A Summary of Advancements in Ultra-High Dimensional Data Analysis

Abstract: Ultra-high dimensional data, due to rapid growing technology, has provided many contemporary statistical problems. With advancements in computers and electronics, data are collected at faster rates and at higher resolutions. Many researchers are working to develop new methodology to analyze these data. Ultra-high dimensional data appear in diverse scientific fields such as high frequency trading, genomics, and brain imaging. As technology continues to advance, this type of data will only become more prevalent. In this paper, we provide a summary of ultra-high dimensional data by discussing some common sources, providing the community agreed definition, and highlighting recent advancements in methodology.

1 Introduction

1.1 Background

In 1978, John Tukey spoke about the rapid expansion of the field of data analysis [Donoho 2000]. Data analysis had an open future and new technology was continuing to create new sources of data; the expansion showed no signs of slowing down. Tukey described how these new sources of data were collecting data in much larger volumes. However, not only was the size of the data increasing but the detail and resolution of the data was becoming finer. One source he focused on were satellites as they collected high-resolution photographs. With satellite images, the size of the data depends heavily on the resolution of the camera. Tukey projected that the size of data will only continue to increase as new electronics are invented providing new ways of collecting data at higher resolutions.

Well, it turns out that Tukey's prediction was indeed correct. The financial world has evolved with the creation of new markets, securities, and derivatives. Pricing a stock depends on more than past prices and economic indicators. Its value depends on its bonds, its derivatives, and related stocks with their bonds and derivatives. The number of influential factors can increase quickly, making identification of relevant information challenging. In another light, developments in technology are now making the sequencing of human genomes an affordable medical procedure. Considering the number of SNPs in the human genome, these data sets are large [Ueki and Tamiya 2012]. With access to not only a patients medical records but to their genetic information, these data may lead to profound advances in personalized medicine and in the treatment of diseases impervious to current medical procedures. These are some common examples of high dimensional data. See Fan and Li [2006] for more detailed examples.

Today, many fields encounter high-dimensional data. Some examples are high resolution images, high frequency trading, genetics, etc. [Fan and Lv 2009]. With satellite images, each pixel is a new piece of data. With high frequency trading, each trade is a new piece of data. With genetics, every SNP recorded is a new piece of data. These types of data can be of a very high dimension; when considering interactions, the dimension grows immensely. Needless to say, the dimension of data can grow quickly and the demand for quality methodology to handle these kind of data is high.

For the rest of section 1, we will discuss ultra-high dimensional data in generality. In section 2, we provide a general approach to variable selection. Section 3 presents advances in methodology for variable selection with discussion of some specific topics of concern in section 4. Lastly, in section 5, we provide some concluding remarks.

1.2 Defining Ultra-High Dimensional Data

We have talked about some sources of high dimensional data and how they are important. Now we will discuss the features that classify these data as high dimensional and when they are considered ultra-high dimensional. The statistical problem where the sample size n and the dimension p have the relationship $n < p$, is a broad class of problems. Specifying the relationship between n and p may change the entire scenario. When $n < p$, some common modeling approaches are the lasso [Tibshirani 1996], the SCAD [Fan and Li 2001] and the elastic net [Zou and Hastie 2005]. When p grows with n, these methods tend to produce less sensible results [Fan and Lv 2009]. High dimensional data is usually classified as the situation where p tends to infinity as n tends to infinity [Fan and Lv 2009]. However, lately, research has been focused on a specific growth rate of p. Ultra-high dimensional data is when p grows at a non-polynomial rate in n, namely $\log(p) = O(n^{\alpha})$ for $\alpha > 0$ [Fan and Lv 2008] [Fan and Lv 2009]. This large dimension produces particularly difficult obstacles to overcome. Donoho [2000] describes the curse of dimensionality as the illusive ability to search through a high-dimensional space in a methodical way.

Recently, some have been challenging the focus of ultra-high dimensional data regarding the relationship between p and n. One assumption typically made is the assumption of sparsity [Fan and Lv 2009]. This will be discussed in more detail later, but for the sake of conversation, the sparsity assumption simply means only a subset of the p predictors are actually related to the response. This idea introduces a new component, the dimension of the support. Verzelen [2012] warns that many are focusing on the wrong relationship between p and n. He suggests defining ultra-high dimensional data as $k[1 + \log(p/k)] = O(n^{\alpha})$ where k is the number of relevant predictors and $\alpha > 0$. In some example problems, through simulation, he shows that the difficulty of a problem, for a fixed p and n, is when k is far away from both 0 and p. Throughout the remaining sections, we will use the previous

definition of ultra-high dimensional data.

