Introduction to High-Dimensional Data

T. V. Ramanathan

ram@unipune.ac.in

Department of Statistics
Savitribai Phule Pune University
Pune - 411007 (India).

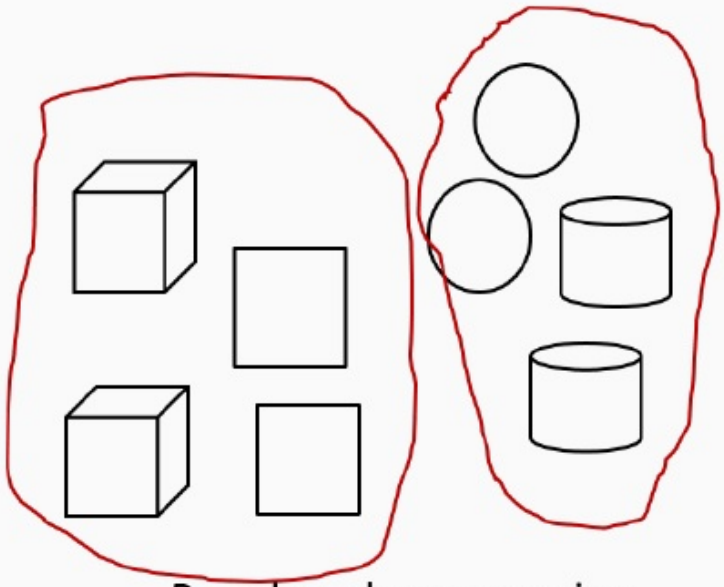
One-Day International Workshop on
High-Dimensional Data and AI/ML Algorithms with Applications

School of Mathematical & Computing Sciences,
Savitribai Phule Pune University, Pune

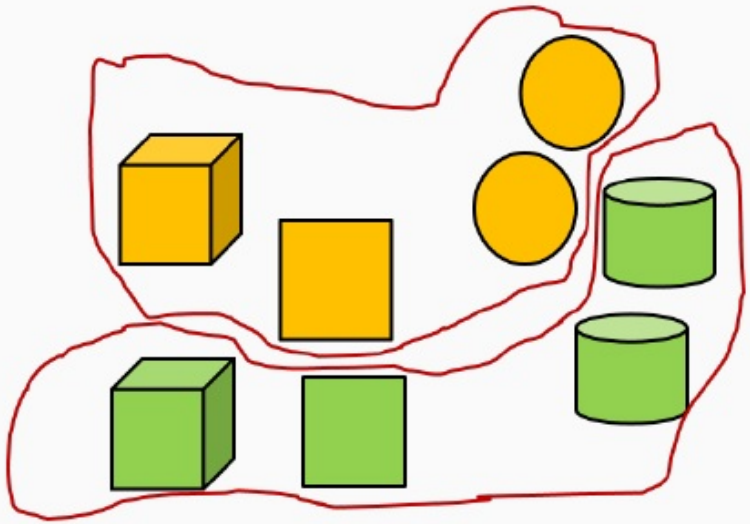
March 27, 2023

HD Data - A Motivating Example

- Let's play the baby shapes game (truly motivating for students ...): Group the items!!!



Based on shape grouping



Based on color grouping

- What about grouping based on both shape and color?
- Lesson to learn: there may be different semantic concepts (and their corresponding patterns) hidden in the data (here: shape and color)

High-Dimensional Data Analysis

- Traditionally, data analysis is a part of the subject of statistics with its basics in probability theory, decision theory and analysis.
- New sources of data — such as from satellites, RFID, Censors etc. — generate automatically huge volumes of data whose summarization called for a wide variety of data processing and analysis tools.
- For such data, traditional ideas of mathematical statistics such as hypothesis testing and confidence intervals do not help.

Some HD Data Problems - Economics, Banking & Finance

- **Integration** of macro-economic, banking, monetary and financial data, huge in size, large number of predictors
- **Risk Management** (Market, Credit, Operational etc.) and their estimation with high dimensional data
  Stock prices, currency and derivative trades, transaction records, high-frequency trades, unstructured news and texts, Claims data across insurance firms,
- **VAR modeling** - with sparsity assumptions (**Large VAR**) - number of parameters grows quadratically - over fitting & bad prediction
- **Portfolio optimization & Risk management** e.g., 1000 stocks means 5,00,500 covariance parameters - sparsity - High dimensional covariance matrix estimation
- **Market micro structure & Duration modeling** with High Frequency Data.

High Dimensionality

- **Curse of dimensionality:** - Richard Bellman - Dynamic Programming

  In optimization: if we must minimize a function f of d variables and we know that it is Lipschitz, that is,

  $$|f(x) - f(y)| \leq C\|x - y\|, \ x, y \in R^d$$

  then we need to order $\left(\frac{1}{\epsilon}\right)^d$ evaluations on a grid in order to approximate the minimizer within error ϵ

High Dimensionality

- Nonparametric regression:

$$X_{i1} = f(X_{i2}, \cdots, X_{id}) + \epsilon_i$$

- Assume that f is Lipschitz and $\epsilon_i \sim iidN(0,1)$. How does the accuracy of the estimate depend on N?
- Let Θ be the class of functions f which are Lipchitz on $[0,1]^d$. Then, it can be shown that

$$sup_{f \in \Theta} E[\hat{f} - f(X)]^2 \geq C \; N^{-2/(2+d)}$$

(cf. Ibragimov & Khasminskii (1981))

- It can be seen that the sample size increases as dimension d increases.

High Dimensionality

- **Blessings of dimensionality:**
- Theoretical benefits due to probability theory. The regularity of having many "identical" dimensions over which one can "average" is a fundamental tool.

High Dimensionality

- **Concentration of measure:** The "concentration of measure phenomenon" is about probabilities on product spaces in high dimension.
- Suppose we have a Lipschitz function f on R^d : Let P be the uniform distribution on the sphere in R^d and let X be a random variable with probability measure P.
- Then,

$$P[|f(X) - E[f(X)]| > t] \leq C_1 e^{-C_2 t^2}$$

  where C_1 and C_2 are constants independent of f and dimension d.
- In other words, a Lipschitz function is almost a constant.

Dimension asymptotics: Another use is that we can obtain results on the phenomenon by letting the dimension go to infinity.

High Dimensionality

- **Approach to continuum:** Some times high dimensional data arises because the underlying objects are really in a continuous space or a continuous phenomena; there is an underlying curve or image that we are sampling such as in functional data analysis or image processing.
- As the measured curves are continuous, there is an underlying compactness to the space of observed data which will be reflected by an approximate finite-dimensionality and an increasing simplicity of analysis for large d.

Example:

- Suppose we have d equi-spaced samples on an underlying curve $B(t)$ on the interval $[0, 1]$ which is a Brownian bridge. We have a d-dimensional data $X_{id} = B(i/d)$
- Suppose we are interested in $max_i\ X_{id}$
- Obviously this tends to $max_{t\in[0,1]}B(t)$ for large d.
- Here we know the exact distribution of $max_{t\in[0,1]}B(t)$ from Kolomogorov-Smirnov.

Challenges

Statistical

- High dimensionality brings **noise accumulation**, **spurious correlations** and **incidental endogeneity**.

Computational

- High dimensionality combined with large sample size creates issues such as **heavy computational cost** and **algorithmic instability**.

Paradigm Shifts in Statistical Thinking

Need for new statistical thinking

- Need for dimension reduction and variable selection to address noise accumulation issues.
- Need for high dimensional classification - new regularization methods
- Need for methods to tackle spurious correlations between response and some unrelated covariates.
- Need for methods to tackle incidental endogeneity (many unrelated covariates may be incidentally correlated with residual noise creates biases and model selection inconsistencies).

Noise Accumulation

- Analysis of High Dimensional Data - Simultaneously estimate or test several parameters
- Severe accumulation effect - noise may dominate the underlying signal - handled by sparse modeling and variable selection
- Variable selection plays a pivotal role in overcoming noise accumulation - but, variable selection in high dimension can bring other issues such as spurious correlation, incidental endogeneity etc.
- E.g., A classification problem with p features with n observations.
- The discriminative power for classification will be low as the number of features (m) in the PC is large due to increased noise accumulation.

