High Dimensional Data and Statistical Learning

Lecture 1: Introduction

Weixing Song

Department of Statistics
Kansas State University

Outline

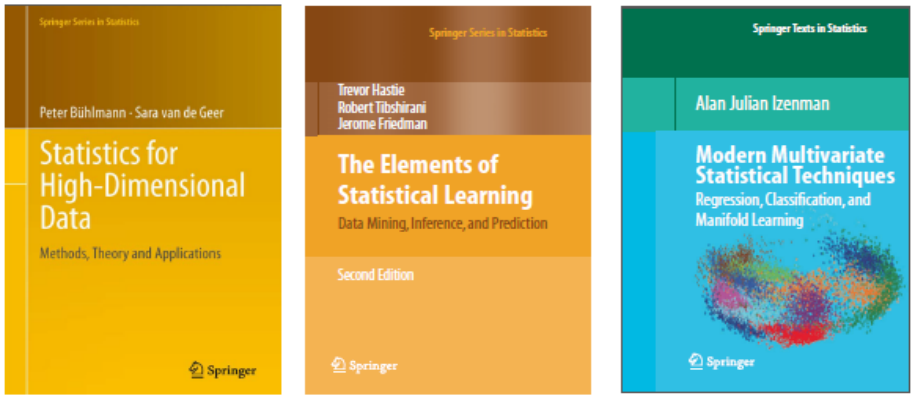
Outline

Course Information

Contact Information: weixing@ksu.edu

Class Place and Time: 302 Dickens Hall, 2:30-3:45, Tuesday, Thursday

References:



- *Statistics for High-Dimensional Data,* by Peter Bühlmann , Sara van de Geer; Springer.
- *The Elements of Statistical Learning,* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman; 2nd Edition, Springer.
  (Book Website: http://statweb.stanford.edu/t̃ibs/ElemStatLearn/)
- *Modern Multivariate Statistics,* by Alan Izenman; Springer.
  (Book Website: http://astro.temple.edu/ alan/MMST/index.html)
- Journal articles.

Special thanks to

Professor Kun Chen

of the University of Connecticut for allowing me to use his lecture notes!

Outline

Introduction

We are entering the era of Big Data — a term that refers to the explosion of available information.

Such a Big Data movement is driven by the fact that massive amounts of very high-dimensional or unstructured data are continuously produced and stored with much cheaper cost than they used to be.

Examples are abundant in genome sequencing, social media analysis, biomedical imaging, high-frequency finance, surveillance videos, retail sales.

The massive amounts of high-dimensional data bring both opportunities and new challenges to data analysis. Valid statistical analysis for Big Data is becoming increasingly important.

In the classical statistical model setting, the number of covariates p is fixed and the sample size n is large.

To be specific, high-dimensional statistics refers to statistical inference when the number of covariates or parameters p is comparable to or much larger than the number of observations n.

- *High dimension:* The dimensionality p grows polynomially with the sample size n, i.e., $p = O(n^\alpha)$ for some $\alpha > 0$.
- *Ultra-high dimension:* The dimensionality p grows non-polynomially with the sample size n, for example, $\log p = O(n^\alpha)$ for some $\alpha > 0$. This is the so-called non-polynomial (NP) dimensionality.

The challenges of analyzing Big Data:

Big data are characterized by high dimensionality and large sample size. These two features raise three unique challenges:

(1). high dimensionality brings noise accumulation, spurious correlations, and incidental homogeneity;

(2). high dimensionality combined with large sample size creates issues such as heavy computational cost and algorithmic instability;

(3). the massive samples in Big Data are typically aggregated from multiple sources at different time points using different technologies. This creates issues of heterogeneity, experimental variations and statistical biases. and requires us to develop more adaptive and robust procedures.

Many traditional methods that perform well for moderate sample size do not scale to massive data. Similarly, many statistical methods that perform well for low-dimensional data are facing significant challenges in analyzing high-dimensional data.

To design effective statistical procedures for exploring and predicting Big Data, we need to address Big Data problems such as heterogeneity, noise accumulation, spurious correlations and incidental heterogeneity, in addition to balancing the statistical accuracy and computational efficiency.

The goals of analyzing Big Data:

(1). to develop effective and robust methods that can accurately predict the future observations;

(2). to gain insight into the relationship between the features and response for scientific purpose;

(3). to understand heterogeneity and commonality across different subpopulation;

