43

Features of Big Data and sparsest solution in high confidence set

Jianqing Fan

Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ

This chapter summarizes some of the unique features of Big Data analysis. These features are shared neither by low-dimensional data nor by small samples. Big Data pose new computational challenges and hold great promises for understanding population heterogeneity as in personalized medicine or services. High dimensionality introduces spurious correlations, incidental endogeneity, noise accumulation, and measurement error. These unique features are very distinguished and statistical procedures should be designed with these issues in mind. To illustrate, a method called a sparsest solution in high-confidence set is introduced which is generally applicable to high-dimensional statistical inference. This method, whose properties are briefly examined, is natural as the information about parameters contained in the data is summarized by high-confident sets and the sparsest solution is a way to deal with the noise accumulation issue.

43.1 Introduction

The first decade of this century has seen the explosion of data collection in this age of information and technology. The technological revolution has made information acquisition easy and cheap through automated data collection processes. Massive data and high dimensionality characterize many contemporary statistical problems from biomedical sciences to engineering and social sciences. For example, in disease classification using microarray or proteomics data, tens of thousands of expressions of molecules or proteins are potential predictors; in genome-wide association studies, hundreds of thousands of SNPs are potential covariates; in machine learning, tens of thousands of features are extracted from documents, images and other objects; in spatial-temporal

problems encountered in economics and earth sciences, time series from hundreds or thousands of regions are collected. When interactions are considered, the dimensionality grows even more quickly. Other examples of massive data include high-resolution images, high-frequency financial data, fMRI data, e-commerce data, marketing data, warehouse data, functional and longitudinal data, among others. For an overview, see Hastie et al. (2009) and Bühlmann and van de Geer (2011).

Salient features of Big Data include both large samples and high dimensionality. Furthermore, Big Data are often collected over different platforms or locations. This generates issues with heterogeneity, measurement errors, and experimental variations. The impacts of dimensionality include computational cost, algorithmic stability, spurious correlations, incidental endogeneity, noise accumulations, among others. The aim of this chapter is to introduce and explain some of these concepts and to offer a sparsest solution in high-confident set as a viable solution to high-dimensional statistical inference.

In response to these challenges, many new statistical tools have been developed. These include boosting algorithms (Freund and Schapire, 1997; Bickel et al., 2006), regularization methods (Tibshirani, 1996; Chen et al., 1998; Fan and Li, 2001; Candès and Tao, 2007; Fan and Lv, 2011; Negahban et al., 2012), and screening methods (Fan and Lv, 2008; Hall et al., 2009; Li et al., 2012). According to Bickel (2008), the main goals of high-dimensional inference are to construct as effective a method as possible to predict future observations, to gain insight into the relationship between features and response for scientific purposes, and hopefully, to improve prediction.

As we enter into the Big Data era, an additional goal, thanks to large sample size, is to understand heterogeneity. Big Data allow one to apprehend the statistical properties of small heterogeneous groups, termed "outliers" when sample size is moderate. It also allows us to extract important but weak signals in the presence of large individual variations.

43.2 Heterogeneity

Big Data enhance our ability to find commonalities in a population, even in the presence of large individual variations. An example of this is whether drinking a cup of wine reduces health risks of certain diseases. Population structures can be buried in the presence of large statistical noise in the data. Nevertheless, large sample sizes enable statisticians to mine such hidden structures. What also makes Big Data exciting is that it holds great promises for understanding population heterogeneity and making important discoveries, say about molecular mechanisms involved in diseases that are rare or affecting small populations. An example of this kind is to answer the question why

chemotherapy is helpful for certain populations, while harmful or ineffective for some other populations.

Big Data are often aggregated from different sites and different platforms. Experimental variations need to be accounted for before their full analysis. Big Data can be thought of as a mixture of data arising from many heterogeneous populations. Let k be the number of heterogeneous groups, $\mathbf{X}$ be a collection of high-dimensional covariates, and y be a response. It is reasonable to regard Big Data as random realizations from a mixture of densities, viz.

$$p_1 f_1(y; \boldsymbol{\theta}_1(\mathbf{x})) + \cdots + p_k f_k(y; \boldsymbol{\theta}_k(\mathbf{x})),$$

in which $f_j(y; \boldsymbol{\theta}_j(\mathbf{x}))$ is the conditional density of Y given $\mathbf{X} = \mathbf{x}$ in population $j \in \{1, \ldots, k\}$, and the function $\boldsymbol{\theta}_j(\mathbf{x})$ characterizes the dependence of the distribution on the covariates. Gaussian mixture models are a typical example; see, e.g., Khalili and Chen (2007) or Städler et al. (2010).

When the sample size is moderate, data from small groups with small p_j rarely occur. Should such data be sampled, they are usually regarded as statistical outliers or buried in the larger groups. There are insufficient amounts of data to infer about $\theta_j(\mathbf{x})$. Thanks to Big Data, when n is so large that np_j is also large, there are sufficient amounts of data to infer about commonality $\boldsymbol{\theta}_j(\mathbf{x})$ in this rare subpopulation. In this fashion, Big Data enable us to discover molecular mechanisms or genetic associations in small subpopulations, opening the door to personalized treatments. This holds true also in consumer services where different subgroups demand different specialized services.

The above discussion further suggests that Big Data are paramountly important in understanding population heterogeneity, a goal that would be illusory when the sample size is only moderately large. Big Data provide a way in which heterogeneous subpopulations can be distinguished and personalized treatments can be derived. It is also an important tool for the discovery of weak population structures hidden in large individual variations.