1.3 Difficulties of Ultra-High Dimensional Data

With any field of data analysis, there are many challenges and difficulties to overcome. Simply speaking, in ultra-high dimensional data new problems arise from the sheer number of variables per observation.

One commonly talked about issue is collinearity. With a large number of predictor variables, it is likely to find completely unrelated variables having high sample correlation. For example, by generating an $n \times p$ matrix of standard normal random variables, the expected maximum correlation between any two columns can be quite high as p grows much larger than n [Fan and Lv 2008]. Notice that these are independent random variables and we still expect high sample correlations. Thus it is quite likely to have high collinearity in ultra-high dimensional data.

Computational expense is a major roadblock when using existing methodology with ultra-high dimensional data. For variable selection, many methods require performing pairwise comparisons. As an example, in clustering a common method is conducting a two sample t-test between any two classes for each feature [Fan and Lv 2008] [Fan and Fan 2008b]. If the number of classes is large, this can be an expensive task. Identifying interactions proposes another difficult problem in a similar way. The number of two-way interactions is $O(p^2)$ which increases tremendously fast in p, making the identification of relevant combinations challenging. Wu et al. [2010] and Ueki and Tamiya [2012] focus on identifying interactions in genome studies with ultra-high dimensional data.

As always, overfitting is of great concern. With a high number of predictors, the possibility of overfitting is also high. This is one reason why the sparsity assumption is helpful. By reducing the dimension, overfitting is not as likely. While this is a focus of many papers,

it seems that it continues to be an unsolved problem. Jin [2009] points out that published classification rules perform poorly when applied to a fresh data set, even when trained using out of sample methods. This shows the tendency to overfit in high dimensions. One other related issue, of particular concern when the data are of ultra-high dimension, is that of indentifiability. Within the set of predictors may lie multiple models that are equally valid.

1.4 Blessings of Ultra-High Dimensional Data

We have pointed out that there are many great uses for ultra-high dimensional data and applications for analyses are vast. However, when talking specifically about analyzing the data, we have only mentioned the challenging aspects and all of the difficulties that arise from working with ultra-high dimensional data.

Donoho [2000] says the blessing of dimensionality is the behavior of randomness in high dimensions. Due to the concentration of measure phenomenon in Banach spaces, randomness tends to behave in predictable ways. In these higher dimensions, asymptotics tend to have nice convergence properties. More formally, Hall et al. [2005] present some simulation studies that display these desirable features. Their experiments show that if one were to sample many random vectors of dimension p and project them onto an arbitrary plane, these vectors would tend to lie at the vertices of a simplex. As p tends to infinity, the clustering at the vertices is more profound. This inherent predictability is highly useful in that within a lower dimensional space, we are quite certain where the projected vector would lie.

2 General Approach for Ultra-High Dimensional Data

With ultra-high dimensional data, existing methodology for problems with $n < p$ fail to produce stable and sensible results simply because of the magnitude of p. However, when using a sufficiently smaller subset of the predictors with size $p_1 \ll p$ they become a potential

possibility again.

The general approach to attacking ultra-high dimensional problems is to act under the sparsity assumption. This assumption says that out of all p predictor variables, only a small subset have a relationship with the response variable [Fan and Lv 2009]. Thus, if one can obtain this subset then methods such as the lasso or elastic net may be used to finish the analysis. In this two-stage process, the first stage, subsetting on the original p variables, is known as screening.

Methods like the lasso, elastic net, and SCAD have great properties. However, in a two-stage process, there is a great possibility that after the first stage the set of variables left may not provide a quality model of the response; a case in which no method will perform well. Thus, the screening stage is essential. In this stage, the typical strategy is to perform a univariate analysis between the response and each individual predictor. From these univariate analyses, a statistic describing the strength of the relationship between the response and the predictor is recorded as a ranking. Some focus on correlation based statistics or some fit univariate models for each predictor and a fit diagnostic is used. Finally, a threshold is employed to decide which variables to keep. There are a number of different ideas for how to select a cutoff. Some use a hard threshold, usually a simple function of n, and others use a soft threshold approximated with a further analysis.

This general approach provides a great framework since it is simple and logical. It is clear that if a predictor variable has a direct relationship with the response, then some univariate methods may indeed identify that. However, often times a variable may have a conditional relationship with the response; its relationship with the response may depend on the value of another variable. If this is the case, then a univariate approach will likely fail to identify the relationship. Some have proposed iterative methods that strive to alleviate this issue.

As previously mentioned, sample correlations between predictors can be quite high in

ultra-high dimensional data even when the variables are truly independent. However, this is not only relevant within the predictor variables. It is not unlikely to see high sample correlations between predictors and the response regardless of their true relationship. In this case, univariate methods may fail to distinguish between an important variable and a non-informative variable. This is another issue addressed by some of the iterative extensions.

