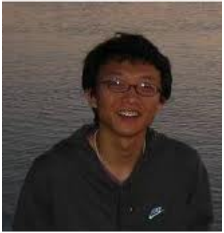
Back to the Future:

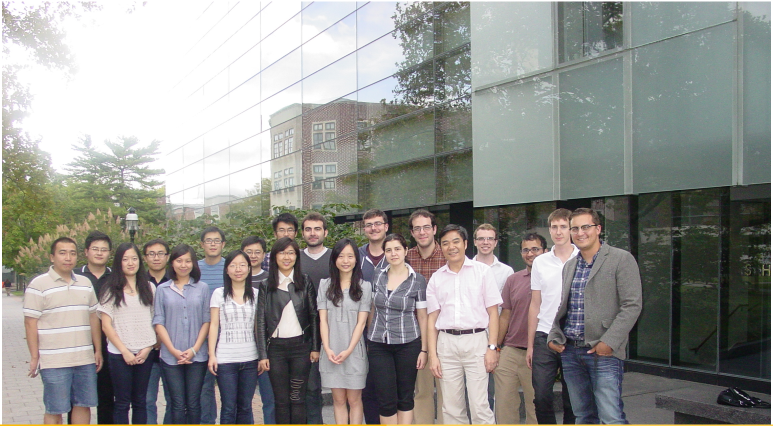
Valid Analysis of Big Data

Jianqing Fan

Princeton University

May 27, 2014





Outline

Are Fundamental Assumptions in High-dimensional Statistics Verifiable?

1. What are Big Data?
2. What are key assumptions in high-dim inference?
3. How to verify them?
4. What are the consequence when violated?
5. How to pose realistic and verifiable assumptions?

Explanation of Title

Most high-dim methods are based on $E(\varepsilon \mathbf{X}) = 0$ **(exogeneity)**.

They are unrealistic, and often wrong.

All high-dim math is beautiful and **correct**!

What is Big Data?

■ Large and Complex Data: ★ Structured (n and p are both large) ★ Unstructured (text, web, videos)

- ★ **Biological Sci.**: Genomics, Medicine, Genetics, Neurosci
- ★ **Engineering**: Machine learning, computer vision, networks.
- ★ **Social Sci.**: Economics, business, and digital humanities.
- ★ **Natural Sci.**: Meteorology, earth science, astronomy.

■ Characterize contemporary scientific and decision problems.

What is Big Data?

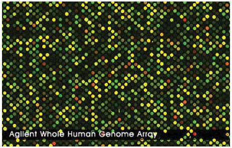
■Large and Complex Data: ★Structured (n and p are both large) ★Unstructured (text, web, videos)

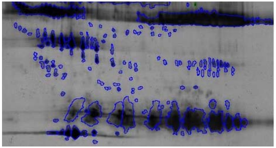
- ★ **Biological Sci.**: Genomics, Medicine, Genetics, Neurosci
- ★ **Engineering**: Machine learning, computer vision, networks.
- ★ **Social Sci.**: Economics, business, and digital humanities.
- ★ **Natural Sci.**: Meteorology, earth science, astronomy.

■Characterize contemporary scientific and decision problems.

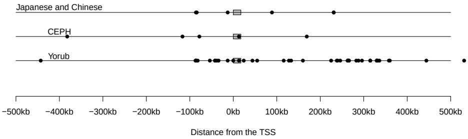
Examples: Biological Sciences

- Bioinformatic: disease classification / predicting clinical outcomes / biological process using microarray or proteomics data.





- Assoc. between phenotypes and SNPs & gene exp (QTL & eQTL).



- Detecting activated voxels after stimulii in neuroscience.

What can big data do?

Hold great promises for understanding

★ **Heterogeneity**: personalized medicine or services

★ **Commonality**: in presence of large variations (noises)

from large pools of variables, factors, genes, environments and their interactions as well as **latent factors**.