Noise Accumulation

High dimensional classification:

- n data points from $N_p(\boldsymbol{\mu}_0, I_p)$; $N_p(\boldsymbol{\mu}_1, I_p)$, $p = 4500$, $\boldsymbol{\mu}_0 = 0$,
- $\boldsymbol{\mu}_1$ - 0 with probability 0.98 and standard DE with probability 0.02 - Most components have no discrimination power
- Even then, some components are very powerful in classification (2 %, or 90 realizations from DE, several components are very large and many are small).
- Distance based classifier put $\boldsymbol{x}$ into class 1 if :

$$\|\boldsymbol{x} - \boldsymbol{\mu}_1\|^2 \leq \|\boldsymbol{x} - \boldsymbol{\mu}_0\|^2 \quad \text{or} \quad \beta^T(\boldsymbol{x} - \boldsymbol{\mu}) \geq 0$$

where $\beta = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)/2$.

- Misclassification rate: $\Phi\left(-\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|/2\right)$
- This is effectively zero (WLLN) as

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

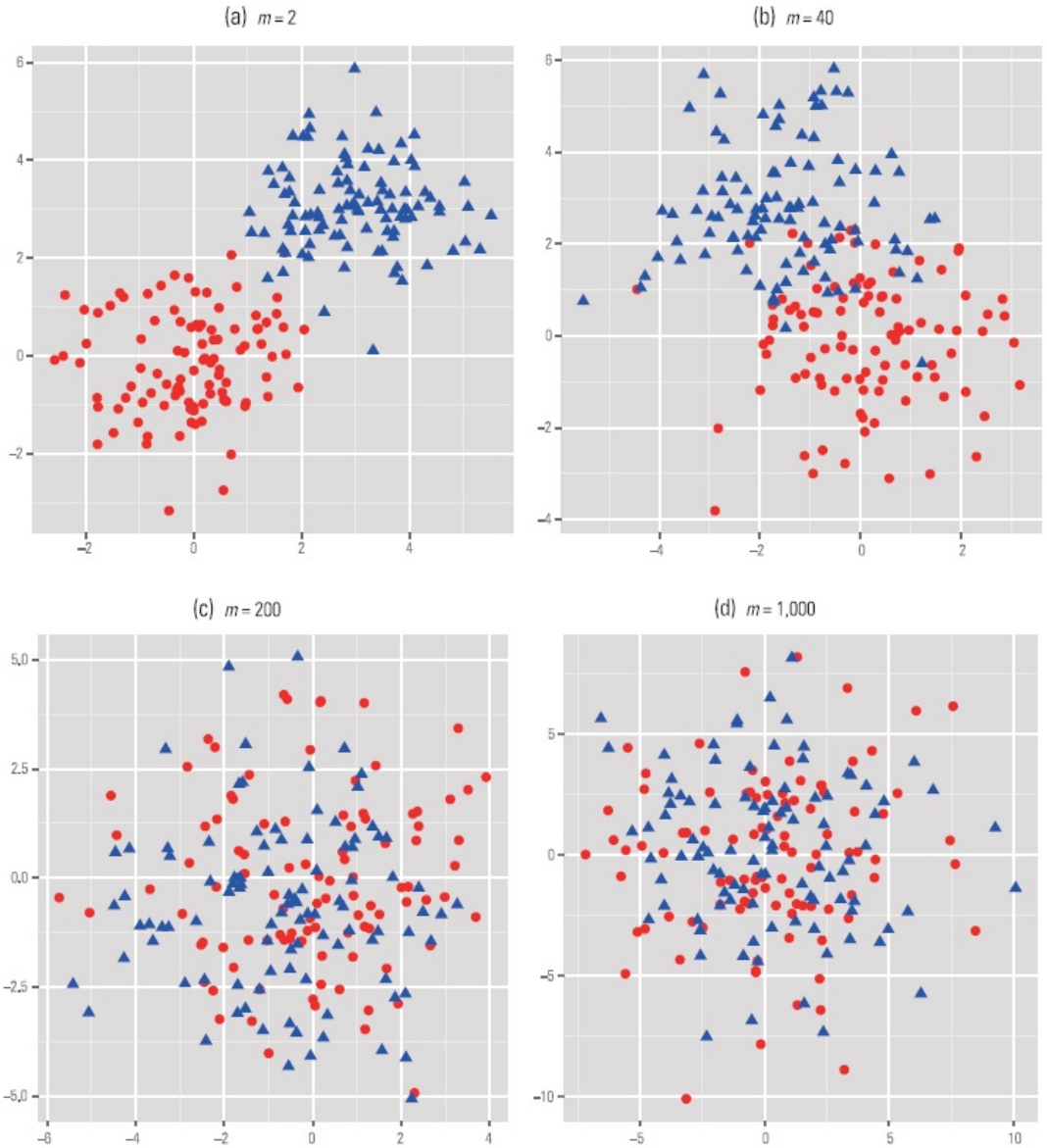
- When we estimate β, resulting classification rule behaves like random guess due to the accumulation of noise.

Noise Accumulation

High dimensional classification: Illustration

- $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n \sim N_p(\boldsymbol{\mu}_1, I_p)$; $\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_n \sim N_p(\boldsymbol{\mu}_2, I_p)$
- To classify $\boldsymbol{Z} \in \mathcal{R}^p$ into one of this.
- Let $p = 1000,\ n = 100,\ \boldsymbol{\mu}_1 = \mathbf{0},\ \boldsymbol{\mu}_2 =$ first 10 entires with value 3, all other entries zero.
- Plot the first two principal components using the first $m = 2,\ 40,\ 200$ and 1000 features.
- The first 10 features contribute to classifications, but when $m > 10$, procedures do not obtain any additional signals, only accumulate noises.
- For $m = 40$, the accumulated signals compensate the accumulated noise, so that the first two principal components still have good discriminative power.
- When $m = 200$, the accumulated noise exceeds the signal gains. - Shows the **need for sparse models in HD classification.**

Noise Accumulation



Spurious Correlation

- A feature of high dimensionality - variables that are not correlated theoretically, but the sample correlation will be very high.
- Important variables can be highly correlated with several spurious variables which are scientifically unrelated.
- **Lead to false scientific discoveries and wrong statistical inference**
- **Impact on variable selection**
- **Variance will be seriously under estimated - bias will be very large.**

Spurious Correlation

(Fan, Han & Liu, 2014 NSR)

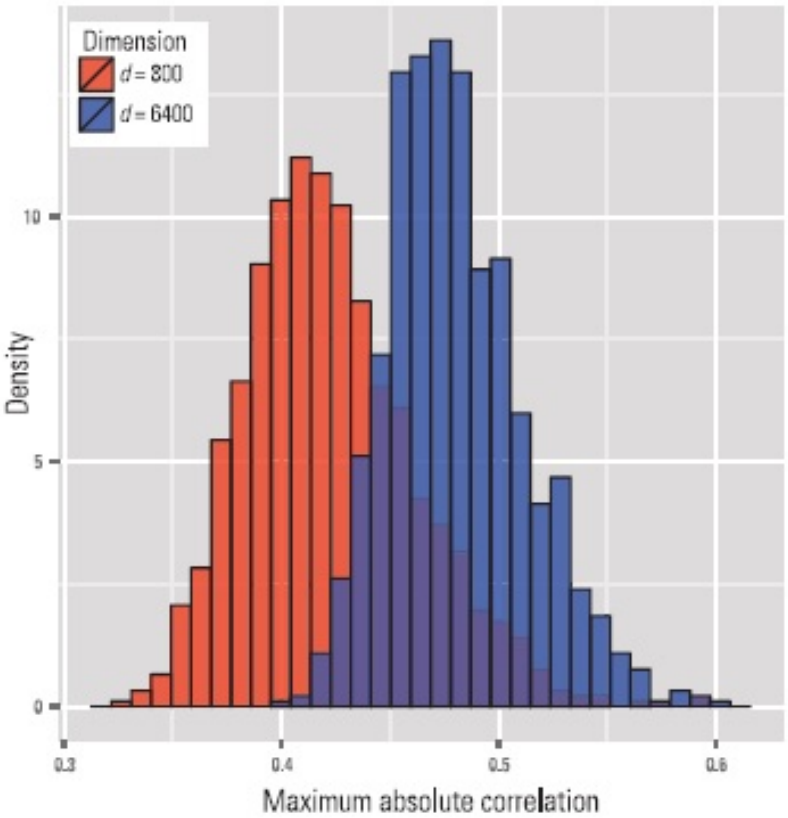
- Consider a random sample of size $n = 60$ of $p-$ dimensional $N(0, I_p)$. - Population correlation between any two is zero.
- Corresponding sample correlation should be also small - indeed the case when p is small, but need not be when p large.
- Compute the maximum of the sample correlation and the maximum of multiple correlation