43.3 Computation

Large-scale computation plays a vital role in the analysis of Big Data. High-dimensional optimization is not only expensive but also unstable in computation, in addition to slowness in convergence. Algorithms that involve iterative inversions of large matrices are infeasible due to instability and computational costs. Scalable and stable implementations of high-dimensional statistical procedures must be sought. This relies heavily on statistical intuition, large-scale screening and small-scale optimization. An example is given in Fan et al. (2009).

Large numbers of observations, which can be in the order of tens of thousands or even millions as in genomics, neuro-informatics, marketing, and online

learning studies, also give rise to intensive computation. When the sample size is large, the computation of summary statistics such as correlations among all variables is expensive. Yet statistical methods often involve repeated evaluations of such functions. Parallel computing and other updating techniques are required. Therefore, scalability of techniques to both dimensionality and the number of cases should be borne in mind when developing statistical procedures.

43.4 Spurious correlation

Spurious correlation is a feature of high dimensionality. It refers to variables that are not correlated theoretically but whose sample correlation is high. To illustrate the concept, consider a random sample of size $n = 50$ of p independent standard $\mathcal{N}(0, 1)$ random variables. Thus the population correlation between any two random variables is zero and their corresponding sample correlation should be small. This is indeed the case when the dimension is small in comparison with the sample size. When p is large, however, spurious correlations start to appear. To illustrate this point, let us compute

$$\hat{r} = \max_{j \geq 2} \hat{\text{corr}}(Z_1, Z_j)$$

where $\hat{\text{corr}}(Z_1, Z_j)$ is the sample correlation between variables Z_1 and Z_j. Similarly, we can compute

$$\hat{R} = \max_{|\mathcal{S}|=5} \hat{\text{corr}}(Z_1, \mathbf{Z}_\mathcal{S}), \tag{43.1}$$

which is the maximum multiple correlation between Z_1 and $\mathbf{Z}_\mathcal{S}$ with $1 \notin \mathcal{S}$, namely, the correlation between Z_1 and its best linear predictor using $\mathbf{Z}_\mathcal{S}$. In the implementation, we use the forward selection algorithm to compute $\hat{R}$, which is no larger than $\hat{R}$ but avoids computing all $\binom{p}{5}$ multiple R^2 in (43.1). This experiment is repeated 200 times.

The empirical distributions of $\hat{r}$ and $\hat{R}$ are shown in Figure 43.1. The spurious correlation $\hat{r}$ is centered around .45 for $p = 1000$ and .55 for $p =$ 10,000. The corresponding values are .85 and .91 when the multiple correlation $\hat{R}$ is used. Theoretical results on the order of the spurious correlation $\hat{r}$ are given in Cai and Jiang (2012) and Fan et al. (2012), but the order of $\hat{R}$ remains unknown.

The impact of spurious correlation includes false scientific discoveries and false statistical inferences. In terms of scientific discoveries, Z_1 and $\mathbf{Z}_{\hat{S}}$ are practically indistinguishable when $n = 50$, given that their correlation is around .9 for a set $\hat{S}$ with $|\hat{S}| = 5$. If Z_1 represents the expression of a gene that is responsible for a disease, we can discover five genes $\hat{S}$ that have a similar predictive power even though they are unrelated to the disease. Similarly,

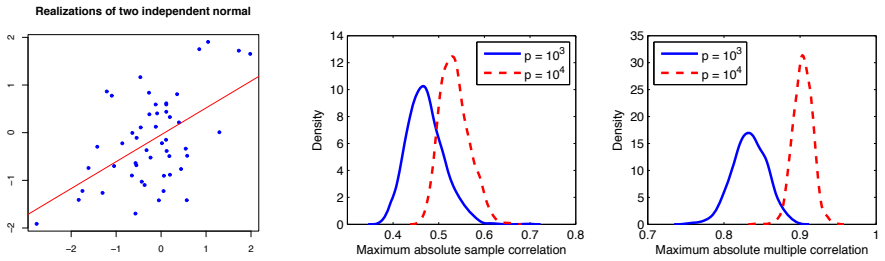

FIGURE 43.1
Illustration of spurious correlation. Left panel: a typical realization of Z_1 with its most spuriously correlated variable ($p = 1000$); middle and right panels: distributions of $\hat{r}$ and $\hat{R}$ for $p = 1000$ and $p = 10{,}000$. The sample size is $n = 50$.

if the genes in $\hat{S}$ are truly responsible for a disease, we may end up wrongly pronouncing Z_1 as the gene that is responsible for the disease.