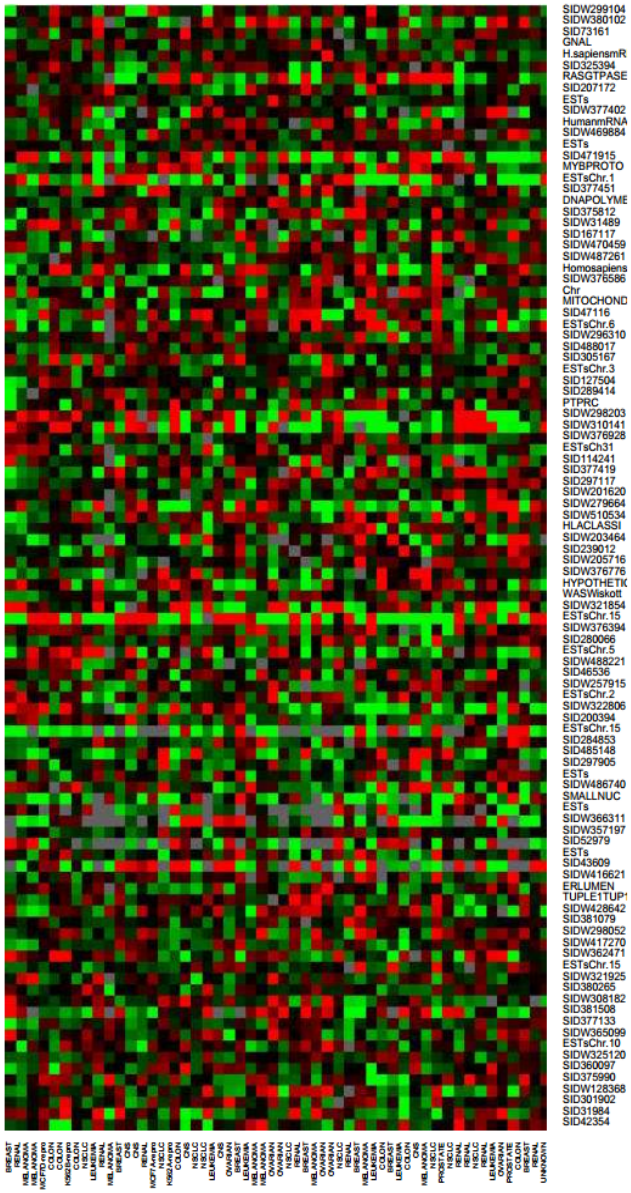
Outline

Examples

Example 1 (DNA Expression Microarrays)

DNA (deoxyribonucleic acid) is the basic material that makes up human chromosomes. DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA (messenger ribonucleic acid) present for that gene.

A gene expression data set collects together the expression values from a series of DNA microarray experiments, with each column representing an experiment. There are therefore several thousand rows representing individual genes and tens of columns representing samples. For example, the figure on the right contains 6830 genes (rows) and 64 samples (columns). For clarity reason, only 100 rows are shown.



Typical questions include:

- which samples are most similar to each other, in terms of their expression profiles across genes?
- which genes are most similar to each other, in terms of their expression profiles across samples?
- do certain genes show very high/low expression for certain cancer samples?

Example 2 (Macy's Inc. and real-time pricing): The retailer adjusts pricing in near-real time for 73 million items, based on demand and inventory, using technology from SAS Institute.

Example 3 (Tipp 24 AG): Tipp24 AG, a platform for placing bets on European lotteries, and prediction. The company uses KXEN software to analyze billions of transactions and hundreds of customer attributes, and to develop predictive models that target customers and personalize marketing messages on the fly. That led to a 90% decrease in the time it took to build predictive models.

Example 4 (Spatial correlation of home price appreciation (HPA)): The consideration of 1000 neighborhoods requires 1 million paramters.

Example 5 (Stock price): Managing 2000 stocks involves over 2 million parameters in the covariance matrix.

Example 6 (Image analysis): High resolution images in signal processing and medical imaging analysis.

Example 7 (Text classification): Text or document classification (email spam, feature extraction via frequency counting).

Outline

Curse of High-Dimensionality

The term "curse of dimensionality" (Bellman, 1961) described how difficult it was to perform high-dimensional numerical integration. This led to the more general use of the term to describe the difficulty of dealing with statistical problems in high dimensions.

- **Sparsity.**

  Suppose we have p input variables. Divide the axis of each of p input variables into k intervals with equal length. Such a partition divides the entire p-dimensional input space into k^p hypercubes. Now we sample uniformly from these hypercubes. Note that in general increasing k reduces the sizes of the hypercubes while increasing the precision of the approximation. If there has to be at least one input point in each hypercube, then the number of such points needed must increase exponentially as p increases. However, in practice, onely a limited number of observations are available, i.e., the data are very sparse in high-dimensional space, and we never have enough data.

Boundary phenomenon.

As the number of dimensions grows larger, almost all the volume inside a hypercubic region of input space lies closer to the boundary or surface of the hypercube rather than near the center.

(i) An p-dimensional hypercube $[0, a]^p$ with each edge of length a has volume a^p.

(ii) Consider a slightly smaller hypercube with each edge of length $a - \varepsilon$, where $\varepsilon > 0$ is small.

(iii) The difference in volume between these tow hypercubes is $a^p - (a - \varepsilon)^p$.

(iv) As $p \to \infty$, the proportion of the volume that is contained between the two hypercubes is

$$\frac{a^p - (a - \varepsilon)^p}{a^p} = 1 - \left(1 - \frac{\varepsilon}{a}\right)^p \to 1.$$

Question: Suppose we have an p-dimensional hypercube $[0, 1]^p$ and observations are uniformly sampled from it. What is the value of ε so that an p-dimensional sub-hypercube $[0, 1 - \varepsilon]^p$ can capture r percent of the data?

Solution: $\varepsilon = 1 - r^{1/p}$.