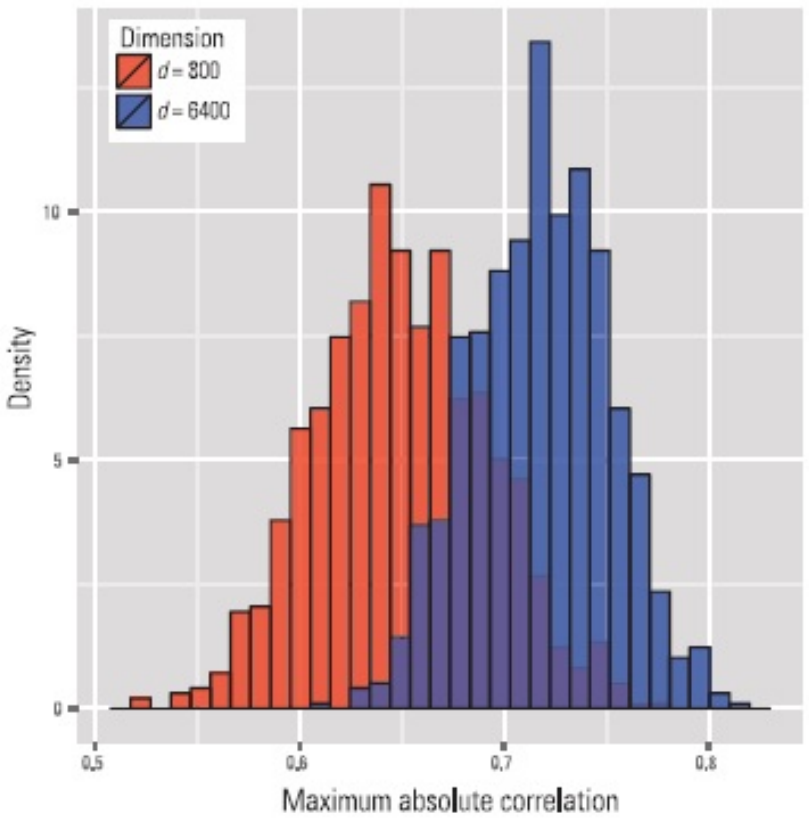
$$\hat{r} = max_{j \geq 2} \ |\hat{corr}(X_1, X_j)|,$$

$$\hat{R} = max_{|\mathcal{S}|=4} \ |\hat{corr}(X_1, \boldsymbol{X}_{\mathcal{S}})|, \quad 1 \notin \mathcal{S}$$

- $\hat{R}$ is nothing but the correlation between X_1 and its best linear predictor using $\boldsymbol{X}_{\mathcal{S}}$ (In the computation, we may use forward selection algorithm to compute $\hat{R}$, which is no larger that $\hat{R}$, but avoids computing all $\binom{p}{4}$ multiple R^2.)
- Suppose we simulate this data for $p = 800$ and $p = 6400$ for 1000 times

Spurious Correlation





Spurious Correlation

- We may also denote the max. abs. multiple correlation as

$$\hat{R} = max\ max_{|\mathcal{S}|=4,\ \{\beta_j\}_{j=1}^{4}} |c\hat{o}rr(X_1, \sum_{j\in\mathcal{S}} \beta_j X_j)|$$

- Note that the empirical distribution of $\hat{r}$ and $\hat{R}$ are not concentrated around zero. In fact, they go away from zero as p increases.
- Theoretical results on $\hat{r}$ can be found in (Cai & Jiang, 2012, JMA and Fan, Guo and Hao, 2012, JRSS B).
- Note that as a consequence of high spurious correlation, X_1 is practically indistinguishable from $\boldsymbol{X}_{\hat{\mathcal{S}}}$ for a set $\hat{\mathcal{S}}$ with $|\hat{\mathcal{S}}| = 4$.
- If X_1 represents the expression level of a gene that is responsible for a disease, we cannot distinguish it from other four genes in $\hat{\mathcal{S}}$ that have a similar predictive power, although they are unrelated to the disease (scientifically irrelevant). (It may happen vice-versa also).

Spurious Correlation

Spurious correlation also affects the statistical inference, besides variable selection.

- $\boldsymbol{Y} = \boldsymbol{X}^T\beta + \epsilon, \quad \sigma^2 = Var(\epsilon)$
- Residual variance based on selected variables

$$\hat{\sigma}^2 = \frac{1}{n - |\hat{\mathcal{S}}|}\boldsymbol{Y}^T(I_n - P_{\hat{\mathcal{S}}})\boldsymbol{Y}, \quad P_{\hat{\mathcal{S}}} = \boldsymbol{X}_{\hat{\mathcal{S}}}(\boldsymbol{X}_{\hat{\mathcal{S}}}{}^T\boldsymbol{X}_{\hat{\mathcal{S}}})^{-1}\boldsymbol{X}_{\hat{\mathcal{S}}}$$

- when the variables are not selected, and the model is unbiased, the d.f. adjustment makes the residual variance unbiased.
- Let $\beta = 0, \ Y = \epsilon$ - all selected variables are spurious
- If the number of selected variable is much less than n,

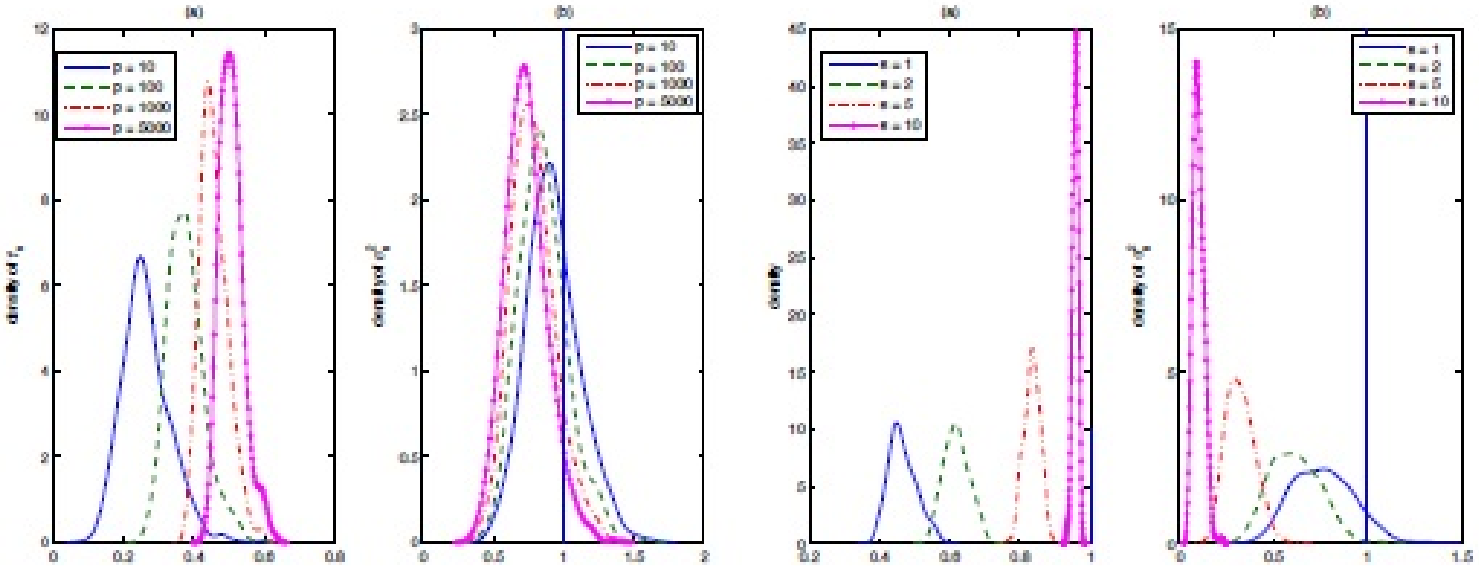
$$\hat{\sigma}^2 = \frac{1}{n - |\hat{\mathcal{S}}|}(1 - \gamma_n^2)\left\|\epsilon\right\|^2 \approx (1 - \gamma_n^2)\sigma^2$$

$$\gamma_n^2 = \epsilon^T P_{\hat{\mathcal{S}}}\epsilon / \left\|\epsilon\right\|^2$$

- σ^2 is under estimated by a factor γ_n^2 - Statistical inference will be in trouble.