3 Methodology

First, we will establish standard notation and some assumptions about the data to be used throughout. Let it be noted that notation may be added as necessary later on.

Denote a response vector as Y with dimension $n \times 1$ and each element be referenced as Y_i. Let X represent the matrix of covariate information and have dimension $n \times p$. The element in the i^{th} row and j^{th} column of X will be denoted by $X_{i,j}$, the i^{th} row will be denoted by $X_{i,\cdot}$ and the j^{th} column will be denoted by $X_{\cdot,j}$. For further specificity, given a set of predictors $\mathcal{S} \subset \{1, \ldots, p\}$, let $X_{\cdot,\mathcal{S}}$ be the matrix constructed by the columns $X_{\cdot,j}$ such that $j \in \mathcal{S}$. We assume that the pairs $(Y_i, X_{i,\cdot})_{i=1}^n$ are independent and identically distributed. Also, let it be assumed that the sample standard deviation of each $X_{\cdot,j}$ is 1 and the sample mean of each $X_{\cdot,j}$ is 0.

In light of the commonly used sparsity assumption, let $\mathcal{F} = \{1, \ldots, p\}$ be the set of all predictors and $\mathcal{M}$ denote the true set of important variables with $s = |\mathcal{M}|$. For a standard linear model $Y = X\beta + \epsilon$, this set would look like $\mathcal{M} = \{j : \beta_j \neq 0\}$. We will denote a subset of $\mathcal{F}$ by $\mathcal{C}$ with possible subscripts and/or superscripts.

With any screening method, it is necessary to assign a rank to each of the predictors. Let ω denote the vector of rankings or utilities. Possible accents may be added to denote an estimated ranking.

The integer part of a random variable is extracted with the operator $\lfloor \cdot \rfloor$.

3.1 Sure Independence Screening

Sure Independence Screening (SIS) was developed by Fan and Lv [2008] and is a screening method as suggested by the name. SIS is based on correlation learning utilizing the sample correlation between the response and a given predictor. The goal of SIS is to reduce the dimension p from a huge scale $\exp\{O(n^{\alpha})\}$, for $\alpha > 0$, to a moderate dimension $O(n)$ by a fast and efficient method [Fan and Lv 2008].

In the original paper, they refer to "sure screening" as a property that the probability of all important variables being left in the model after screening converges to one. This is a desirable property in a screening method.

SIS ranks based on the magnitude of the sample correlation between a given predictor and the response. However, since it is assumed that $X_{\cdot,j}$ has sample mean 0 and sample variance 1 for all j, the sample correlation between the j^{th} predictor and the response is equivalent to $X_{\cdot,j}^{\intercal}Y$ scaled by the sample correlation of Y. Since the missing scalar does not depend on j, this provides an equivalent ranking. Taking the absolute value of these inner products we obtain the rankings

$$\omega_j = |X_{\cdot,j}^{\intercal}Y|, \quad 1 \leq j \leq p. \tag{1}$$

Each element in ω is a feature describing how correlated the variable is to the response. These features are calculated independently of the other variables.

Using ω select the variables with the largest ω_j. For some proportion $\gamma \in (0, 1)$ let

$$\mathcal{C}_{\gamma} = \left\{k : \omega_k \in \{\omega_{[p-d+1]}, \ldots, \omega_{[p]}\}, d = \lfloor \gamma n \rfloor\right\} \tag{2}$$

where $[\cdot]$ denotes the order statistics. Thus $\mathcal{C}_{\gamma}$ has cardinality $d < n$. This shrinks the model down to a manageable size. Selecting variables in this way is called SIS.

Screening using SIS is simple and straight forward. A large concern with ultra-high dimensional data is computational expense. However, calculating ω is np complex and finding the largest d elements is $O(dp)$ complex, thus the total complexity is $O(np)$ for a fixed d. This is generally a manageable cost.

SIS uses a hard threshold to determine the set of variables to keep. Since a standard linear model with $p > n$ is unidentifiable, Fan and Lv [2008] suggest $d = n - 1$ or $d = n/\log(n)$. While using a larger d increases the probability of including the true model $\mathcal{M}$ in $\mathcal{C}_\gamma$, the number of false positives is much larger as well. One must keep this trade off in mind.

In the development of screening algorithms for ultra-high dimensional data, SIS was one of the first widely recognized algorithms. Many build off the ideas in SIS incorporating more complex techniques of correlation learning. Direct extensions of SIS are mentioned later in the discussion.