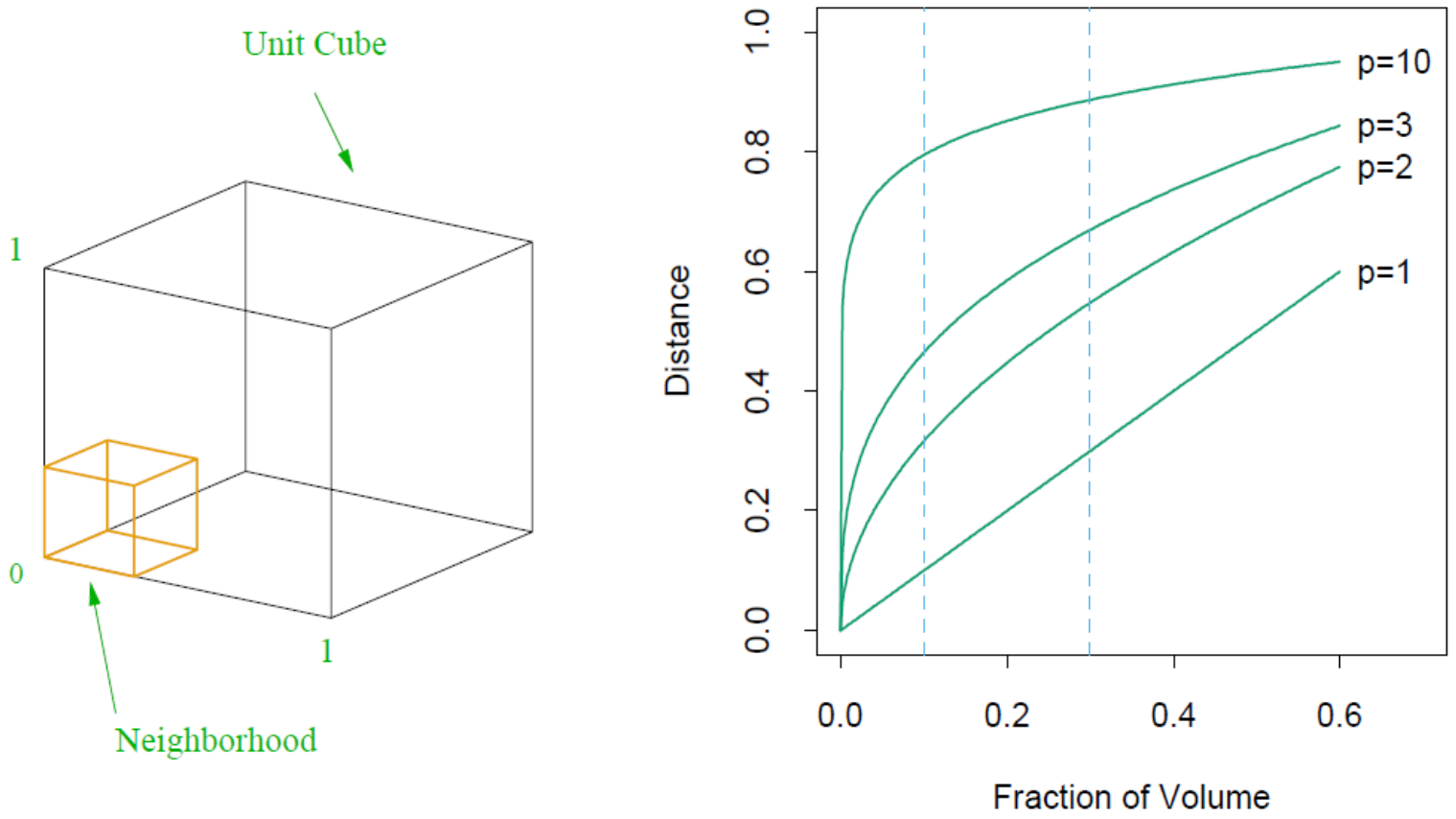


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover* 80% *of the range of each coordinate to capture* 10% *of the data.*

- **Heterogeneity.** Big Data are often created via aggregating many data sources corresponding to different subpopulations. Each subpopulation might exhibit some unique features not shared by others. So heterogeneity is very common in such contexts.

  The following mixture model is often used to describe data with many subpopulations:

  $$\lambda_1 p_1(y; \theta_1(x)) + \cdots \lambda_m p_m(y; \theta_m(x)),$$

  where $\lambda_j \geq 0$ represents the proportion of the j-th subpopulation, $p_j(y, \theta_j(x))$ is the probability distribution of the response of the j-th subpopulation given the covariates x with $\theta_j(x)$ as the parameter vector.

  Inferring the mixture model for large datasets requires sophisticated statistical and computational methods. In low dimensions, standard techniques such as the EM algorithm for finite mixture models can be applied. In high dimensions, we need to carefully regularize the estimating procedure to avoid overfitting or noise accumulation and to devise good computation algorithms.

- **Noise accumulation.**

  Noise accumulation is a common phenomenon in high-dimensional prediction.

  Example 1: Consider a linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{y}$ is an n-vector of response, $\mathbf{X}$ is an $n \times p$ design matrix, β is a p-vector of regression coefficients with the true value β_0 having only s nonzero components, and ε is an n-vector of random error with mean 0 and variance σ^2.

  The model built on all regressors usually has prediction error of order $(1 + p/n)\sigma^2$ rather than $(1 + s/n)\sigma^2$ when $p \leq n$ and there are only s intrinsic predictors.

  The ordinary least squares (OLS) estimator is not well behaved when p is comparable to n and the OLS is not applicable when $p > n$.

Remark: For an arbitrary test point x_0, we have $\hat{y}_0 = x_0^T \hat{\beta}$, which can be written as

$$\hat{y}_0 = x_0^T \beta + x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon.$$

Therefore, the conditional expected prediction error (EPE) given x_0 is

$$\text{EPE}(x_0) = E_{y_0|x_0}(y_0 - \hat{y}_0)^2 = \sigma^2 + \sigma^2 E_{\mathcal{T}} x_0^T (\mathbf{X}^T \mathbf{X})^{-1} x_0 + 0^2.$$

If n is large and $\mathcal{T}$ is selected at random, and assuming $EX = 0$, then $\mathbf{X}^T \mathbf{X} \xrightarrow{a.s} n\text{Cov}(X)$. Therefore, the EPE of $\hat{y}_0$ is

$$E_{x_0} \text{EPE}(x_0) \sim \sigma^2 + \frac{\sigma^2}{n} E_{x_0} x_0^T \text{Cov}^{-1}(X) x_0 = \sigma^2(1 + p/n).$$

Example 2: Consider a classification problem where the data come from two classes:

$$X_1, \ldots, X_n \sim N_d(\mu_1, I_d), \quad Y_1, \ldots, Y_n \sim N_d(\mu_2, I_d).$$

To illustrate the impact of noise accumulation in classification, set $n = 100$, and $d = 1000$. We set $\mu_1 = 0$ and μ_2 to be sparse, i.e., only the first 10 entries of μ_2 are nonzero with value 3, and all other entries are zero.

The first two principal components by using the first $m = 2, 40, 200$ features and the whole 1000 features are plotted in the following figure. As illustrated in these plots, when $m = 2$ we obtain high discriminative power. However, the discriminative power becomes very low when m is too large due to noise accumulation.