Spurious Correlation

Left panel represents distributions of γ_n and $\sigma^2 = 1$ when $|\hat{\mathcal{S}}| = 1$. In the other case, $Y = 2X_1 + .3X_2 + \epsilon$, $p = 1000$, $n = 50$.



Incidental Endogeneity

- A researcher collects information about covariates which are potentially related to the response.
- A regressor is said to be *endogenous* when it is correlated with the error term, and *exogenous* otherwise.
- Consider the conventional sparse model, which assumes

$$Y = \sum_{j=1}^{p} \beta_j X_j + \epsilon, \quad \text{with} \quad E(\epsilon X_j) = 0, \quad j = 1, 2, ..., p.$$

  with a small set $\mathcal{S} = \{j : \beta_j \neq 0\}$.
- Some predictors are correlated with residual noise.
- Whenever more covariates are collected or measured, hardly the exogenous assumption is satisfied.
- Unlike spurious correlation, incidental endogeneity refers to the genuine existence of correlations between variables unintentionally, both due to high dimensionality.

Incidental Endogeneity

- Endogeneity occurs as a result of selection biases, measurement errors and omitted variables. Also, it could be incidental (as a consequence of large number of predictors available)
- Big Data are usually aggregated from multiple sources with potentially different data generating schemes. This increases the possibility of selection bias and measurement errors, which also cause potential incidental endogeneity.
- Consequence : **Endogeneity causes the inconsistency of the penalized least-squares method and possible false scientific discoveries.** (Fan & Liao, 2014, AS)

Incidental Endogeneity

- **How to test this in practice?** The problem of dealing with endogenous variables is not well understood in high-dimensional statistical analysis.

- The condition $E(\epsilon X_j) = 0, j = 1, 2, ..., p$ is too restrictive for real applications. A more realistic model assumption would be

$$E(\epsilon | \{X_j\}_{j \in \mathcal{S}}) = 0.$$

- (Fan & Liao, 2014, AS) considered still a weaker condition ("over identification") viz.,

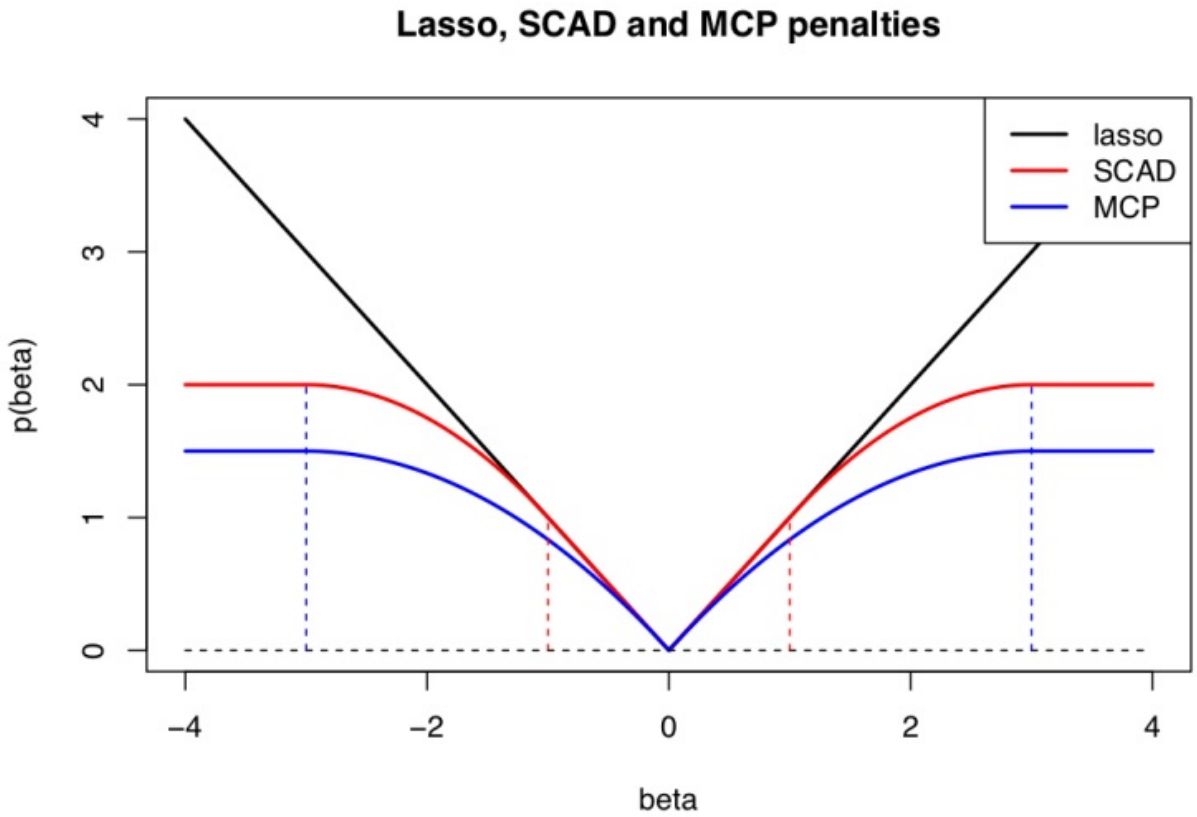
$$E(\epsilon X_j) = 0 \text{ and } E(\epsilon X_j^2) = 0, \ j \in \mathcal{S}.$$

These authors have showed that under the abouve condition, classical penalized least squares methods such as LASSO, SCAD and MCP are no more consistent.

Sparsity

- $Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i, \ \ i = 1, 2, ..., n$ with $p \geq n$.
- Sparsity can be quantified in terms of ℓ_q-norm for $1 \leq q \leq \infty$ analogue to a ℓ_0, which is not a norm.
- $\|\beta\|_0^0 = |\{j; \beta_j \neq 0\}| = \sum_{j=1}^{p} |\beta_j|^0$ $(0^0 = 0)$ - count the number of non-zero entries
- In analogy, $\|\beta\|_q^q = \sum_{j=1}^{p} |\beta_j|^q$ for $0 < q < \infty$. (or $q = 1$ it measures the sparsity in a different way and has the computational advantage of being convex in β.
- Roughly, high-dimensional statistical inference is possible, in the sense of leading to reasonable accuracy or asymptotic consistency, if $log(p)$. $(sparsity(\beta)) << n$, depending on how we define sparsity.

Penalties


Lasso, SCAD and MCP penalties
lasso
SCAD
MCP
p(beta)
0
1
2
3
4
−4
−2
0
2
4
beta

Incidental Endogeneity

- (Fan & Liao, 2014, AS) introduced a penalized method, called **focused generalized method of moments** (FGMM). The FGMM effectively achieves the dimension reduction and applies the instrumental variable methods.
- They have shown that FGMM possesses the oracle property even in the presence of endogenous predictors, and that the solution is also near global minimum under the over-identification assumption.