3.2 Iterated Sure Independence Screening

Fan and Lv [2008] point out that there are some flaws with the methodology of SIS. Screening based strictly on correlation will fail to remove unimportant variables that are highly correlated to important variables or to the response. Also, a variable may not be directly correlated with the response, but conditional on a separate variable it is. Thus, focus on sample correlations will not place high importance on said variables. Additionally, the SIS method fails to eliminate collinearity among the variables which are kept.

To remedy these three issues, Iterative Sure Independence Screening (ISIS), an extension of SIS, is proposed by Fan and Lv [2008]. This extension utilizes the well known fact that in a linear model, the residuals are uncorrelated with the predictor variables.

As previously mentioned, screening methods are typically coupled with a standard variable selection method such as the lasso or SCAD. The first step in ISIS is to perform SIS combined

with one of these classical methods. From this step, we obtain a set of selected variables $\mathcal{C}_1 \subset \mathcal{F}$. For each $\ell \geq 1$, fit a linear model with $X_{\cdot,\mathcal{C}_\ell}$ as the design matrix to obtain the vector of residuals denoted by ξ_ℓ. Perform a two-stage screening with SIS using $\mathcal{F} \setminus \bigcup_{k \leq \ell} \mathcal{C}_k$ as the available predictors and ξ_ℓ as the response. This results in a new set $\mathcal{C}_{\ell+1}$. Repeat this process until $|\bigcup_k \mathcal{C}_k| \geq d$ for some d. Note that $\mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset$ for all $k \neq k'$. Then $\bigcup_k \mathcal{C}_k$ is the final set of screened variables.

One benefit of SIS is the low computational expense. With this iterative procedure, each $\mathcal{C}_\ell$ requires using the lasso, SCAD, or some other similar technique, thus the computational cost is not nearly as low and depends heavily on the algorithm used. However, since the variables are screened prior to using these methods, the cost is dramatically reduced when compared to applying them to the full data set.

In this iterative process, the important component is using the residuals as the response in subsequent steps. This strives to alleviate the three issues described above. Due to its success, other methods adopt similar iterative extensions to improve their screening techniques.

3.3 Model-Free Screening Procedure

For variable selection and screening, a common approach is to make a model assumption about the data. Zhu et al. [2011] propose a screening procedure which is essentially model-free. The method uses empirical estimates for conditional densities of the response given the predictors. They state that the method is robust to outliers and heavy-tailed responses. It is proven that the method enjoys the consistency in ranking (CIR) property, a similar property to the sure screening property. CIR says that the employed utility function will rank an unimportant predictor lower than an important predictor with probability tending to one.

Define $\Omega(y) = \mathbb{E}[X_{1,\cdot}F(Y_1|X_{1,\cdot})] = \text{Cov}(X_{1,\cdot}, \mathbb{1}_{\{Y_1<y\}})$. The second equality holds since $X_{1,j}$ is assumed to have mean zero. Here $\Omega(y)$ is a vector, so let $\Omega_j(y)$ be the j^{th} element. With that, define

$$\omega_j = \mathbb{E}[\Omega_j^2(Y_1)]. \tag{3}$$

This is the proposed value to be used for ranking predictors. To estimate this quantity, the expectations are replaced by empirical expectations to remain model free. The resulting estimator is

$$\tilde{\omega}_j = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{i'=1}^{n} X_{i',j}\mathbb{1}_{\{Y_{i'}<Y_i\}}\right]^2. \tag{4}$$

Zhu et al. [2011] show that scaling $\tilde{\omega}_j$ by the proper factor results in a U-statistic. The final scaled result is

$$\hat{\omega}_j = \frac{n^3}{n(n-1)(n-2)}\tilde{\omega}_j. \tag{5}$$

For screening, rank the predictors in descending order according to $\hat{\omega}$ and select the first d predictors. Proposed are two different techniques for selecting d. The first method is a hard threshold similar to that recommended by Fan and Lv [2008]. The second is to introduce artificial independent predictors. This is based off the ideas from Luo et al. [2004]. One possibility is to generate them from independent standard normal random variables. For example, let Z be an $n \times q$ matrix where $Z_{i,j} \sim \mathcal{N}(0,1)$. Append Z to X and calculate the weights $\hat{\omega}$ for all $p+q$ predictors. Since Z has absolutely nothing to do with the response, take all $X_{\cdot,j}$ such that $\hat{\omega}_j > \max_{1\le k\le q}\hat{\omega}_{p+k}$. However, with this soft thresholding rule, a new problem has been introduced. Namely that of choosing q. Zhu et al. [2011] mention they have had success with $q = p$. No concrete simulation results are provided to support this claim and they suggest it to be a topic of further research.