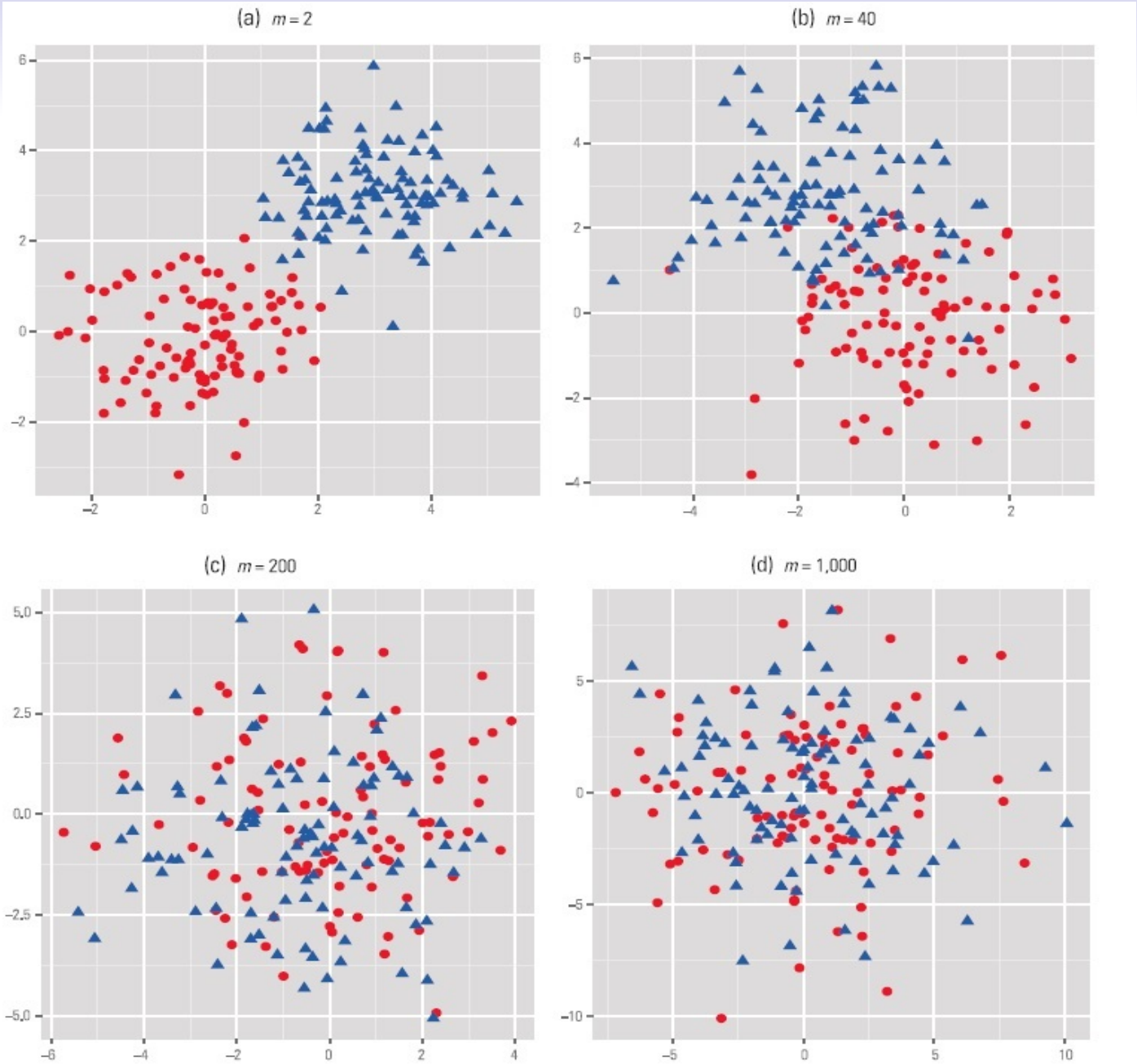


Figure 1. Scatter plots of projections of the observed data ($n = 100$ from each class) onto the first two principal components of the best m-dimensional selected feature space. A projected data with the filled circle indicates the first class and the filled triangle indicates the second class.

- **Spurious correlation.**

  Variables tend to be highly correlated in high dimensions.

  From the geometrical point of view, the correlation among variables increases with dimensionality, more and more variables packed together.

  High collinearity and spurious correlation make high-dimensional variable selection intrinsically difficult:

  - some true or important variables can have a weaker relationship with the response than some noise variables;
  - some noise variables can have a strong relationship with the response.

Example 1: Let $x_1, x_2, \ldots, x_n$ be n independent observations of a d-dimensional Gaussian random vector $X = (X_1, \ldots, X_d)' \sim N_d(0, I_d)$. Repeatedly simulate the dat with $n = 60$ and $d = 800, 6400$ for 1000 times. Consider the empirical distribution of the maximum absolute sample correlation coefficient between the first variable with the remaining ones defined as

$$\hat{r} = \max_{2 \le j \le d} |\widehat{\text{Corr}}(X_1, X_j)|.$$

Furthermore, we compute the maximum absolute multiple correlation between X_1 and linear combinations of several irrelevant spurious variables:

$$\hat{R} = \max_{|S=4|} \max_{\{\beta_j\}_{j=1}^4} |\widehat{\text{Corr}}(X_1, \sum_{j \in S} \beta_j X_j)|,$$

where S is any size of four subset of $\{2, 3, \ldots, d\}$ and β_j is the LS regression coefficient of X_j when regressing X_1 on $\{X_j\}_{j \in S}$.