We now examine the impact of spurious correlation on statistical inference. Consider a linear model

$$Y = \mathbf{X}^\top \beta + \varepsilon, \qquad \sigma^2 = \text{var}(\varepsilon).$$

The residual variance based on a selected set $\hat{\mathcal{S}}$ of variables is

$$\hat{\sigma}^2 = \frac{1}{n - |\hat{\mathcal{S}}|} \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{P}_{\hat{\mathcal{S}}}) \mathbf{Y}, \qquad \mathbf{P}_{\hat{\mathcal{S}}} = \mathbf{X}_{\hat{\mathcal{S}}} (\mathbf{X}_{\hat{\mathcal{S}}}^\top \mathbf{X}_{\hat{\mathcal{S}}})^{-1} \mathbf{X}_{\hat{\mathcal{S}}}^\top.$$

When the variables are not data selected and the model is unbiased, the degree of freedom adjustment makes the residual variance unbiased. However, the situation is completely different when the variables are data selected. For example, when $\beta = 0$, one has $\mathbf{Y} = \epsilon$ and all selected variables are spurious. If the number of selected variables $|\hat{\mathcal{S}}|$ is much smaller than n, then

$$\hat{\sigma}^2 = \frac{1}{n - |\hat{\mathcal{S}}|} (1 - \gamma_n^2) \|\epsilon\|^2 \approx (1 - \gamma_n^2) \sigma^2,$$

where $\gamma_n^2 = \epsilon^\top \mathbf{P}_{\hat{\mathcal{S}}} \epsilon / \|\epsilon\|^2$. Therefore, σ^2 is underestimated by a factor of γ_n^2.

Suppose that we select only one spurious variable. This variable must then be mostly correlated with $\mathbf{Y}$ or, equivalently, ϵ. Because the spurious correlation is high, the bias is large. The two left panels of Figure 43.2 depict the distributions of γ_n along with the associated estimates of $\hat{\sigma}^2$ for different choices of p. Clearly, the bias increases with the dimension, p.

When multiple spurious variables are selected, the biases of residual variance estimation become more pronounced, since the spurious correlation gets larger as demonstrated in Figure 43.1. To illustrate this, consider the linear

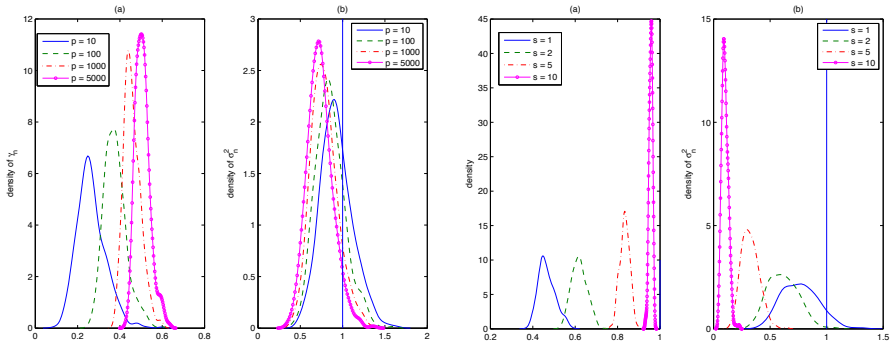

FIGURE 43.2
Distributions of spurious correlations. Left panel: Distributions of γ_n for the null model when $|\hat{\mathcal{S}}| = 1$ and their associated estimates of $\sigma^2 = 1$ for various choices of p. Right panel: Distributions of γ_n for the model $Y = 2X_1 + .3X_2 + \varepsilon$ and their associated estimates of $\sigma^2 = 1$ for various choices of $|\hat{\mathcal{S}}|$ but fixed $p = 1000$. The sample size $n = 50$. Adapted from Fan et al. (2012).

model $\mathbf{Y} = 2\mathbf{X}_1 + .3\mathbf{X}_2 + \epsilon$ and use the stepwise selection method to recruit variables. Again, the spurious variables are selected mainly due to their spurious correlation with ϵ, the unobserved but realized vector of random noises. As shown in the two right panels of Figure 43.2, the spurious correlation is very large and $\hat{\sigma}^2$ gets notably more biased when $|\hat{\mathcal{S}}|$ gets larger.

Underestimation of residual variance leads the statistical inference astray. Variables are declared statistically significant that are not in reality, and this leads to faulty scientific conclusions.

43.5 Incidental endogeneity

High dimensionality also gives rise to incidental endogeneity. Scientists collect covariates that are potentially related to the response. As there are many covariates, some of those variables can be incidentally correlated with the residual noise. This can cause model selection inconsistency and incorrect

selection of genes or SNPs for understanding molecular mechanism or genetic associations.

Let us illustrate this problem using the simple linear model. The idealized model for variable selection is that there is a small subset $\mathcal{S}_0$ of variables that explains a large portion of the variation in the response Y, viz.

$$Y = \mathbf{X}^\top \beta_0 + \varepsilon, \quad \mathrm{E}(\varepsilon \mathbf{X}) = 0, \tag{43.2}$$

in which the true parameter vector β_0 has support $\mathcal{S}_0$. The goal of variable selection is to find the set $\mathcal{S}_0$ and estimate the regression coefficients β_0.

To be more concrete, let us assume that the data generating process is $Y = X_1 + X_2 + \varepsilon$, so that $\mathcal{S}_0 = \{1, 2\}$. As we do not know which variables are related to Y in the joint model, we collect as many covariates as possible that we deem to be potentially related to Y, in the hope of including all members in $\mathcal{S}_0$. Some of those X_j are incidentally correlated with $Y - X_1 - X_2$ or ε. This makes model (43.2) invalid. The rise of incidental endogeneity is due to high dimensionality, making the specifications $\mathrm{E}(\varepsilon \mathbf{X}) = 0$ invalid for some collected covariates, unintentionally. The more covariates are collected, the more unlikely this assumption is.