3.4 Nonparametric Independence Screening

Often times assumptions of normality or assumptions that a linear model is correct are invoked. However, in practice, there is not always evidence to support these assumptions. Thus it is of interest to allow for these assumptions to be broken when screening for important predictors. The original SIS procedure uses marginal correlations between a predictor and the response. This is comparable to fitting a simple linear model. Fan et al. [2011] expand on this idea by using a nonparametric model, adding much flexibility, which they call Nonparametric Independence Screening (NIS). They propose using model diagnostics to rank the predictors. Some mentioned options are magnitude of marginal estimators, nonparametric marginal correlations, and marginal residual sum of squares.

The NIS procedure begins by assuming the response has the form

$$Y_i = g(X_{i,\cdot}) + \epsilon \tag{6}$$

where ϵ has conditional mean zero. To identify the important predictors, solve the marginal least squares problems. For each j, find

$$g_j = \underset{f \in L_2(P)}{\arg\min}\, \mathbb{E}[Y_1 - f(X_{1,j})]^2$$

where P is the joint distribution of $X_{1,\cdot}$ and Y_1 and $L_2(P)$ is the class of square-integrable functions under P.

To estimate g_j, Fan et al. [2011] use a B-spline polynomial basis. The approximation of g_j will take the form

$$g_j(x) = \sum_{\ell=1}^{q} \beta_\ell \phi_{j,\ell}(x) \tag{7}$$

where β_ℓ are estimated using least squares and $\phi_{j,\ell}(x)$ is the value of the ℓ^{th} spline basis

function for predictor j evaluated at x. Now, rank the variables in decreasing order by

$$\hat{\omega}_j = \frac{1}{n}\sum_{i=1}^{n}[\hat{g}_j(X_{i,j})]^2. \tag{8}$$

Compared to the previous methods, NIS is more complex. There are more parameters to tune such as the polynomial degree of the B-spline basis and the number of knots. However, Fan et al. [2011] address these particular parameters and provide some conditions that must hold for the method to achieve desirable properties.

3.5 Screening with Generalized Correlation

With similar aims as NIS, Hall and Miller [2009] propose a screening technique that uses a generalized correlation to rank the individual predictors. The technique was motivated by the fact that an assumption of the response being linear in the predictors will often times overlook many important variables.

First, define a set of real valued functions $\mathcal{G}$. The goal is to maximize the correlation between Y_1 and $g(X_{1,j})$ by searching over all $g \in \mathcal{G}$ for each $j \in \mathcal{F}$. A generalized correlation is defined by

$$\sup_{g\in\mathcal{G}} \frac{\mathrm{Cov}\{g(X_{1,j}), Y_1\}}{\sqrt{\mathrm{Var}\{g(X_{1,j})\}\mathrm{Var}(Y_1)}} \tag{9}$$

with the empirical estimate

$$\sup_{g\in\mathcal{G}} \frac{\sum_i [g(X_{i,j}) - \bar{g}_j][Y_i - \bar{Y}]}{\sqrt{\sum_i [g(X_{i,j})^2 - \bar{g}_j^2]\sum_i [Y_i - \bar{Y}]^2}} \tag{10}$$

where $\bar{g}_j = \frac{1}{n}\sum_i g(X_{i,j})$. However, note that the variance and the estimate of the variance of Y_1, do not depend on j, thus it is equivalent to rank the predictor variables using the

quantity

$$\hat{\omega}_j = \sup_{g \in \mathcal{G}} \frac{\sum_i [g(X_{i,j}) - \bar{g}_j][Y_i - \bar{Y}]}{\sqrt{n \sum_i [g(X_{i,j})^2 - \bar{g}_j^2]}}. \tag{11}$$

While solving for $\hat{\omega}_j$ is a rather difficult task, Hall and Miller [2009] prove, for finite dimensional $\mathcal{G}$, if there exists a $g \in \mathcal{G}$ that obtains the supremum in equation 11, then the least squares solution

$$\arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} [Y_i - g(X_{i,j})]^2 \tag{12}$$

is also a solution to obtain $\hat{\omega}_j$.

If $\mathcal{G}$ is restricted to linear functions, $\hat{\omega}_j$ reduces to the conventional correlations making the rankings equivalent to SIS. Similarly, if a B-spline basis is used, the rankings are similar to that of NIS.

Just as other methods employ an iterative procedure to alleviate certain downfalls of the method, Hall and Miller [2009] propose bootstrapping the rankings ω_j and creating confidence intervals. Based on the width and relative center of the interval, more accurate conclusions can be made about the utility of each predictor.