Example Code:

```
maxcorr=function(n,d)
 {
 X=matrix(rnorm(n*d),nrow=n);
 max(abs(cor(X[,1],X[,-1])))
 }
mcsample1=mcsample2=vector()
for(i in 1:1000)
 {
 mcsample1[i]=maxcorr(60,800);
 mcsample2[i]=maxcorr(60,6400);
 }
lend=min(c(mcsample1,mcsample2));
rend=max(c(mcsample1,mcsample2));
hist(mcsample1,col=rgb(0,0,1,1/4),freq=F,xlim=c(lend,rend),ylim=c(0,13))
hist(mcsample2,col=rgb(1,0,0,1/4),freq=F,add=T)
```

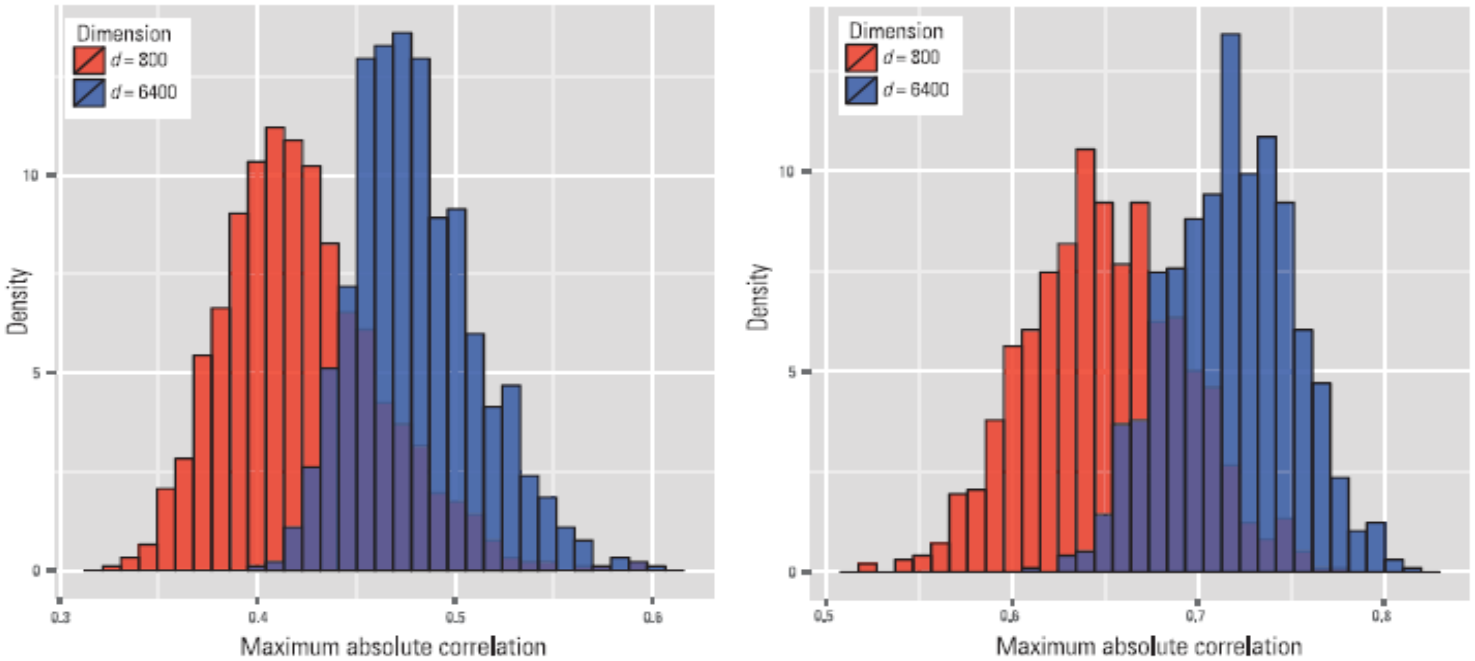


Figure 2. Illustration of spurious correlation. (a) Distribution of the maximum absolute sample correlation coefficients between X_1 and $\{X_j\}_{j\neq 1}$. (b) Distribution of the maximum absolute sample correlation coefficients between X_1 and the closest linear projections of any four members of $\{X_j\}_{j\neq 1}$ to X_1. Here the dimension d is 800 and 6400, the sample size n is 60. The result is based on 1000 simulations.

Even though X_1 is utterly independent of $X_2, \ldots, X_d$, the correlation between X_1 and other variables, or the closest linear combination of any other four variables ca be very high.

Some bad consequences of spurious correlation:

- False discovery.

  Let $\mathbf{X}_S = \{X_j\}_{j \in S}$ be the sub-random vector indexed by S and let $\hat{S}$ be the selected set that has the higher spurious correlation with X_1. Then when d is large, if X_1 represents the expression level of a gene that is responsible for a disease, we cannot distinguish it from the other four genes in $\hat{S}$ that have a similar predictive power although they are scientifically irrelevant.

- Wrong statistical inference.

  Consider a linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, $\mathrm{Var}(\varepsilon) = \sigma^2 I_d$. We would like to estimate the standard error σ of the residual, which is a prominently featured in statistical inferences of regression coefficients, model selection, goodness-of-fit test and marginal regression. Let $\hat{S}$ be a set of selected variables and $P_{\hat{S}}$ be the projection matrix on the column space of $\mathbf{X}_{\hat{S}}$. The standard residual variance estimator, based on the selected variable, is

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T (I_n - P_{\hat{S}})\mathbf{y}}{n - |\hat{S}|}.$$

  The above estimator is unbiased when the variables are not selected by data and the model is correct. However, the situation is completely different when the variables are selected by the data.

Example 2: Using the same set up as in Example 1. Let $Y = X_1$ and we fit a linear regression model using the four selected variables in the set $\hat{S}$, the residual variance is

$$\hat{\sigma}^2 = \frac{RSS}{n - |\hat{S}|} \approx \frac{60(1 - 0.7^2)}{56} = 0.55. \quad (\hat{\sigma} = 0.74).$$

The error standard deviation is deflated by a factor of more than $1/4$.

As a result, in variable selection context, more variables would be declared statistically significant.

- **Incidental endogeneity.**

  In a regression setting $Y = \sum_{j=1}^{d} \beta_j X_j + \varepsilon$, the term "endogeneity" means that some predictors $\{X_j\}$ correlate with the residual noise ε.