Does incidental endogeneity arise in practice? Can the exogeneity assumption $\mathrm{E}(\varepsilon \mathbf{X}) = 0$ be validated? After data collection, variable selection techniques such as the Lasso (Tibshirani, 1996; Chen et al., 1998) and folded concave penalized least squares (Fan and Li, 2001; Zou and Li, 2008) are frequently used before drawing conclusions. The model is rarely validated. Indeed, the residuals were computed based only on a small set of the selected variables. Unlike with ordinary least squares, the exogeneity assumption in (43.2) cannot be validated empirically because most variables are not used to compute the residuals. We now illustrate this fact with an example.

Consider the gene expressions of 90 western Europeans from the international "HapMap" project (Thorisson et al., 2005); these data are available on `ftp://ftp.sanger.ac.uk/pub/genevar/`. The normalized gene expression data were generated with an Illumina Sentrix Human-6 Expression Bead Chip (Stranger et al., 2007). We took the gene expressions of *CHRNA6*, cholinergic receptor, nicotinic, alpha 6, as the response variable, and the remaining expression profiles of 47,292 as covariates. The left panel of Figure 43.3 presents the correlation between the response variable and its associated covariates.

Lasso is then employed to find the genes that are associated with the response. It selects 23 genes. The residuals $\hat{\varepsilon}$ are computed, which are based on those genes. The right panel of Figure 43.3 displays the distribution of the sample correlations between the covariates and the residuals. Clearly, many of them are far from zero, which is an indication that the exogeneity assumption in (43.2) cannot be validated. That is, incidental endogeneity is likely present. What is the consequence of this endogeneity? Fan and Liao (2012) show that this causes model selection inconsistency.

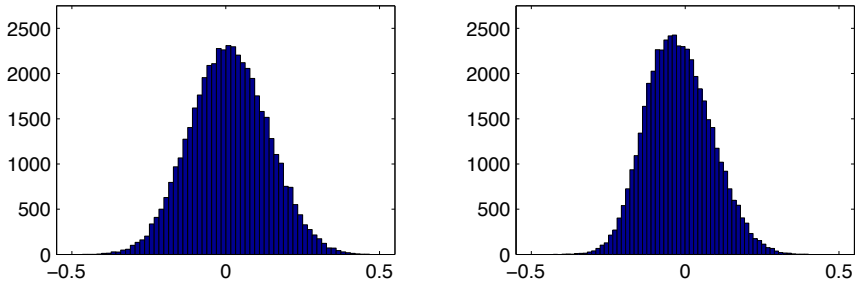

FIGURE 43.3
Distributions of sample correlations. Left panel: Distributions of the sample correlation $\hat{\text{corr}}(X_j, Y)$ ($j = 1, \ldots,$ 47,292). Right panel: Distribution of the sample correlation $\hat{\text{corr}}(X_j, \hat{\varepsilon})$, in which $\hat{\varepsilon}$ represents the residuals after the Lasso fit.

How do we deal with endogeneity? Ideally, we hope to be able to select consistently $\mathcal{S}_0$ under only the assumption that

$$Y = \mathbf{X}_{\mathcal{S}_0}^\top \beta_{\mathcal{S}_0,0} + \varepsilon, \quad \mathrm{E}(\varepsilon \mathbf{X}_{\mathcal{S}_0}) = 0,$$

but this assumption is too weak to recover the set $\mathcal{S}_0$. A stronger assumption is

$$Y = \mathbf{X}_{\mathcal{S}_0}^\top \beta_{\mathcal{S}_0,0} + \varepsilon, \quad \mathrm{E}(\varepsilon | \mathbf{X}_{\mathcal{S}_0}) = 0. \tag{43.3}$$

Fan and Liao (2012) use over identification conditions such as

$$\mathrm{E}(\varepsilon \mathbf{X}_{\mathcal{S}_0}) = 0 \quad \text{and} \quad \mathrm{E}(\varepsilon \mathbf{X}_{\mathcal{S}_0}^2) = 0 \tag{43.4}$$

to distinguish endogenous and exogenous variables, which are weaker than the condition in (43.3). They introduce the Focused Generalized Method of Moments (FGMM) which uses the over identification conditions to select consistently the set of variables $\mathcal{S}_0$. The readers can refer to their paper for technical details. The left panel of Figure 43.4 shows the distribution of the correlations between the covariates and the residuals after the FGMM fit. Many of the correlations are still non-zero, but this is fine, as we assume only (43.4) and merely need to validate this assumption empirically. For this data set, FGMM

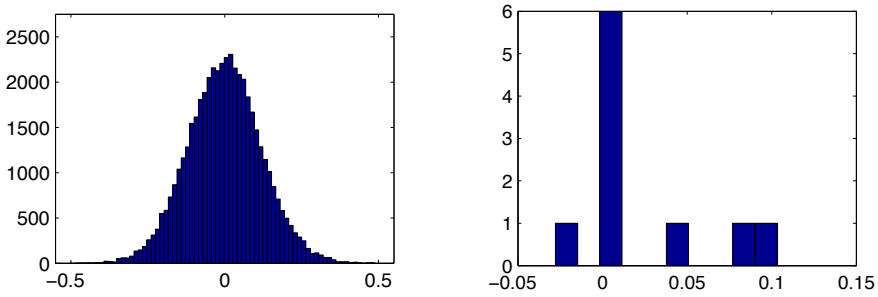

FIGURE 43.4
Left panel: Distribution of the sample correlation $\hat{\text{corr}}(X_j, \hat{\varepsilon})$, in which $\hat{\varepsilon}$ represents the residuals after the FGMM fit. Right panel: Distribution of the sample correlation $\hat{\text{corr}}(X_j, \hat{\varepsilon})$ for only selected 5 genes by FGMM.

selects five genes. Therefore, we need only validate 10 empirical correlations specified by conditions (43.4). The empirical correlations between the residuals after the FGMM fit and the five selected covariates are zero, and their correlations with squared covariates are small. The results are displayed in the right panel of Figure 43.4. Therefore, our model assumptions and model diagnostics are consistent.