- Penalized Quasi Likelihood

- Classical model selection: Minimize the quasi likelihood

$$-QL(\beta) + \lambda \|\beta\|_0$$

- $\|.\|_0$ - l_0-pseudo-norm (number of non-zero entries in a vector)
- $\lambda > 0$ - regularization parameter - control bias variance trade-off
- More general form:

$$l_n(\beta) + \sum_{j=1}^{p} \rho_{\lambda,\gamma}(|\beta_j|)$$

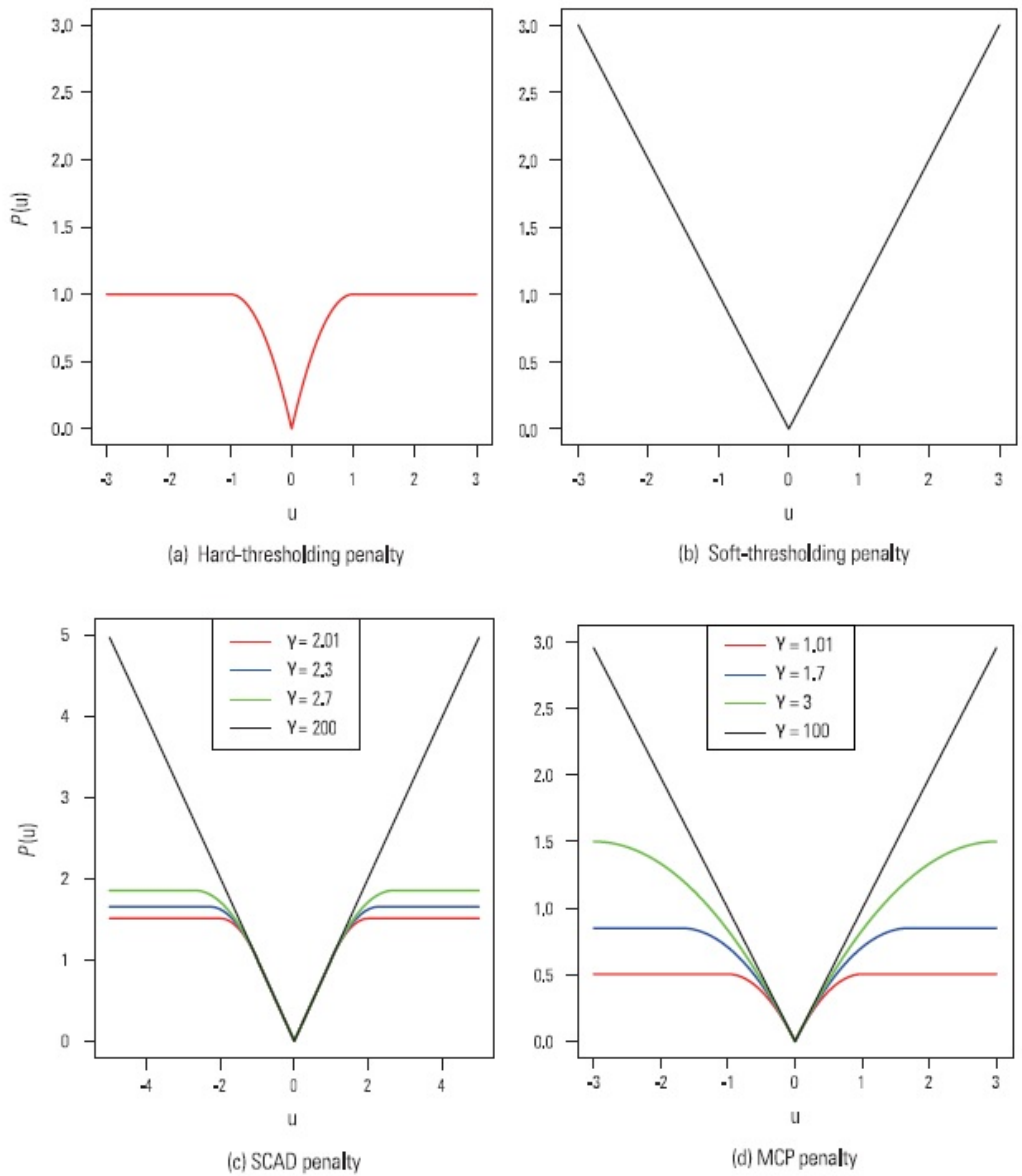
- $l_n(\beta)$ measures the goodness-of-fit of the model with parameter β
- $\sum_{j=1}^{p} \rho_{\lambda,\gamma}(|\beta_j|)$ - Sparsity inducing penalty (encourages sparsity)
- λ - tuning parameter that controls the bias-variance trade-off
- γ- a possible fine-tune parameter which controls the degree of concavity of the penalty function.

Penalized Quasi Likelihood

Some of the commonly used penalty functions are given below.

- Bridge: $\rho_{\lambda,\gamma}(|\beta|) = |\beta|^q$
- Ridge: Same as above, but with $q = 2$.
- Lasso: $\rho_{\lambda,\gamma}(|\beta|) = |\beta|$ (Least absolute selection and shrinkage operator)
- SCAD: $\rho_{\lambda,\gamma}(|\beta|) = a_1(\lambda,\gamma)I[0 \leq \beta < \lambda] + a_2(\lambda,\gamma)I[\lambda \leq \beta \leq k\lambda] + a_3(\lambda,\gamma)I[k\lambda < \beta]$ (Smoothly clipped absolute deviation)
- The penalties are nondifferentiable at 0, which is necessary for sparsity.
- The Lasso is convex while the bridge and SCAD penalties are nonconvex.
- Nonconvexity is necessary for unbiasedness of the estimated coefficients.

Penalized Quasi Likelihood



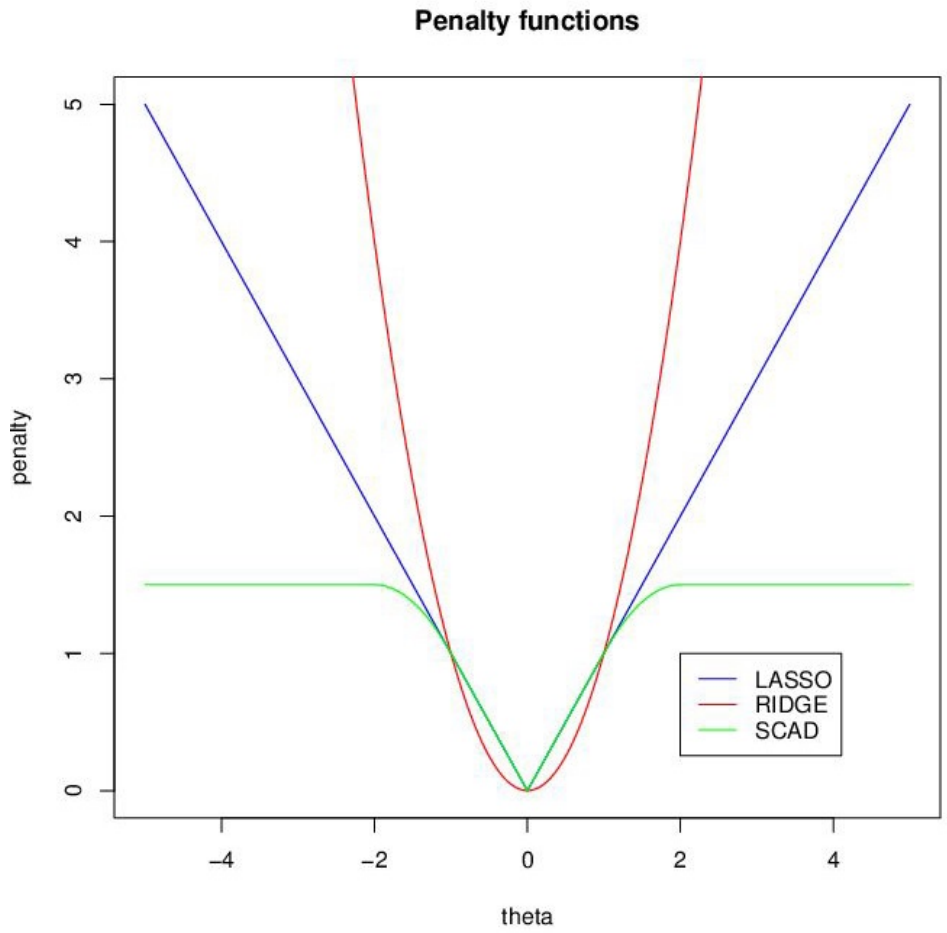
(a) Hard-thresholding penalty

(b) Soft-thresholding penalty

(c) SCAD penalty

(d) MCP penalty

- Penalized Quasi Likelihood



Penalized Quasi Likelihood

- How shall we choose among these penalty functions?
- Sparsity and computing time should be the decisive factors.
- In applications, it is recommended to use either SCAD or MCP (minimax concavity penalty) thresholding, since they combine the advantages of both hard- and soft-thresholding operators.
- Many efficient algorithms have been proposed for solving the optimization problem in using different penalties. (Candes and Tao, 2007, AS)

Penalized Quasi Likelihood

- The **oracle property** means that the penalized estimator is asymptotically equivalent to the oracle estimator that is the ideal estimator obtained only with signal variables without penalization.
- Many noncovex penalties such as the bridge and SCAD penalties possess the oracle property.
- In practice, however, only a local minimum (of the penalized sum of squared residuals) is given, and it is extremely difficult (almost impossible) to check if a given local minimum is (asymptotically) the oracle estimator.
- In this sense, the oracle property of a nonconvex penalty is practically meaningful only when reasonable local minima are asymptotically equivalent to the oracle estimator.