  The conventional sparse model assumes

$$Y = \sum_{j=1}^{d} \beta_j X_j + \varepsilon, \quad E(\varepsilon X_j) = 0, j = 1, 2, \ldots, d,$$

  with a small set $S = \{j : \beta_j \neq 0\}$.

  The exogenous assumption in $E(\varepsilon X_j) = 0$ is crucial for validity of most existing statistical procedures, including variable selection consistency.

  This assumption can be easily violated in high dimensions as some variables $\{X_j\}$ are incidentally correlated with ε, making most high-dimensional procedures statistically invalid.

A quote from Fan, et al. (2014):

> *Unlike spurious correlation, incidental endogeneity refers to the genuine existence of correlations between variables unintentionally, both due to high dimensionality.The former is analogous to find two persons look alike but have no genetic relation, whereas the latter is similar to bumping into an acquaintance, both easily occurring in a big city. More generally, endogeneity occurs as a result of selection biases, measurement errors and omitted variables.*

Outline

Blessing of High Dimensionality

A quote from Yi Ma:

> *In computer vision, you are routinely dealing with images or videos that are very high-dimensional. Fortunately, it turns out that when the dimensionality is high enough, if you have the right computation tool, you can harness rich redundancy in the data that gives you a very, very good chance of solving some of the hardest problems in the world. That's why its called the blessing of dimensionality.*

For example, the main advantage brought by Big Data is to understand the heterogeneity of subpopulations, such as the benefits of certain personalized treatments, which are infeasible when sample size is small or moderate.

Big Data allow us to unveil weak commonality across whole population. For example, the benefit of one drink of red wine per night on heart can be difficult to assess without large sample size. Similarly, heath risks to exposure of certain environmental factors can only be more convincingly evaluated when the sample sizes are sufficiently large.

Donoho (2000):

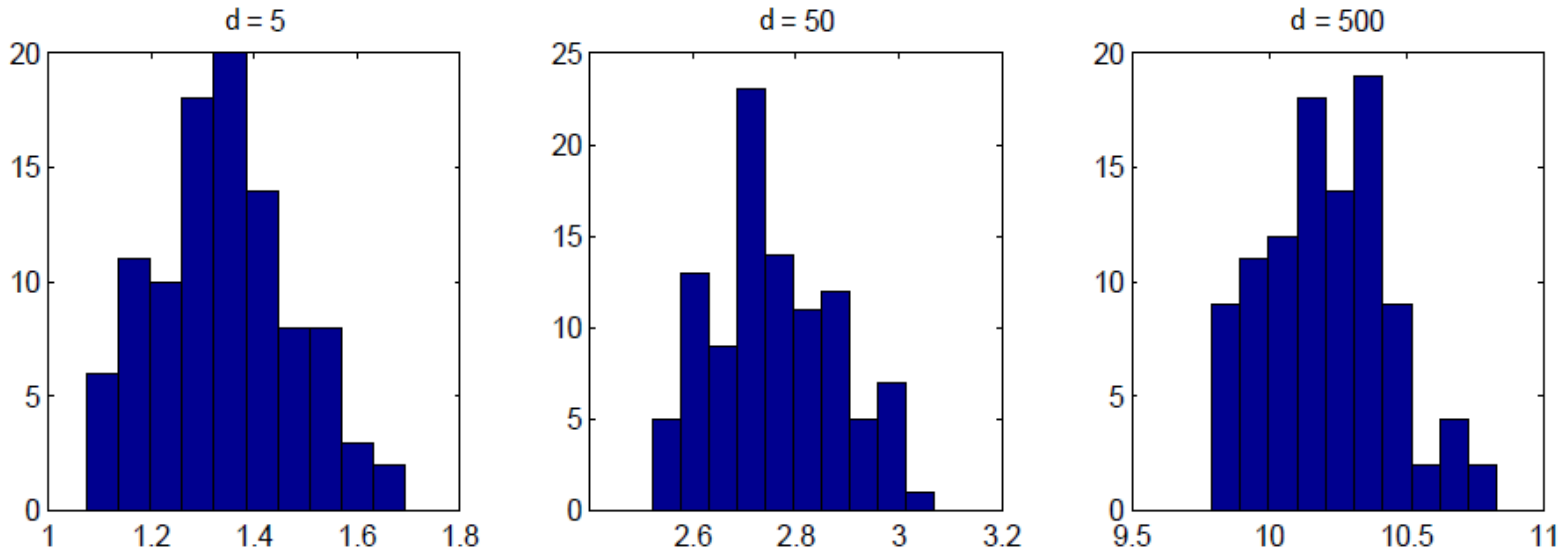
- **Concentration of Measure.** Various large deviation results state that the probability measures often concentrate on small sets, especially in high-dimensions. To try to get a commercial slogan out of it, we could say that in many cases, there are really "few things that matter" and that the function will be constant on most of the space.
- **Dimension Asymptotics.** Many results such as asymptotic distributions already exist in mathematical analysis obtained by letting the number of dimension go to infinity.
- **Approach to Continuum.** Many times high dimension data are collected from objects that are really continuous-space or continuous-time phenomena, e.g., we are sampling a curve or an image. Therefore, there is an underlying compactness to the space of observed data which will be reflected by an increasing simplicity of analysis for large p.

Consider the case of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $\sim N(0, I_d)$ and let $\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ be the sample covariance matrix (without estimating the mean vector).

Question: Does the distribution of the largest eigenvalue $\lambda_{\max}$ of $\hat{\Sigma}$ concentrate around the true value 1 ?

The answer is YES in fixed dimensions, but how about in high dimensions?

The following are histograms of $\lambda_{\max}$ for $n = 100$ and different d.



The histograms show that the distribution shifts further to the right of 1 when d becomes larger.

Gelman (1980)

As $n/d \to \gamma \geq 1$, $\lambda_{\max} \xrightarrow{a.s} (1+\gamma^{-1/2})^2$.