43.6 Noise accumulation

When a method depends on the estimation of many parameters, the estimation errors can accumulate. For high-dimensional statistics, noise accumulation is more severe and can even dominate the underlying signals. Consider, for example, a linear classification which assigns the class label $\mathbf{1}(\mathbf{x}^\top \beta > 0)$ for each new data point $\mathbf{x}$. This rule can have high discrimination power when β is known. However, when an estimator $\hat{\beta}$ is used instead, the classification rule can be as bad as a random guess due to the accumulation of errors in estimating the high-dimensional vector $\hat{\beta}$.

As an illustration, we simulate n data points respectively from the population $\mathcal{N}(\boldsymbol{\mu}_0, \mathbf{I}_p)$ and $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_p)$, in which $p = 4500$, $\boldsymbol{\mu}_0 = 0$, and $\boldsymbol{\mu}_1$ is a realization of a mixture of point mass 0 with probability .98 and the standard double exponential distribution with probability .02. Therefore, most components have no discriminative power, yet some components are very powerful in classification. Indeed, among 2% or 90 realizations from the double exponential distributions, several components are very large, and many components are small.

The distance-based classifier, which classifies $\mathbf{x}$ to class 1 when

$$\|\mathbf{x} - \boldsymbol{\mu}_1\|^2 \le \|\mathbf{x} - \boldsymbol{\mu}_0\|^2 \quad \text{or} \quad \beta^\top(\mathbf{x} - \boldsymbol{\mu}) \ge 0,$$

where $\beta = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)/2$. Letting Φ denote the cumulative distribution function of a standard Normal random variable, we find that the misclassification rate is $\Phi(-\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|/2)$, which is effectively zero because by the Law of Large Numbers,

$$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\| \approx \sqrt{4500 \times .02 \times 1} \approx 9.48.$$

However, when β is estimated by the sample mean, the resulting classification rule behaves like a random guess due to the accumulation of noise.

To help the intuition, we drew $n = 100$ data points from each class and selected the best m features from the p-dimensional space, according to the absolute values of the components of $\boldsymbol{\mu}_1$; this is an infeasible procedure, but can be well estimated when m is small (Fan and Fan, 2008). We then projected the m-dimensional data on their first two principal components. Figure 43.5 presents their projections for various values of m. Clearly, when $m = 2$, these two projections have high discriminative power. They still do when $m = 100$, as there are noise accumulations and also signal accumulations too. There are about 90 non-vanishing signals, though some are very small; the expected values of those are approximately 9.48 as noted above. When $m = 500$ or 4500, these two projections have no discriminative power at all due to noise accumulation. See also Hall et al. (2005) for a geometric representation of high dimension and low sample size data for further intuition.

43.7 Sparsest solution in high confidence set

To attenuate the noise accumulation issue, we frequently impose the sparsity on the underlying parameter β_0. At the same time, the information on β_0 contained in the data is through statistical modeling. The latter is summarized by confidence sets of β_0 in $\mathbb{R}^p$. Combining these two pieces of information, a general solution to high-dimensional statistics is naturally the sparsest solution in high-confidence set.

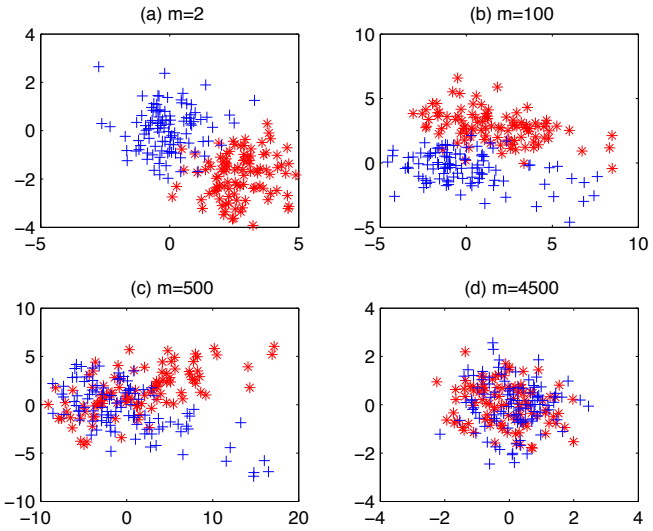

FIGURE 43.5
Scatter plot of projections of observed data ($n = 100$ from each class) onto the first two principal components of the m-dimensional selected feature space.

