{V}}_n^{-1}(\hat{\boldsymbol{\phi}})(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}), \tag{16}$$

where $\boldsymbol{\phi} = 0$ under $H_0$, and $\hat{\mathbf{V}}_n^{-1}(\hat{\boldsymbol{\phi}}) = \sum_{i=1}^{n} \mathbf{f}_{y_i} \mathbf{f}_{y_i}^{\mathrm{T}} B''(\hat{\boldsymbol{\phi}})$, with $B''(u)$ being the second derivative of $B(u)$. The next proposition sets conditions for the consistency and Chernoff-consistency of $T_n$.

THEOREM 5.3 ([16, p. 452]) *Consider model (10) and the test statistic $T_n$ in Equation (16) for the hypotheses (8). Let $\lambda_n^{\max}$ be the largest eigenvalue of $\hat{\mathbf{V}}_n$.*

(i) *If $\lim_{n\to\infty} \lambda_n^{\max} \to 0$ then $T_n$ is consistent.*
(ii) *Assume that $\alpha = \alpha_n \to 0$ as $n \to \infty$, and $\chi^2_{r,1-\alpha_n} \lambda_n^{\max} = o(1)$, then $T_n$ is Chernoff-consistent.*

### 5.3 *Power of the tests*

In screening the set of $p$ predictors, our interest is mostly in capturing the true active predictors. Two types of error entail the hypothesis testing (8): selecting a predictor as active while it is not (type I error) and declaring a predictor to be inactive when it is active (type II error). While selecting an inactive predictor could be inconsequential, failing to select an active predictor should be avoided, thus statistical tests with high powers should be sought.

The power of the test for a given $\boldsymbol{\phi}_1$ is $P(R(T_n) = 1 \mid \boldsymbol{\phi} = \boldsymbol{\phi}_1)$. In general there is no uniformly most powerful test for the multi-parameter exponential family with composite hypotheses. However, optimal tests could be constructed [4,15]. Under the alternative hypothesis, $T_n$ has approximately a $\chi_r^2(\xi_n)$, with the noncentrality parameter

$$\xi_n(\boldsymbol{\phi}) = \boldsymbol{\phi}^{\mathrm{T}} \hat{\mathbf{V}}_n^{-1} \boldsymbol{\phi}. \tag{17}$$

For consistent tests, $\xi_n(\boldsymbol{\phi}) \to \infty$ as $n \to \infty$. Thus, there is no limiting distribution of $T_n$ under $H_a$. Pitman alternative hypotheses of the form $H_a : \boldsymbol{\phi}_n = \boldsymbol{\psi}/\sqrt{n}$ for some $\boldsymbol{\psi} \in \mathbf{R}^r$ could be considered to help estimate the limiting power.

## 6. Concluding remarks

We present a screening method for predictors that, given the response, follow an exponential family distribution. A main feature of the method is that it is likely to capture any predictor that is marginally related to the response with the use of basis functions.

The method is shown to subsume the SIS and also $t$ statistics-based screening methods. The number of relevant predictors to be selected is tied to test statistics. Thus, the significance level can be chosen for the optimal power.

There is a number of basis functions to choose from. However, for continuous predictors, the best choice of basis functions may be data set-specific. Any arbitrary marginal relationship between active predictors and the response can be accommodated with the appropriate choice of the basis function.

Results in Section 5 are essentially due to a large sample theory where the sample size is assumed to be growing to infinity. In practice, sample sizes are finite and they almost never grow. In a finite sample setting, the tests obtained in this Section 5 may not always be exact and nothing can be said in general about their exact distribution under the null or alternative hypotheses. Simulation studies could always be used to evaluate the performance of the tests. Re-sampling methods, such as the bootstrap hypothesis testing [12] can be used.



## Acknowledgement

The author thanks Tomas Oberg for providing the *OH* data set.