Aims of High-dimensional statistical inference

- **Risk property**: To construct as effective a method as possible to predict future observations. ★**Correlation**

- **Feature selection and risk property**: To gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction method. ★**Causation**

★Fan and Li (2006), Bickel (2008, JRSS-B)

Aims of High-dimensional statistical inference

- **Risk property**: To construct as effective a method as possible to predict future observations. ★**Correlation**

- **Feature selection and risk property**: To gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction method. ★**Causation**

★Fan and Li (2006), Bickel (2008, JRSS-B)

Impact of Big Data

- **Data Acquisition**: Multiple platforms, bias sampling, experimental variations, measurement errors.
- **Data Management**: Storage, memory, preprocessing, queries.
- **Computing infrastructure**: distributed file systems and cloud computing
- **Computation**: new paradigms on optimization and computing: high-performance and parallel computing.
- **Data analysis**: Noise accumulation, spurious correlations, incidental endogeneity, measurement errors, and heterogeneity.

Impact of Big Data

- **Data Acquisition**: Multiple platforms, bias sampling, experimental variations, measurement errors.
- **Data Management**: Storage, memory, preprocessing, queries.
- **Computing infrastructure**: distributed file systems and cloud computing
- **Computation**: new paradigms on optimization and computing: high-performance and parallel computing.
- **Data analysis**: Noise accumulation, spurious correlations, incidental endogeneity, measurement errors, and heterogeneity.

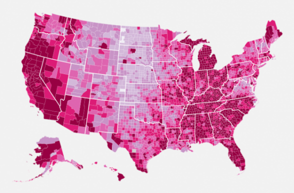
Impact of Big Data

- **Data Acquisition**: Multiple platforms, bias sampling, experimental variations, measurement errors.
- **Data Management**: Storage, memory, preprocessing, queries.
- **Computing infrastructure**: distributed file systems and cloud computing
- **Computation**: new paradigms on optimization and computing: high-performance and parallel computing.
- **Data analysis**: Noise accumulation, spurious correlations, incidental endogeneity, measurement errors, and heterogeneity.

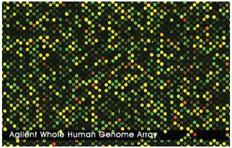
Are our assumptions verifiable?

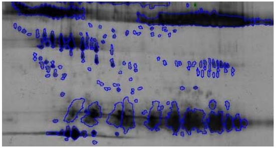
Analysis of High-dim Data

Collect data: e.g. Unemployment rates



Bioinformatic: disease classs. / clinical outcomes w/ "-omics" data.





Regularization: Use PLS (Lasso & Scad) to get $\mathcal{S}_0$ and β_0.

Done!

Key Assumptions: Exogeneity

Stylized Model: $Y = \mathbf{X}^T \beta_0 + \varepsilon$, **$\beta_0$ sparse**

$$E\varepsilon\mathbf{X} = 0 \quad \textbf{or} \quad E(\varepsilon|\mathbf{X}) = 0$$

There are tens of thousand of equations!

- Related to identifiability!

Are X_j and $\hat{\varepsilon}$ uncorrelated?

What consequence if not?

How to do it right?

Are X_j and $\hat{\varepsilon}$ uncorrelated?

What consequence if not?

How to do it right?

Are X_j and $\hat{\varepsilon}$ uncorrelated?

What consequence if not?

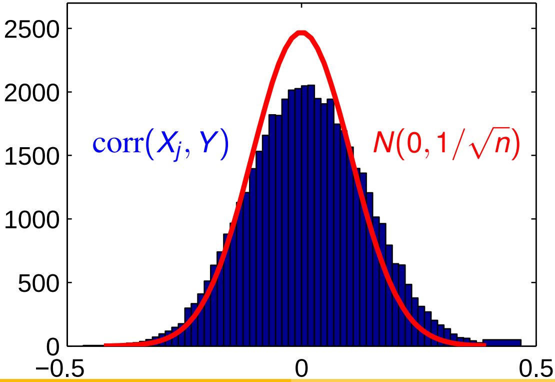
How to do it right?

Example: Distribution of correlations

Data: 90 western Europeans from 'HapMap' project

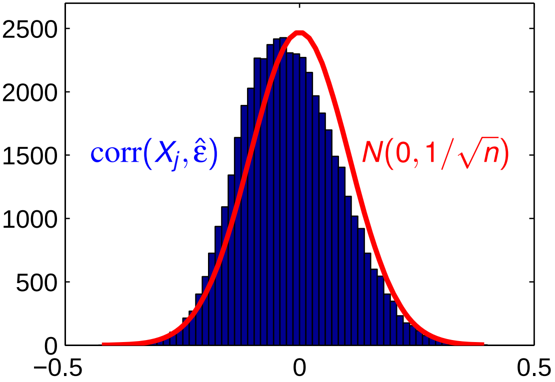
Response: expressions of *CHRNA6*, cholinergic receptor, nicotinic, alpha 6 (554 SNPs within 1MB).