43.7.1 A general setup

We now elaborate the idea. Assume that the Big Data are collected in the form $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, which can be regarded as a random sample from the population $(\mathbf{X}, Y)$. We wish to find an estimate of the sparse vector $\beta_0 \in \mathbb{R}^p$ such that it minimizes $L(\beta) = \mathrm{E}\{L(\mathbf{X}^\top \beta, Y)\}$, in which the loss function is assumed convex in the first argument so that $L(\beta)$ is convex. The setup encompasses the generalized linear models (McCullagh and Nelder, 1989) with $L(\theta, y) = b(\theta) - \theta y$ under the canonical link where $b(\theta)$ is a model-dependent convex function, robust regression with $L(\theta, y) = |y - \theta|$, the hinge loss $L(\theta, y) = (1 - \theta y)_+$ in the support vector machine (Vapnik, 1999) and exponential loss $L(\theta, y) = \exp(-\theta y)$ in AdaBoost (Freund and Schapire, 1997; Breiman, 1998) in classification in which y takes values ± 1, among others. Let

$$L_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{X}_i^\top \beta, Y_i)$$

be the empirical loss and $L_n'(\beta)$ be its gradient. Given that $L'(\beta_0) = 0$, a natural confidence set is of form

$$\mathcal{C}_n = \{\beta \in \mathbb{R}^p : \|L_n'(\beta)\|_\infty \le \gamma_n\}$$

for some given γ_n that is related to the confidence level. Here $L_n'(\beta) = 0$ can be regarded as the estimation equations. Sometimes, it is handy to construct the confidence sets directly from the estimation equations.

In principle, any norm can be used in constructing confidence set. However, we take the L_∞-norm as it is the conjugate norm to the L_1-norm in Hölder's inequality. It also makes the set $\mathcal{C}_n$ convex, because $|L_n'(\beta)|$ is nondecreasing in each argument. The tuning parameter γ_n is chosen so that the set $\mathcal{C}_n$ has confidence level $1 - \delta_n$, viz.

$$\Pr(\beta_0 \in \mathcal{C}_n) = \Pr\{\|L_n'(\beta_0)\|_\infty \le \gamma_n\} \ge 1 - \delta_n. \tag{43.5}$$

The confidence region $\mathcal{C}_n$ is called a high confidence set because $\delta_n \to 0$ and can even be zero. Note that the confidence set is the interface between the data and parameters; it should be applicable to all statistical problems, including those with measurement errors.

The set $\mathcal{C}_n$ is the summary of the data information about β_0. If in addition we assume that β_0 is sparse, then a natural solution is the intersection of these two pieces of information, namely, finding the sparsest solution in the high-confidence region, viz.

$$\min_{\beta \in \mathcal{C}_n} \|\beta\|_1 = \min_{\|L_n'(\beta)\|_\infty \le \gamma_n} \|\beta\|_1. \tag{43.6}$$

This is a convex optimization problem. Here, the sparsity is measured by the L_1-norm, but it can also be measured by other norms such as the weighted L_1-norm (Zou and Li, 2008). The idea is related to that in Negahban et al. (2012), where a nice framework for analysis of high-dimensional M-estimators with decomposable regularizers is established for restricted convex losses.

43.7.2 Examples

The Danzig selector (Candès and Tao, 2007) is a specific case of problem (43.6) in which the loss is quadratic $L(x, y) = (x - y)^2$ and $\delta_n = 0$. This provides an alternative view to the Danzig selector. If $L(x, y) = \rho(|x - y|)$ for a convex function ρ, then the confidence set implied by the data is

$$\mathcal{C}_n = \{\beta \in \mathbb{R}^p : \|\rho'(|\mathbf{Y} - \mathbf{X}\beta|)\mathbf{X}^\top \operatorname{svn}(\mathbf{Y} - \mathbf{X}\beta)\|_\infty \le \gamma_n\}$$

and the sparsest solution in the high confidence set is now given by

$$\min \|\beta\|_1, \quad \text{subject to } \|\rho'(|\mathbf{Y} - \mathbf{X}\beta|)\mathbf{X}^\top \operatorname{svn}(\mathbf{Y} - \mathbf{X}\beta)\|_\infty \le \gamma_n.$$

In particular, when $\rho(\theta) = \theta$ and $\rho(\theta) = \theta^2/2$, they correspond to the L_1-loss and L_2-loss (the Danzig selector).

Similarly, in construction of sparse precision $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ for the Gaussian graphic model, if $L(\Theta, \mathbf{S}_n) = \|\Theta\mathbf{S}_n - \mathbf{I}_p\|_F^2$ where $\mathbf{S}_n$ is the sample covariance matrix and $\|\cdot\|_F$ is the Frobenius norm, then the high confidence set provided by the data is

$$\mathcal{C}_n = \{\boldsymbol{\Theta} : \|\mathbf{S}_n \cdot (\boldsymbol{\Theta}\mathbf{S}_n - \mathbf{I}_p)\|_\infty \le \gamma_n\},$$

where $\cdot$ denotes the componentwise product (a factor 2 of off-diagonal elements is ignored). If we construct the high-confidence set based directly on the estimation equations $L_n'(\boldsymbol{\Theta}) = \boldsymbol{\Theta}\mathbf{S}_n - \mathbf{I}_p$, then the sparse high-confidence set becomes

$$\min_{\|\boldsymbol{\Theta}\mathbf{S}_n - \mathbf{I}_p\|_\infty \le \gamma_n} \|\mathrm{vec}(\boldsymbol{\Theta})\|_1.$$

If the matrix L_1-norm is used in (43.6) to measure the sparsity, then the resulting estimator is the CLIME estimator of Cai et al. (2011), viz.

$$\min_{\|\boldsymbol{\Theta}\mathbf{S}_n - \mathbf{I}_p\|_\infty \le \gamma_n} \|\boldsymbol{\Theta}\|_1.$$

If we use the Gaussian log-likelihood, viz.