## References

[1] K.P. Adragni and R.D. Cook, *Sufficient dimension reduction and prediction in regression*, Phil. Trans. R. Soc. A 367 (2009), pp. 4385–4405.
[2] K.P. Adragni and A. Raim, Methods for likelihood-based dimension reduction in regression, 2014. Available at http://CRAN.R-project.org/package=ldr.
[3] E. Bair, T. Hastie, D. Paul and R. Tibshirani, *Prediction by supervised principal components*, J. Amer. Statist. Assoc. 101(473) (2006), pp. 119–137.
[4] P.J. Bickel and K.A. Doksum, *Mathematical Statistics – Basis Ideas and Selected Topics*, 2d ed., Pearson Prentice Hall, Englewood Cliffs, NJ, 2007.
[5] R.D. Cook, *Fisher lecture – dimension reduction in regression (with discussion)*, Statist. Sci. 22 (2007), pp. 1–26.
[6] R.D. Cook and L. Forzani, *Principal fitted components for dimension reduction in regression*, Statist. Sci. 23 (2009), pp. 485–501.
[7] R.D. Cook and L. Li, *Dimension reduction in regressions with exponential family predictors*, J. Comput. Graph. Statist. 18(3) (2009), pp. 774–791.
[8] J. Fan, Y. Feng, and R. Song, *Nonparametric independence screening in sparse ultra-high dimensional additive models*, J. Amer. Statist. Assoc. 106(494) (2011), pp. 544–557.
[9] J. Fan and J. Lv, *Sure independence screening for ultrahigh dimensional feature space (with discussion)*, J. R. Stat. Soc. Ser. B 70 (2008), pp. 849–911.
[10] J. Fan and R. Song, *Sure independence screening in generalized linear models with NP-dimensionality*, Ann. Statist. 38(6) (2010), pp. 3567–3604.
[11] I. Guyon and A. Elisseeff, *An introduction to variable and feature selection*, J. Mach. Learn. Res. 3 (2003), pp. 1157–1182.
[12] P. Hall and S.R. Wilson, *Two guidelines for bootstrap hypothesis testing*, Biometrics 47(2) (1991), pp. 757–762.
[13] C. Lai, M.J.T. Reinders, and L.F.A. Wessels, *Multivariate gene selection: Does it help?* IEEE Computational Systems Biology Conference, Stanford, 2005.
[14] S. Roberts and S.J. Mukherjee, *A theoretical analysis of gene selection*, Proceedings of the IEEE Computer Society Bioinformatics Conference, 2004.
[15] S.G. Self, R.H. Mauritsen, and J. Ohara, *Power calculations for likelihood ratio tests in generalized linear models*, Biometrics 48(3) (1992), pp. 1–39.
[16] J. Shao, *Mathematical Statistics*, 2nd ed., Springer, New York, NY, 2003.
[17] H. Wang, *Forward regression for ultra-high dimensional variable screening*, J. Amer. Statist. Assoc. 104(488) (2009), pp. 1512–1524.
[18] L.-P. Zhu, L. Li, R. Li, and L.-X. Zhu, *Model-free feature screening for ultrahigh dimensional data*, J. Amer. Statist. Assoc. 106(496) (2011), pp. 1464–1475.

## Appendix

### A.1 *Proof of Proposition 2.2*

*Proof* From $\mathrm{Cov}(\eta_Y, \mathbf{f}_Y) = \boldsymbol{\gamma}^{\mathrm{T}}\mathrm{Cov}(\boldsymbol{\nu}_Y, \mathbf{f}_Y) = \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\beta}\mathrm{Var}(\mathbf{f}_Y) = \boldsymbol{\phi}\mathrm{Var}(\mathbf{f}_Y)$, we have,

$$\boldsymbol{\phi} = \boldsymbol{\gamma}^{\mathrm{T}}\mathrm{Cov}(\boldsymbol{\nu}_Y, \mathbf{f}_Y)[\mathrm{Var}(\mathbf{f}_Y)]^{-1}.$$

Suppose that $X \perp\!\!\!\perp Y$. Then $\eta_Y$ does not depend on $Y$, and $\boldsymbol{\gamma} = 0$, which implies $\boldsymbol{\phi} = 0$.

Conversely, suppose that $\boldsymbol{\phi} = 0$. If $\mathrm{Cov}(\boldsymbol{\nu}_Y, \mathbf{f}_Y)$ is of full rank and $\mathrm{Var}(\mathbf{f}_Y)$ is nonsingular with finite elements, then $\boldsymbol{\gamma} = 0$. Thus the distribution of $X_y$ does not depend upon $Y$ and $X \perp\!\!\!\perp Y$. ■