Johnstone (2001)

As $n/d \to \gamma \geq 1$,

$$\frac{n\lambda_{\max} - \mu_{nd}}{\sigma_{nd}} \xrightarrow{\mathscr{L}} F_1 \quad \text{(Tracy-Widom distribution of order 1)}$$

where $\mu_{nd} = (\sqrt{n-1}+\sqrt{d})^2$ and $\sigma_{nd} = \sqrt{\mu_{nd}}(1/\sqrt{n-1}+1/\sqrt{d})^{1/3}$.

Principal component analysis can fail in high dimensions. The estimated eigenvectors of leading eigenvalues can have a positive angle with the true directions asymptotically.

Outline

General Thinking on Dealing with High Dimensionality

An optimal statistical model is characterized by at least the following three fundamental attributes:

- Parsimony (model simplicity)
- Goodness-of-fit (conformity of the fitted model to the data at hand)
- Generalizability (applicability of the fitted model to describe or predict new data)

Occam's Razor: Laws of Parsimony.

Occam's razor is a philosophical principle credited to the medieval English philosopher and Franciscan monk William of Ockham (1285-1349).

Principle of Occam's razor:

> *Plurality should not be posited without necessity.*
>
> *Entities should not be multiplied beyond necessity.*

Occam's Razor recommends that we "shave of" extraneous ideas to better reveal the truth.

A Video on Occam's Razor: www.youtube.com/watch?v=9XEA3k_QIKo

Quote from Thomas Aquinas (1225-1274):

> *If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices.*

Quote from Isaac Newton (1643-1727):

> *We are to admit no more causes of natural things than such are both true and sufficient to explain their appearances.*

Quote from Albert Einstein (1879-1955):

> *Everything should be made as simple as possible, but not simpler.*

A key idea in high-dimensional modeling is **Sparsity**.

It is often desirable to build more interpretable models involving fewer variables.

Though sparsity is considered a prior belief, it is often a reasonable one and it can work beautifully in many applications.

An example: Several methods have been developed to fit linear regression models with thousands of predictors within seconds, which can work as if the true sparse model were known in advance.

Sparsity should be understood more widely in transformed or enlarged feature spaces: some grouping or transformation of the input variables; enlarge the feature space by adding interactions and higher order terms to reduce the model bias.

Sparsity can also be viewed in the context of dimensionality reduction by introducing a sparse representation, for example

- Fama-French three-factor or five factor models
- Use the factor model to reduce the number of effective parameters in high-dimensional covariance matrix estimation for portfolio selection
- Multivariate reduced-rank estimation
- Matrix completion (Netflix million dollar challenge)

Fama-French Three-factor Model

In asset pricing and portfolio management, the "FamaâĂŞFrench three-factor model" is a model designed by Eugene Fama and Kenneth French to describe stock returns. Fama and French were professors at the University of Chicago Booth School of Business.

The traditional asset pricing model, known formally as the "capital asset pricing model" (CAPM) uses only one variable to describe the returns of a portfolio or stock with the returns of the market as a whole. In contrast, the FamaâĂŞFrench model uses three variables. Fama and French started with the observation that two classes of stocks have tended to do better than the market as a whole: (i) small caps and (ii) stocks with a low Price-to-Book ratio (P/B, customarily called value stocks, contrasted with growth stocks). They then added two factors to CAPM to reflect a portfolio's exposure to these two classes :

$$r = R_f + \beta_3(K_m - R_f) + b_s \cdot SMB + b_v \cdot HML + \alpha$$

Here r is the portfolio's expected rate of return, R_f is the risk-free return rate, and K_m is the return of the market portfolio. The "three factor" β is analogous to the classical β but not equal to it, since there are now two additional factors to do some of the work. "SMB" stands for "small minus big" and "HML" for "high minus low"; they measure the historic excess returns of small caps over big caps and of value stocks over growth stocks. These factors are calculated with combinations of portfolios composed by ranked stocks (BtM ranking, Cap ranking) and available historical market data.

A General Framework of Sparse Modeling

Regularized Estimation Procedure

Goodness-of-Fit + Penalty on Model Complexity

- The goodness-of-fit term can be RSS, negative of the log-likelihood, empirical risk.
- Bayesian interpretation: the procedure is equivalent to maximum a posterior (MAP) with a suitably chosen proper or improper prior.
- Important questions: the limit of dimensionality; the role of penalty function; statistical properties; efficient implementation.

To be specific, we consider the following probability model $\{\mathbb{P}_\theta : \theta \in \Omega\}$.

Suppose sample $Z^n = \{(x_i, y_i)\}_{i=1}^n$ drawn from unknown $\mathbb{P}_{\theta_0}$.

The regularized estimation procedure is

$$\hat{\theta}_n \in \text{argmin}_{\theta \in \Omega} \left[L(\theta; Z^n) + \lambda_n r(\theta) \right]$$

The theory to be developed for the regularized estimation procedure includes

- Upper bounds on $\|\hat{\theta}_n - \theta_0\|$.
- Asymptotic results allowing for $(n, p, s_k) \to \infty$, where n is the sample size, p is the dimension of Ω, and s_k are the structural parameters, such as sparsity, rank, etc.

Examples:

Example 1 (Lasso Regression): Consider the linear regression model

$$Y_{n\times 1} = X_{n\times p}\theta_{p\times 1} + \varepsilon_{n\times 1}.$$

Lasso estimation procedure:

$$\hat{\theta}_n \in \text{argmin}_\theta \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \lambda_n \sum_{j=1}^{p}|\theta_j| \right].$$