$$L_n(\boldsymbol{\Theta}) = -\ln(|\boldsymbol{\Theta}|) + \mathrm{tr}(\boldsymbol{\Theta}\mathbf{S}_n),$$

then $L_n'(\boldsymbol{\Theta}) = -\boldsymbol{\Theta}^{-1} + \mathbf{S}_n$ and $\mathcal{C}_n = \{\|\boldsymbol{\Theta}^{-1} - \mathbf{S}_n\|_\infty \le \gamma_n\}$. The sparsest solution is then given by

$$\min_{\|\boldsymbol{\Theta}^{-1} - \mathbf{S}_n\|_\infty \le \gamma_n} \|\boldsymbol{\Theta}\|_1.$$

If the relative norm $\|\mathbf{A}\|_\infty = \|\boldsymbol{\Theta}^{1/2}\mathbf{A}\boldsymbol{\Theta}^{1/2}\|_\infty$ is used, the solution can be more symmetrically written as

$$\min_{\|\boldsymbol{\Theta}^{1/2}\mathbf{S}_n\boldsymbol{\Theta}^{1/2} - \mathbf{I}_p\|_\infty \le \gamma_n} \|\boldsymbol{\Theta}\|_1.$$

In the construction of the sparse linear discriminant analysis from two Normal distributions $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, the Fisher classifier is linear and of the form $\mathbf{1}\{\beta^\top(\mathbf{X} - \boldsymbol{\mu}) > 0\}$, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)/2$, $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, and $\beta = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ can easily be estimated from the sample. The question is how to estimate β, which is assumed to be sparse. One direct way to construct confidence interval is to base directly the estimation equations $L_n'(\beta) = \mathbf{S}_n\beta - \hat{\boldsymbol{\delta}}$, where $\mathbf{S}_n$ is the pooled sample covariance and $\hat{\boldsymbol{\delta}}$ is the difference of the two sample means. The high-confidence set is then

$$\mathcal{C}_n = \{\beta : \|\mathbf{S}_n\beta - \hat{\boldsymbol{\delta}}\|_\infty \le \gamma_n\}. \tag{43.7}$$

Again, this is a set implied by data with high confidence. The sparsest solution is the linear programming discriminant rule by Cai et al. (2011).

The above method of constructing confidence is neither unique nor the smallest. Observe that (through personal communication with Dr Emre Barut)

$$\|\mathbf{S}_n\beta - \hat{\boldsymbol{\delta}}\|_\infty = \|(\mathbf{S}_n - \boldsymbol{\Sigma})\beta + \boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|_\infty \le \|(\mathbf{S}_n - \boldsymbol{\Sigma})\|_\infty\|\beta\|_1 + \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|_\infty.$$

Therefore, a high confidence set can be taken as

$$\mathcal{C}_n = \{\|\mathbf{S}_n\beta - \hat{\boldsymbol{\delta}}\|_\infty \le \gamma_{n,1}\|\beta\|_1 + \gamma_{n,2}\}, \tag{43.8}$$

where $\gamma_{n,1}$ and $\gamma_{n,2}$ are the high confident upper bound of $\|(\mathbf{S}_n - \boldsymbol{\Sigma})\|_\infty$ and $\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|_\infty$. The set (43.8) is smaller than the set (43.7), since a further bound $\|\beta\|_1$ in (43.8) by a constant $\gamma_{n,3}$ yields (43.7).

43.7.3 Properties

Let $\hat{\beta}$ be a solution to (43.6) and $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}} - \beta_0$. As in the Danzig selection, the feasibility of β_0 implied by (43.5) entails that

$$\|\beta_0\|_1 \geq \|\hat{\boldsymbol{\beta}}\|_1 = \|\beta_0 + \hat{\boldsymbol{\Delta}}\|_1. \tag{43.9}$$

Letting $\mathcal{S}_0 = \text{supp}(\beta_0)$, we have

$$\|\beta_0 + \hat{\boldsymbol{\Delta}}\|_1 \quad = \quad \|(\beta_0 + \hat{\boldsymbol{\Delta}})_{\mathcal{S}_0}\|_1 + \|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_0^c}\|_1 \geq \|\beta_0\|_1 - \|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_1 + \|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_0^c}\|_1.$$

This together with (43.9) yields

$$\|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_1 \geq \|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_0^c}\|_1, \tag{43.10}$$

i.e., $\hat{\boldsymbol{\Delta}}$ is sparse or "restricted". In particular, with $s = |\mathcal{S}_0|$,

$$\|\hat{\boldsymbol{\Delta}}\|_2 \geq \|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_2 \geq \|\hat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_1/\sqrt{s} \geq \|\hat{\boldsymbol{\Delta}}\|_1/(2\sqrt{s}), \tag{43.11}$$

where the last inequality uses (43.10). At the same time, since $\hat{\boldsymbol{\beta}}$ and β_0 are in the feasible set (43.5), we have

$$\|L_n'(\hat{\boldsymbol{\beta}}) - L_n'(\beta_0)\|_\infty \leq 2\gamma_n$$

with probability at least $1 - \delta_n$. By Hölder's inequality, we conclude that

$$|[L_n'(\hat{\boldsymbol{\beta}}) - L_n'(\beta_0)]^\top \hat{\boldsymbol{\Delta}}| \leq 2\gamma_n \|\hat{\boldsymbol{\Delta}}\|_1 \leq 4\sqrt{s}\gamma_n \|\hat{\boldsymbol{\Delta}}\|_2 \tag{43.12}$$

with probability at least $1 - \delta_n$, where the last inequality utilizes (43.11). By using the Taylor's expansion, we can prove the existence of a point β^* on the line segment between β_0 and $\hat{\boldsymbol{\beta}}$ such that $L_n'(\hat{\boldsymbol{\beta}}) - L_n'(\beta_0) = L_n''(\beta^*)\hat{\boldsymbol{\Delta}}$. Therefore,

$$|\hat{\boldsymbol{\Delta}}^\top L_n''(\beta^*)\hat{\boldsymbol{\Delta}}| \leq 4\sqrt{s}\gamma_n \|\hat{\boldsymbol{\Delta}}\|_2.$$