3.6 Forward Regression in Ultra-High Dimensions

A popular model selection method is best subset regression. However, one downfall of this method is that the computational cost increases combinatorially in p. Thus, one can imagine the difficulty this would propose in an ultra-high dimensional setting. An alternative to best subset selection is forward regression. Wang [2009] proposes using forward regression (FR) in ultra-high dimensional settings as a way to shrink the dimension down to a manageable size. All of the preceding methods, with the exception of ISIS and the other iterative extensions, employ an independent screening process where the utility for a given predictor does not depend on any other predictor. FR deviates from this trend.

Using FR, the result is a sequence of nested models each with one more predictor than the last. These variables are added according to which one will decrease the regression sum of squares the most. Define $\mathcal{C}_0 = \emptyset$. For each $\ell = 1, \ldots, n$ repeat the following steps. With each $j \in \mathcal{F} \setminus \mathcal{C}_{\ell-1}$ fit a model using $\mathcal{C}_{\ell-1} \cup \{j\}$ as the set of predictors. From this model compute the residual sum of squares

$$\omega_{j(\ell)} = Y^\intercal (I_n - \mathbf{H}_{j(\ell-1)}) Y \tag{13}$$

where $\mathbf{H}_{j(\ell-1)} = X_{\cdot,\mathcal{C}_{\ell-1}\cup j}(X^\intercal_{\cdot,\mathcal{C}_{\ell-1}\cup j} X_{\cdot,\mathcal{C}_{\ell-1}\cup j})^{-1} X^\intercal_{\cdot,\mathcal{C}_{\ell-1}\cup j}$. Now, in this case a smaller ω_j is more desirable. Select

$$a_\ell = \underset{j \in \mathcal{F}\setminus\mathcal{C}_{\ell-1}}{\arg\min}\ \omega_{j(\ell)} \tag{14}$$

and create the next set in the solution path $\mathcal{C}_\ell = \mathcal{C}_{\ell-1} \cup \{a_\ell\}$. The final result is $\mathbb{C} = \{\mathcal{C}_\ell : 1 \leq \ell \leq n\}$ where $\mathcal{C}_\ell = \{a_1, \ldots, a_\ell\}$. Note that $|\mathcal{C}_\ell| \leq n$ for all ℓ allows use of classical techniques for the second stage variable selection regardless of which $\mathcal{C}_\ell$ is chosen.

With the entire solution path, one option is to use the final set, $\mathcal{C}_n$, as the model for further variable selection. Another option is to select a set based on some criterion. A possible option would be the BIC criterion proposed by Chen and Chen [2008]. However, Wang [2009] points out that this criterion has only been proven to have the consistency in selection property when $p = O(n^\alpha)$ and in order to prove the property for $\log(p) = O(n^\alpha)$, some extra conditions are required.

For classification problems in high-dimensional data, Hall and Miller [2010] also suggest using a variant of forward selection. They say it aids in separating the predictors into groups depending on if they are important alone, important in combination with others, or not important.

3.7 Methods for Varying Coefficient Models

3.7.1 Conditional Correlation Sure Independent Screening

For varying coefficient models, Liu et al. [2014] develop an extension of SIS called Conditional Correlation Sure Independence Screening (CC-SIS). Consider the model

$$Y(u) = X(u)\beta(u) + \epsilon \tag{15}$$

where u is a univariate unknown variable, $\mathbb{E}[\epsilon|X, u] = 0$ and $\beta(u) = (\beta_1(u), \ldots, \beta_p(u))^\intercal$ is a vector of p unknown smooth functions. Conditional on u, the above equation becomes a traditional linear model. Thus it would be reasonable to consider the conditional correlation between response and predictor

$$\rho(X_{1,j}, Y_1|u) = \frac{\text{Cov}(X_{1,j}, Y_1|u)}{\sqrt{\text{Var}(X_{1,j}|u)\text{Var}(Y_1|u)}}. \tag{16}$$

However, this is still a function of u. Thus, define the utility function for each covariate as

$$\omega_j = \mathbb{E}[\rho^2(X_{1,j}, Y_1|u)]. \tag{17}$$

Before estimating ω_j, an estimator of $\rho(X_{1,j}, Y_1|u)$ is needed, which requires estimation of five conditional moments, $\mathbb{E}[Y_1|u], \mathbb{E}[Y_1^2|u], \mathbb{E}[X_{1,j}|u], \mathbb{E}[X_{1,j}^2|u]$, and $\mathbb{E}[X_{1,j}Y_1|u]$. Each of these moments is assumed to be a non-parametric smooth function in u. To estimate these moments, Liu et al. [2014] suggest using the kernel smoothing method from Fan and Gijbels [1996]. Kernel regression guarantees the estimates of $\text{Var}(Y_1|u)$ and $\text{Var}(X_{1,j}|u)$ are non-negative and it is required to use the same bandwidth when estimating the five moments to ensure the estimate of $\rho(X_{1,j}, Y_1|u)$ is bounded between -1 and 1.