Portfolio Optimization with High Dimensional Data

- **Portfolio optimization:**
  Minimize

$$w^{'}\Sigma w$$

  such that

$$\sum_{i=1}^{p} w_i = 1,$$

  where, w_i's are weights associated with i^{th} asset and Σ is the variance covariance matrix of the assets.

- Σ is to be estimated (as it is unknown).

- When the number of parameter increases, the estimation can be difficult and the accuracy of the estimate may not be maintained.

New Methods of Covariance Estimation

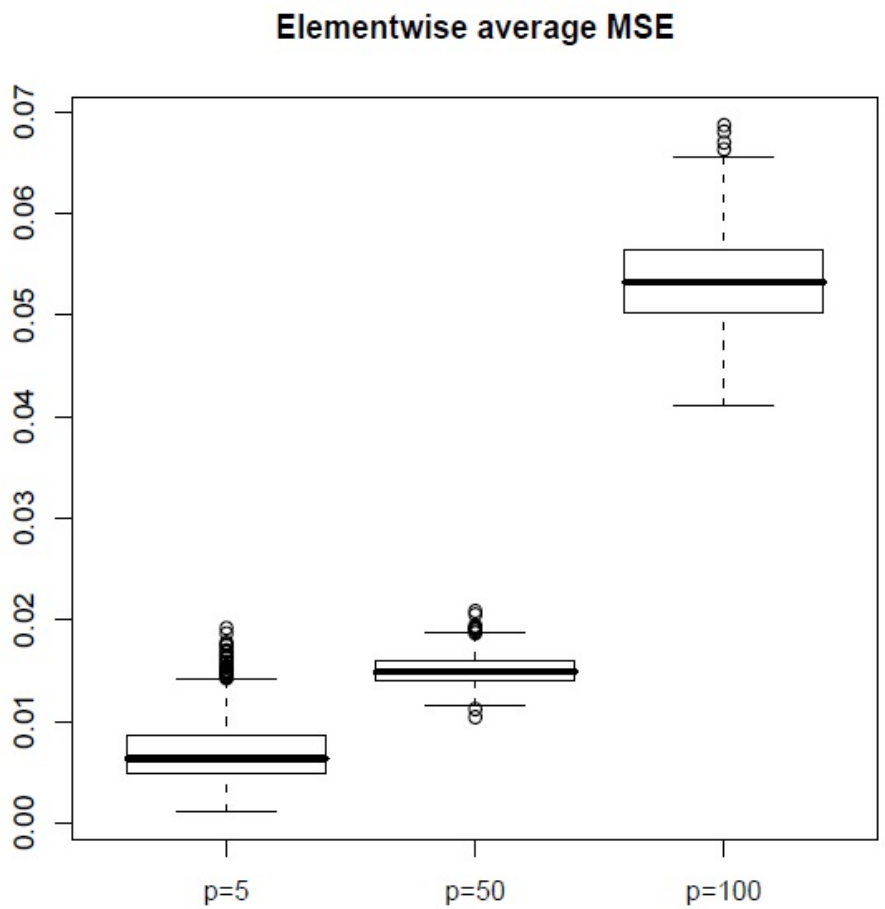
- **The shrinkage method:** A linear combination of the sample estimator and another estimator.
- **Factor models:** Matrices implied by the large dimensional factor models - observable or latent - principal components method and the maximum likelihood method.
- **Bayesian and empirical Bayes estimators:** Related to shrinkage estimator - provide alternative interpretations for the shrinkage method.
- **Method based on Random Matrix Theory:** Aims to attenuate the randomness of the sample covariance S using the theory of random matrices of high dimension. (El Karoui (2008), Artur Kotlicky (2015), Recent Papers by Arup Bose)

1. Sample Variance-Covariance Matrix

- Simple to construct and unbiased.
- When invertible, the sample covariance coincides with the classical maximum likelihood estimate.
- Contains a lot of estimation error when the number of observations n is less than the number of variables say p, in which case, it is not invertible, even though the underlying true covariance matrix is invertible.
- When n is comparable to p, it has significant amount of sampling error.
- Extremely sensitive to outliers.
- Simulation Study: $N(0, \Sigma)$, $\Sigma = I$, $ER = \Sigma^{-1} - S^{-1}$,
  $(1/p^2)\,\|ER\|_F^2 = (1/p^2) \sum_{i=1}^{p} \sum_{j=1}^{p} er_{ij}^2$,
  er_{ij} is the $(i,j)^{th}$ element of ER,

1. Sample Covariance Matrix - Element wise MSE of Precision Matrix

Sample size = 200, Simulation = 1000.



2. Penalized Estimation Using Matrix Log Transformation

- To obtain a positive definite estimate of the covariance matrix, which is an accurate estimate with a well-structured eigen-system.
- A regularized approach may be adopted to estimate Σ using the approximate log-likelihood function of $l_n(A)$, where $\Sigma = exp(A)$.
- The penalty function $\|A\|_F^2$ i.e the Frobenius norm of A, which is equivalent to $tr(A^2)$ is used.
- Estimate Σ, or equivalently A, by minimizing

$$l_{n,\lambda}(A) = l_n(A) + \lambda tr(A^2)$$

where λ is a tuning parameter.

2. Penalized Estimation Using Matrix Log Transformation

- Tuning parameter is a trade-off between the likelihood function and the penalty function.
- Set $A_0 = log(\Sigma_0)$, where $\Sigma_0 = S + \epsilon I$, ϵ is a pre-specified small positive quantity.
- Spectral decomposition of Σ_0, we get $T_0 D_0 T_0'$.
- Get $\hat{B}$ by minimizing $l_{n,\lambda}$, where $B = T_0'(A - A_0)T_0$.
- Get $\hat{A} = T_0 \hat{B} T_0' + A_0$ and estimate $\hat{\Sigma} = exp(\hat{A})$.
- Stop if $\left\| \hat{\Sigma} - \Sigma_0 \right\|_F^2 < \delta$ (pre sepcified).

3. Shrinkage Estimator

- The shrinkage estimator is a linear combination of the sample covariance matrix S and a highly structured estimator F.
- Compromise between the two by computing a convex linear combination

$$\Sigma = \delta F + (1 - \delta) S$$

- δ - Shrinkage constant, $(0 < \delta < 1)$.
- Here the sample covariance matrix is 'shrunk' towards the structured estimator.
- We consider a constant correlation model for F.

3. Optimal Shrinkage Estimator - Ledoit & Wolf (LW)

- The Structured estimator is F is given by

$$f_{ii} = s_{ii}, \quad f_{ij} = \bar{r}\sqrt{s_{ii}s_{jj}}, \; r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

and $\bar{r} = 2((N-1)N)^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} r_{ij}$

- The shrinkage constant is
$\hat{\delta}^* = max(0, min(1, \hat{\kappa}/T))$, where $\hat{\kappa} = (\hat{\pi} - \hat{\rho})/\hat{\gamma}$.