Example 2 (Low Rank Matrix Approximation): Consider the multivariate regression model

$$Y_{n\times q} = X_{n\times p}\Theta_{p\times q} + E_{n\times q}.$$

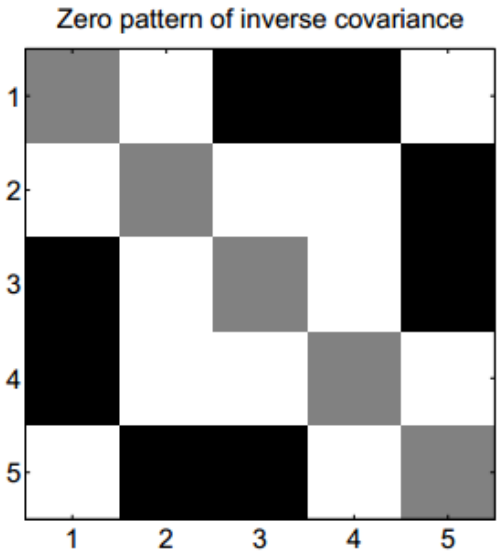
The low rank estimation procedures for Θ:

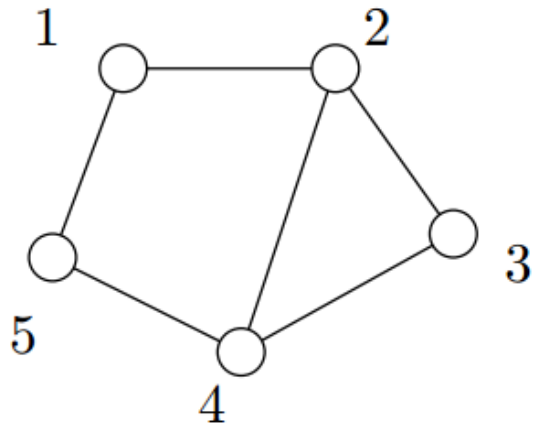
$$\hat{\Theta} \in \text{argmin}_\Theta \left[\| Y - X\Theta\|_F^2 + P_\lambda(\Theta) \right],$$

where $P_\lambda(\Theta)$ can be chosen as

$$\lambda_n \sum_{j=1}^{p\wedge q} I(\sigma_j(\Theta) \neq 0), \quad \lambda_n \sum_{j=1}^{p\wedge q} \sigma_j(\Theta), \quad \lambda_n \sum_{j=1}^{p\wedge q} w_j\sigma_j(\Theta).$$

Example 3 (Structured Inverse Covariance Matrix Estimation):





Set up: Samples from random vector with sparse covariance Σ or sparse inverse covariance Θ.

Estimation Procedure:

$$\hat{\Theta} \in \text{argmin}_{\Theta} \left[\ll \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T, \Theta \gg - \log\det(\Theta) + \lambda_n \sum_{b \in B} \|\Theta_b\|_F \right].$$

A Fact on Multivariate Normal.

Claim: Suppose $X \sim N_d(\mu, \Sigma)$, $\Theta = \Sigma^{-1}$. If the (i,j)-th component of Θ is zero, then the i-th component and j-th component of X are conditionally independent.

Proof: The proof is a immediate consequence of the following argument. Note that if

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \sim N_d \left(\begin{pmatrix} \mu^{(1)} \\ \mu^{(b)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

where $X^{(1)}$ is d_1-dimensional, and $X^{(2)}$ is $d_2 = d - d_1$-dimensional, then

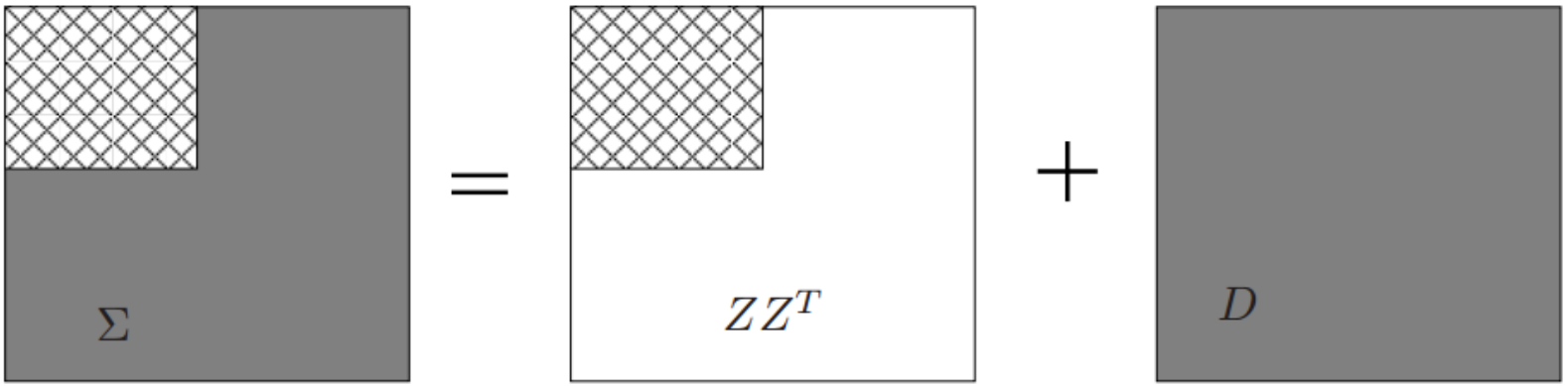
$$X^{(1)}|X^{(2)} = N_{d_1}(\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(X^{(2)} - \mu^{(2)}), \Sigma_{11.2}),$$

where $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Also write $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Then

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{11.2}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{21})^{-1} & \Sigma_{22.1}^{-1} \end{pmatrix}.$$

If $\Sigma_{11.2}^{-1}$ is diagonal, so is $\Sigma_{11.2}$.

Example 4 (Sparse PCA):



Set up: Covariance matrix $\Sigma = ZZ^T + D$, where the leading eigenspace Z has sparse columns.

Goal: Produce an estimate $\hat{Z}$ based on samples X_i with covariance matrix Σ.

Outline

References

- Bellman, R.E. (1961). Adaptive control processes: a guided tour. Princeton University Press.
- Fan,J.Q., Han, F. and Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*,1, 293-314.
- Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36, 2605-2637.
- Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20, 101-148.
- Donoho, D.L. (2000). High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Lecture given in the conference of "Math Challenges of the 21st Century" held by the American Math. Society in Los Angeles, August 6-11.