Assuming a random sample $\{(Y_i, X_{i,\cdot}, u_i), i = 1, \ldots, n\}$ from equation 15, CC-SIS uses

Kernel estimation to obtain $\hat{\rho}(x_j, y|u)$, the estimate of the conditional correlation coefficient as a function of u. Using this estimate, average the squared values over the samples of u to obtain the estimate of the utility function

$$\hat{\omega}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}(X_{1,j}, Y_1|u_i)^2. \tag{18}$$

Using these estimated utilities, order the covariates in descending order and select those exceeding a specified threshold.

3.7.2 Varying-coefficient Independence Screening

Varying-coefficient Independence Screening (VIS) is an extension of SIS for varying coefficient models proposed by Song et al. [2012]. Again, consider the model in equation 15. Focus on the marginal nonparametric regression problem

$$\min_{\beta_j(u)\in L_2(P)} \mathbb{E}[Y(u) - X_{\cdot,j}(u)\beta_j(u)]^2 \tag{19}$$

where $L_2(P)$ is the class of square integrable functions under the measure P, the joint distribution of $X(u)$ and $Y(u)$.

With a random sample $\{(Y_i, X_{i,\cdot}, u_i), i = 1, \ldots, n\}$ estimate $\beta_j(u)$ using a linear combination of B-spline basis functions. This is done by solving the least squares problem

$$\hat{\gamma}_j = \underset{\gamma_j\in\mathbb{R}^q}{\arg\min} \sum_i [Y_i(u_i) - \sum_{\ell=1}^{q} X_{i,j}(u_i)\phi_{j,\ell}(u_i)\gamma_{j,\ell}]^2 \tag{20}$$

where $\phi_{j,\ell}$ is the ℓ^{th} B-spline basis function for feature j. Using the estimates $\hat{\gamma}_j$, let $\hat{\beta}_j(u) =$

$\sum_\ell \phi_{j,\ell}(u)\hat{\gamma}_{j,\ell}$ and define the estimated utility function

$$\hat{\omega}_j = \frac{1}{|\mathcal{U}|}\int_{\mathcal{U}} \hat{\beta}_j(u)^2 du. \tag{21}$$

Using the estimated utilities, rank the covariates in descending order and select the top values according to a specified threshold.

4 Discussion

4.1 Extensions of SIS

Originally, SIS had been designed for linear models with a normally distributed response. Many have taken the concepts in SIS and have extended them to accommodate other scenarios. Fan and Song [2010b] extend SIS to use a more general utility for generalized linear models. Instead of ranking based on marginal correlations, they propose using marginal maximum likelihood estimates to order the predictors. Fan et al. [2009b] expand ISIS to a general pseudo-likelihood framework. In addition, they add the possibility of removing predictors as part of the iterative process, thus directly improving on ISIS. For use in Cox's proportional hazard models, Fan et al. [2010] broaden SIS and ISIS to handle censored data.

4.2 Ultra-High Dimensional Screening for Interactions

With ultra-high dimensional data, the task of selecting the best subset of predictors is a daunting one. Beyond deciding between individual predictors comes the problem of selectively choosing interaction terms. In ultra-high dimensional data, this proposes many computational issues given the rapidly growing combinatorial complexity. Wu et al. [2010] have developed a procedure called screen and clean which utilizes the lasso to identify impor-

tant SNPs and possible interactions. With yet another extension of SIS, Ueki and Tamiya [2012] expand the popular framework to identification of gene-gene interactions. Most of the research in ultra-high dimensional problems have focused on screening for individual predictors rather than interactions. The topic of screening for interactions is important and is likely to be a topic of future research.

4.3 Use of GPUs in Ultra-High Dimensional Problems

Many papers cite a difficulty of ultra-high dimensional data to be the computational cost associated with a large p. This is a definite concern when trying to implement a classical method such as best subset regression. However, there is little discussion of the possibility to code the algorithms to be suitable for a graphical processing unit (GPU). When the average consumer CPU has as many as 8-12 active threads and the average consumer GPU has 1000s of active threads, the comparison of parallel computing power has a clear winner. While a GPU suffers when performing one task alone, where the GPU prevails is repeatedly executing the same task. Many algorithms for variable selection have this type of construction. Consider best subset regression as an example. For each set of predictors, a linear model is fit and some model diagnostic is recorded for each set. This is a perfect setup to be executed on a GPU. Each thread can run a different subset with the GPU tackling 1000s simultaneously.