Covariates: All other expressions ($p = 47292$)



Validating Exogeneity Assumption

Lasso: Select 23 variables.



Moral: High-dimensionality is a source of incidental endogeneity

Incidental Endogeneity

An Illustration

True model: $Y = 2X_1 + X_2 + \varepsilon$, $\quad \mathrm{corr}(X_1, \varepsilon) = 0, \mathrm{corr}(X_2, \varepsilon) = 0$

Netting: Collecting many variables $\{X_j\}_{j=1}^p$.
Incidentally,

$$\mathrm{corr}(X_j, \underbrace{Y - 2X_1 - X_2}_{\varepsilon}) \neq 0. \quad \textbf{Endogeneity}$$

■Many X_j's related to Y, hence to ε incidentally due to **large** p.

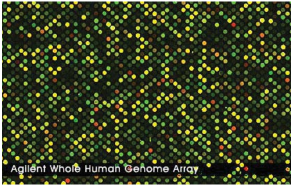
High dim causes incidental endogeneity

Outcome: $Y =$ clinical, biological, or health, credit

Exogenous model: $Y = \underbrace{\mathbf{X}_{S_0}^T \beta_0 + \varepsilon}_{E(\varepsilon|\mathbf{X}_{S_0})=0}$, **unknown S_0**. collect many

e.g. gene expressions

e.g. microecon/risk factors, **related to Y**



Hard to make: $E\underbrace{\left(Y - \mathbf{X}_{S_0}^T \beta_0\right)}_{\varepsilon} X_j = 0$ **for all j**

H_1: high-dim causes endogeneity

Any tools to test?

What are verifiable assumptions?

H_1: high-dim causes endogeneity

Any tools to test?

What are verifiable assumptions?

H_1: high-dim causes endogeneity

Any tools to test?

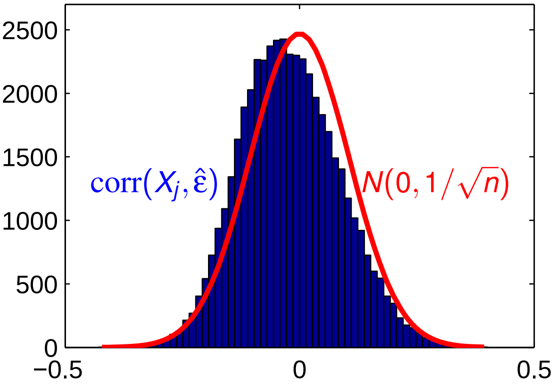
What are verifiable assumptions?

Test against Exogeneity

Raw Materials and Visualization

Raw materials: Residuals $\hat{\varepsilon}$ after regularized fit:

$\{\mathbf{r_j} = \mathrm{corr}(\hat{\varepsilon}, \mathbf{X_j})\}_{j=1}^{p}$ Visualized by histogram



Example: Apply Lasso to 'HapMap' project data

Test statistics and null distributions

■What is **null dist.** of the histogram? $N(0, 1/\sqrt{n})$?

★**KS test**: $T_1 = \|\hat{F}_n(x) - F_0(x)\|_\infty$,

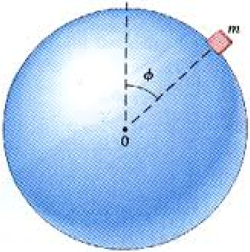
★**CVM test** $T_2 = \|\hat{F}_n(x) - F_0(x)\|_2^2$.

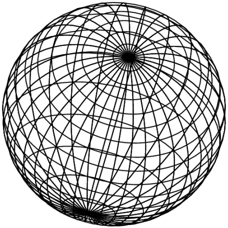
■What are the null distributions when p is large?

What is new: $\{\mathbf{X}_j\}_{j=1}^p$ are **correlated**!

Relation to random geometry

■What is the empirical dist of angles between p **random points** on the **n-dim** unit sphere and the north pole?





■What are the dist. of the min angle or ave angle?

■See Cai, Fan, and Jiang (13) for both large n and small n when $p \to \infty$, but for **independent** random points.