- $\pi-$ sum of asymptotic variances of entries of S, $\rho-$ asymptotic covariances of entries of F with entries of S and $\gamma-$ mis specification of shrinkage target. (all the three can be consistently estimated)

- The optimal shrinkage estimator is given by

$$\hat{\Sigma}_{LW} = \hat{\delta}^* F + \left(1 - \hat{\delta}^*\right) S.$$

3. Shrinkage Estimator

- Performs better than the sample variance covariance matrix.
- An additional advantage of shrinkage estimator is that it is always positive definite i.e shrinkage estimator is a convex combination of an estimator that is positive definite (the shrinkage target F) and an estimator that is positive semi definite (the sample covariance matrix)
- LW Shrinkage estimator is distribution free.
- Here we consider the constant correlation model which gives comparable performance but is easier to implement. The model states that all the (pairwise) correlations are identical

4. Rao-Blackwell Ledoit-Wolf Estimator

- If the Gaussian assumption is true, then the LW estimator can be improved upon by applying the Rao-Blackwell theorem to the LW method, which results in a new estimator *RBLW*:

$$\hat{\Sigma} = \lambda^* S + (1 - \lambda^*) T$$

- T is a structured estimator defined as $T = \frac{tr(S)}{p} I$.
- λ^*_{RBLW} is the Rao-Blackwell Optimal Shrinkage Intensity, given by

$$\lambda^*_{RBLW} = \frac{\frac{n-2}{n} tr(S) + tr^2(S)}{(n+2)\left[tr(S^2) - \frac{tr(S)}{p}\right]}$$

$$\hat{\Sigma}_{RBLW} = E\left[\hat{\Sigma}_{LW} | S\right]$$

4. Rao-Blackwell Ledoit-Wolf Estimator

- The shrinkage intensity is modified to avoid over shrinkage:

$$\lambda^*_{RBLW} = min(1, \lambda^*_{RBLW})$$

- The RBLW estimator is

$$\hat{\Sigma}_{RBLW} = (1 - \lambda^*_{RBLW})S + \lambda^*_{RBLW} T$$

5. Factor Models

- Returns have factor structure - Risk can be expressed as a linear function of factor loadings
- The number of factors can be allowed to grow with the number of parameters.
- Asset returns are linear functions of k unobservable factors, $k < p$:
- $$X = \underline{\mu} + \Lambda \underline{f} + \underline{\epsilon}$$

  $X = (\underline{X}_1, \underline{X}_2, ..., \underline{X}_p)$ is a $n \times p$ matrix of asset returns, Λ is a $p \times k$ matrix of factor loadings.
- The implied covariance matrix is $\Sigma = \Lambda \Omega_f \Lambda^{'} + \Omega_\epsilon$ where, $\Omega_f = var(\underline{f})$, and $\Omega_\epsilon = var(\underline{\epsilon})$.

- Because both Λ and $\underline{f}$ are unobservable and they enter the model in a multiplicative way - cannot be identified separately without restrictions.
- Normalization is done such that $\Omega_f = I$, implying $\Sigma = \Lambda\Lambda^{'} + \Omega_\epsilon$.
- Standard methods provide an estimate of Σ.
- Ω_ϵ - a diagonal matrix. If not diagonal, but maximum eigen value is bounded, then it is called as an approximate factor model.

Data Description

Portfolio: Consists of 300 stocks.
Source: Yahoo Finance, National Stock Exchange (NSE), Bombay stock exchange (BSE) and Data market.
Few names of the sectors and companies considered.

- Banking sector: ICICI, HDFC, IDBI, Axis Bank, SBI, UBI, CBI etc.
- Automobile sector: Bajaj auto, Maruti Suzuki, Honda, Tata Motors
- Pharmacy sector: Glenmark, Cipla, Dr.Reddy's Ranbaxy etc.
- Financial sector: Bajaj finance, India bulls, Mahindra finance.
- Exchange rates of foreign currencies to INR : GBP, US dollar, Canadian dollar, Yen, Swiss franc, Euro etc.

Note: The data is of daily returns for the stocks and daily exchange rates for the financial year 2014-2015.

Realized Risk

P(Number of Stocks)	100	150	200	250	300
Methods	Realised Risk				
Sample Covariance Matrix	0.978226	0.806533	0.694745	0.641807	0.47658
Penalized Covariance Matrix Estimator	0.78956	0.60589	0.58587	0.50947	0.38457
Ledoit Wolf Shrinkage Estimator	0.022076	0.01342	0.011373	0.010764	0.016257
Rao Blackwell Ledoit-Wolf Shrinkage Estimator	0.019364	0.01192	0.00996	0.009291	0.009006
Latent Factor Model	0.252041	0.236381	0.205314	0.195873	0.182776