For detection of gene-gene interactions, Ueki and Tamiya [2012] implemented their algorithm in EPISIS, a software program capable of using GPUs. More recently, Dashti et al. [2013] present a brute force k nearest neighbor graph implementation for ultra-high dimensional data which is computed on a GPU. The brute force method has a computational complexity of $O(n^2)$, but with the power of a GPU, this is less of an issue. This helps to demonstrate the potential of these powerful devices. Hopefully in the future the use of GPUs

will be considered more often when designing new methodology for ultra-high dimensional data.

5 Conclusion

Data is being collected in mass quantities and with tremendous resolution. Ultra-high dimensional data combine new challenges with amplified versions of classical difficulties. With the increase in computing power, we have the ability to tackle these problems. Most have adopted the sparsity condition as a means of analyzing this type of data. While there is a multitude of methods for variable selection, the classical methods lose stability and are far too computationally expensive to run naively. However, the general approach to ultra-high dimensional data, that of a two-stage screening process, provides for a simple framework and the ability to recycle classical methods.

A common approach for screening is to use a utility function independent across predictors. This approach has been expanded to handle many types of data and various model assumptions. However, with this technique, high collinearity and conditional relationships among the predictors pose problems. Most independent techniques have been adapted to use iterative processes that aim to alleviate some of these concerns. Some have taken different approaches that do not rely on independent utilities.

While there have been many advances in the analysis of ultra-high dimensional data over the last decade, in the coming years, with increasing technology and advanced methodology, it will be interesting to see the progress.

References

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008. doi: 10.1093/biomet/asn034.

Ali Dashti, Ivan Komarov, and Roshan M DSouza. Efficient computation of k-nearest neighbour graphs for large high-dimensional data sets on gpu clusters. *PLOS ONE*, 8(9):e74113, 2013.

D. L. Donoho. High-dimensional data analysis: the curses and blessings of dimensionality. In *American Mathematical Society Conf. Math Challenges of the 21st Century*, 2000.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1996. ISBN 9780412983214. URL `http://books.google.com/books?id=BM1ckQKCXP8C`.

J. Fan and R. Li. Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *ArXiv Mathematics e-prints*, February 2006.

J. Fan and J. Lv. A Selective Overview of Variable Selection in High Dimensional Feature Space (Invited Review Article). *ArXiv e-prints*, October 2009.

Jianqing Fan and Yingying Fan. High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605–2637, 12 2008b. doi: 10.1214/07-AOS504.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00674.x.

Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 12 2010b. doi: 10.1214/10-AOS798.

Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.*, 10:2013–2038, December 2009b. ISSN 1532-4435.

Jianqing Fan, Yang Feng, and Yichao Wu. *High-dimensional variable selection for Coxs proportional hazards model*, volume Volume 6 of *Collections*, pages 70–86. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2010. doi: 10.1214/10-IMSCOLL606.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011. doi: 10.1198/jasa.2011.tm09779.

Peter Hall and Hugh Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18 (3):533–550, 2009. doi: 10.1198/jcgs.2009.08041.

Peter Hall and Hugh Miller. Sequential, bottom-up variable selection for high-dimensional classification. *Australian and New Zealand Journal of Statistics*, 52(4):403–421, 2010. ISSN 1467-842X. doi: 10.1111/j.1467-842X.2010.00594.x.

Peter Hall, J. S. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00510.x.

Jiashun Jin. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 106(22):8859–8864, 2009. doi: 10.1073/pnas.0903931106.

Jingyuan Liu, Runze Li, and Rongling Wu. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109(505):266–274, 2014. doi: 10.1080/01621459.2013.850086.

Xiaohui Luo, Leonard A. Stefanski, and Dennis D. Boos. Tuning variable selection procedures by adding noise. *Technometrics, May 2006, Vol*, pages 165–175, 2004.

Rei Song, Feng Yi, and Hui Zou. On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 2012. doi: 10.5705/ss.2012.299.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996. ISSN 00359246.

Masao Ueki and Gen Tamiya. Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis. *BMC Bioinformatics*, 2012.

Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012. doi: 10.1214/12-EJS666.

Hansheng Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009. doi: 10.1198/jasa.2008.tm08516.

Jing Wu, Bernie Devlin, Steven Ringquist, Massimo Trucco, and Kathryn Roeder. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology*, 34(3):275–285, 2010. ISSN 1098-2272. doi: 10.1002/gepi.20459.

Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011. doi: doi:10.1198/jasa.2011.tm10563.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x.