Since $\mathcal{C}_n$ is a convex set, $\beta^* \in \mathcal{C}_n$. If we generalize the restricted eigenvalue condition to the generalized restricted eigenvalue condition, viz.

$$\inf_{\|\boldsymbol{\Delta}_{\mathcal{S}_0}\|_1 \geq \|\boldsymbol{\Delta}_{\mathcal{S}_0^c}\|_1} \inf_{\beta \in \mathcal{C}_n} |\boldsymbol{\Delta}^\top L_n''(\beta)\boldsymbol{\Delta}|/\|\boldsymbol{\Delta}\|_2^2 \geq a, \tag{43.13}$$

then we have

$$\|\hat{\boldsymbol{\Delta}}\|_2 \leq 4a^{-1}\sqrt{s}\gamma_n. \tag{43.14}$$

The inequality (43.14) is a statement on the L_2-convergence of $\hat{\boldsymbol{\beta}}$, with probability at least $1 - \delta_n$. Note that each component of

$$L_n'(\hat{\boldsymbol{\beta}}) - L_n'(\beta_0) = L_n'(\beta_0 + \hat{\boldsymbol{\Delta}}) - L_n'(\beta_0)$$

in (43.12) has the same sign as the corresponding component of $\hat{\boldsymbol{\Delta}}$. Condition (43.13) can also be replaced by the requirement

$$\inf_{\|\boldsymbol{\Delta}_{\mathcal{S}_0}\|_1 \geq \|\boldsymbol{\Delta}_{\mathcal{S}_0^c}\|_1} \left| [L_n'(\beta_0 + \Delta) - L_n'(\beta_0)]^\top \Delta \right| \geq a\|\Delta\|^2.$$

This facilitates the case where L_n'' does not exist and is a specific case of Negahban et al. (2012).

43.8 Conclusion

Big Data arise from many frontiers of scientific research and technological developments. They hold great promise for the discovery of heterogeneity and the search for personalized treatments. They also allow us to find weak patterns in presence of large individual variations.

Salient features of Big Data include experimental variations, computational cost, noise accumulation, spurious correlations, incidental endogeneity, and measurement errors. These issues should be seriously considered in Big Data analysis and in the development of statistical procedures.

As an example, we offered here the sparsest solution in high-confidence sets as a generic solution to high-dimensional statistical inference and we derived a useful mean-square error bound. This method combines naturally two pieces of useful information: the data and the sparsity assumption.

Acknowledgement

This project was supported by the National Institute of General Medical Sciences of the National Institutes of Health through Grants R01–GM072611 and R01–GMR01GM100474. Partial funding in support of this work was also provided by National Science Foundation grant DMS–1206464. The author would like to thank Ahmet Emre Barut, Yuan Liao, and Martin Wainwright for help and discussion related to the preparation of this chapter. The author is also grateful to Christian Genest for many helpful suggestions.

References

Bickel, P.J. (2008). Discussion on the paper "Sure independence screening for ultrahigh dimensional feature space" by Fan and Lv. *Journal of the Royal Statistical Society, Series B*, 70:883–884.

Bickel, P.J., Ritov, Y., and Zakai, A. (2006). Some theory for generalized boosting algorithms. *The Journal of Machine Learning Research*, 7:705–732.

Breiman, L. (1998). Arcing classifier. *The Annals of Statistics*, 26:801–849.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin.

Cai, T. and Jiang, T. (2012). Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis*, 107:24–39.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607.

Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351.

Chen, S.S., Donoho, D.L., and Saunders, M.A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36:2605.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society, Series B*, 74:37–65.

Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, J. and Liao, Y. (2012). Endogeneity in ultrahigh dimension. *Available at SSRN 2045864*.

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

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57:5467–5484.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038.

Freund, Y. and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139.

Hall, P., Marron, J.S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B*, 67:427–444.

Hall, P., Titterington, D.M., and Xue, J.-H. (2009). Tilting methods for assessing the influence of components in a classifier. *Journal of the Royal Statistical Society, Series B*, 71:783–803.

Hastie, T., Tibshirani, R.J., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102:1025–1038.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107:1129–1139.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, London.

Negahban, S.N., Ravikumar, P., Wainwright, M.J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27:538–557.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models (with discussion). *Test*, 19:209–256.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., and Dermitzakis, E.T. (2007). Population genomics of human gene expression. *Nature Genetics*, 39:1217–1224.

Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. (2005). The International HapMap Project Web Site. *Genome Research*, 15:1592–1593.

Tibshirani, R.J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Vapnik, V. (1999). *The Nature of Statistical Learning Theory*. Springer, Berlin.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533.