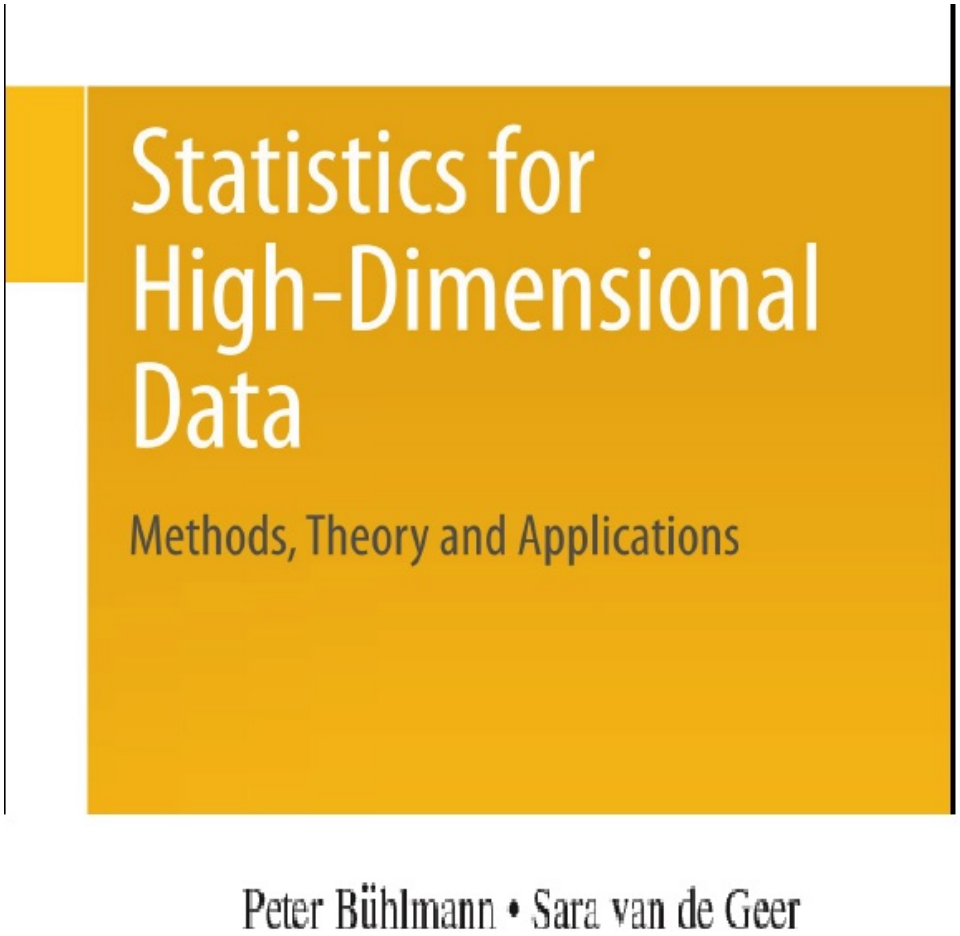
Data Analysis: Some Observations

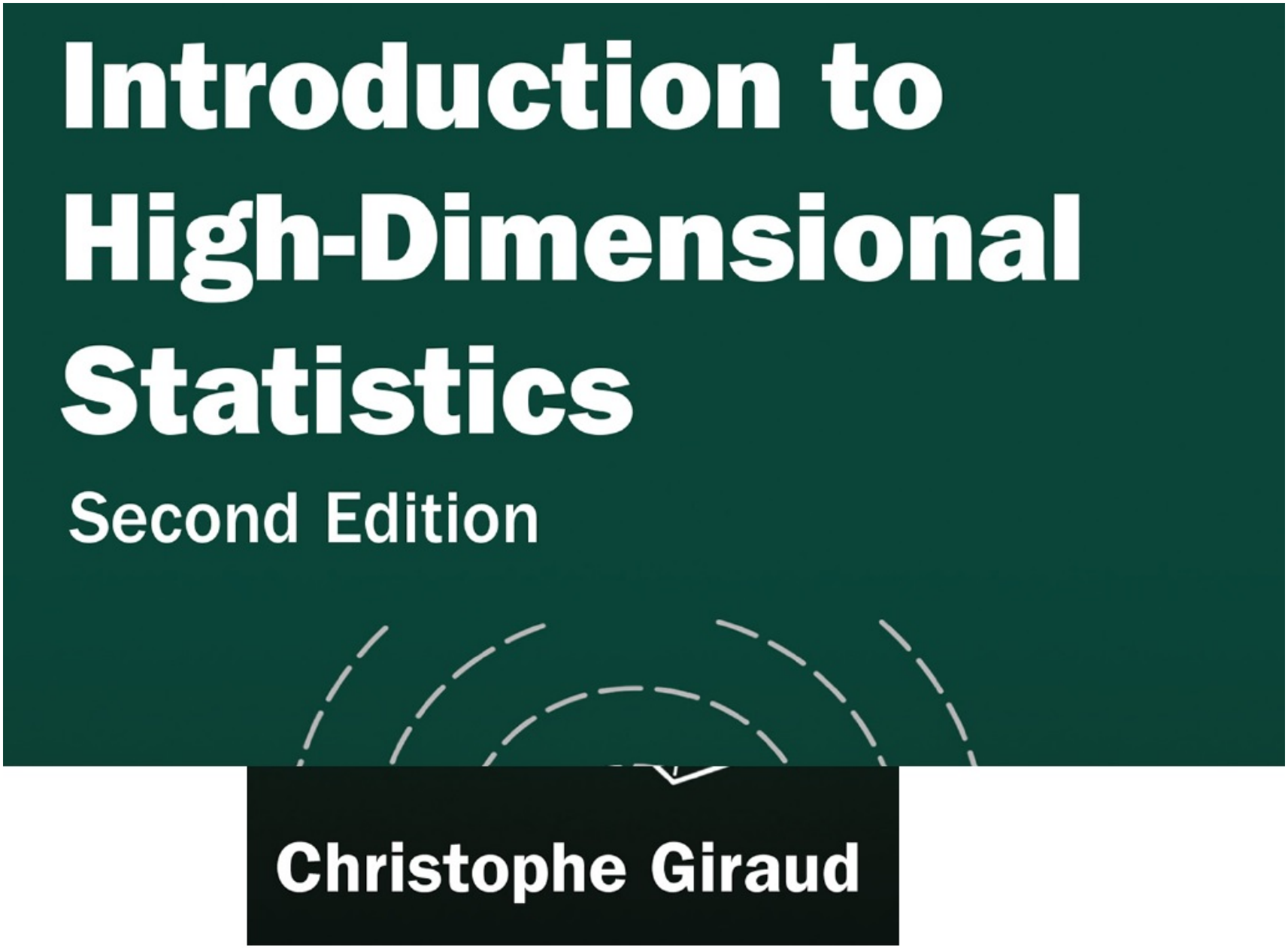
- Normality tests confirmed Gaussian nature of the returns.
- Rao-Blackwell Ledoit-Wolf Shrinkage estimator appears to be performing better compared to others.
- Ledoit-Wolf Shrinkage estimator also performs well. (It is distribution-free in nature)

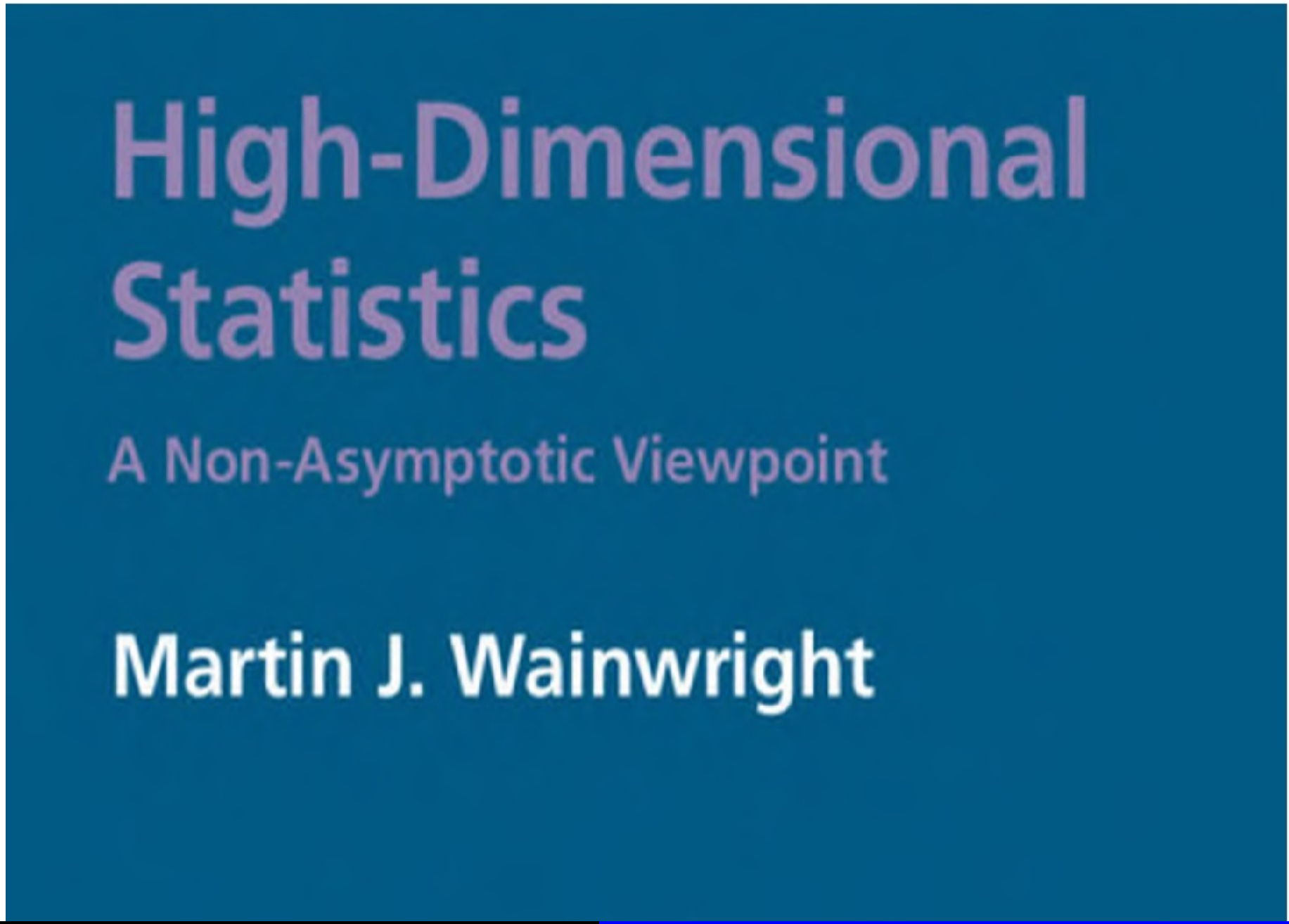
  The performance of factor model estimators are not very bad. In fact, these are better than the penalized methods.
- The sample covariance estimators are the worst.

Some Future Directions

- HD Covariance Estimation: Relaxing the assumption of normality, and instead, can we have heavy-tailed distributions - GHD would be a good choice.
- HD Covariance Estimation: Behaviour, when we use Value-at-Risk(VaR) and its generalizations, instead of realized risk.
- Application of multivariate GARCH or multivariate SV models to address the time dependency of variance.
- Model selection in HD - Use of FIC (Pandhare & Ramanathan, *Statistics*, 2022)
- Large VAR modeling
- Use of High Frequency Data for understanding Market micro structure - Duration modeling (DST Project, 2014-2017)


Statistics for
High-Dimensional
Data
Methods, Theory and Applications
Peter Bühlmann • Sara van de Geer


Introduction to
High-Dimensional
Statistics
Second Edition
Christophe Giraud


High-Dimensional
Statistics
A Non-Asymptotic Viewpoint
Martin J. Wainwright

THANK YOU

FOR YOUR ATTENTION