Other test statistics

$$T_3 = p^{-1} \sum_{j=1}^{p} r_j^q, \qquad\qquad T_4 = \max_{1 \le j \le p} |r_j|$$

- They are empirical q-th moment and ∞-moment of $\hat{F}_n(x)$, corresponding to the ave ($q = 1$) and min angles.
- ★ More powerful for a small fraction of departures, but can not give an estimate of the proportion of violations.
- Their distributions under depend. covariates.

Consequence of Endogeneity

Consequence of Endogeneity

■Necessary condition for any PLS consistent is **exogeneity:** $EX_j\varepsilon = 0, \forall j$ (*Fan and Yuan, 14*).

Scientific Implications: Can choose wrong sets of genes or SNPs using LASSO/SCAD in **presence of endogeneity**.

■Related to model identifiability, e.g.

$$\begin{aligned} Y &= 2X_1 + X_2 + \varepsilon, & EX_1\varepsilon = EX_2\varepsilon = 0 \\ &= a_3X_3 + a_4X_4 + a_5X_5 + \varepsilon^*, & EX_j\varepsilon^* = 0, j = 3,4,5. \end{aligned}$$

Consequence of Endogeneity

■Necessary condition for any PLS consistent is **exogeneity:** $EX_j\varepsilon = 0, \forall j$ (*Fan and Yuan, 14*).

Scientific Implications: Can choose wrong sets of genes or SNPs using LASSO/SCAD in **presence of endogeneity**.

■Related to model identifiability, e.g.

$$\begin{aligned} Y &= 2X_1 + X_2 + \varepsilon, & EX_1\varepsilon = EX_2\varepsilon = 0 \\ &= a_3X_3 + a_4X_4 + a_5X_5 + \varepsilon^*, & EX_j\varepsilon^* = 0, j = 3, 4, 5. \end{aligned}$$

Simulation Results

True model: $\beta_S^0 = (5, -4, 7, -1, 1.5)$, $\mathbf{Z} \sim N(0, \boldsymbol{\Sigma}), \sigma_{ij} = 0.5^{|i-j|}$

$X_j = Z_j$ for $j \leq 100$(exogenous), $\quad X_j = (Z_j + 5)(\varepsilon + 1)$, (endogenous).

- $n = 200$, $p = 300$, 100 replicates.

	PLS		FGMM			
	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.1$	post-FGMM	$\lambda = 0.2$	post-FGMM
MSE_S	0.278	0.712	**0.215**	**0.190**	**0.241**	**0.188**
MSE_N	0.541	0.118	0.018		0.006	
TP-Mean	5	4.733	5		4.97	
FP-Mean	206.26	31.14	3.56		3.58	

Verifiable Assumptions

Low dimensional assumption

Model selection consistency under

$$Y = \mathbf{X}_{S_0}^T \beta_0 + \varepsilon, \qquad \mathbf{E}(\varepsilon | \mathbf{X}_{\mathbf{S_0}}) = \mathbf{0}$$

or weaker, e.g. $E\mathbf{X}_{S_0}\varepsilon = 0$, $E\mathbf{X}_{S_0}^2\varepsilon = 0$.

- Easier to validate: only $2|S_0|$ correlations to be validated.
- Use over-identification to screen endogeneious variables: FGMM (*Fan&Liao, 14*)

Low dimensional assumption

Model selection consistency under

$$Y = \mathbf{X}_{S_0}^T \beta_0 + \varepsilon, \qquad \mathbf{E}(\varepsilon | \mathbf{X}_{\mathbf{S_0}}) = \mathbf{0}$$

or weaker, e.g. $E\mathbf{X}_{S_0}\varepsilon = 0$, $E\mathbf{X}_{S_0}^2\varepsilon = 0$.

- Easier to validate: only $2|S_0|$ correlations to be validated.
- Use over-identification to screen endogeneious variables: FGMM (*Fan&Liao, 14*)

Focussed GMM

focused on endogeneity screening by

$$L_{\text{FGMM}}(\beta) = \left\| \frac{1}{n} \sum_{i=1}^{n} \overbrace{(Y_i - \mathbf{X}_{S,i}^T \beta_S)}^{\varepsilon_i} \begin{pmatrix} \mathbf{X}_{S,i} \\ f(\mathbf{X}_{S,i}) \end{pmatrix} \right\|_w .$$

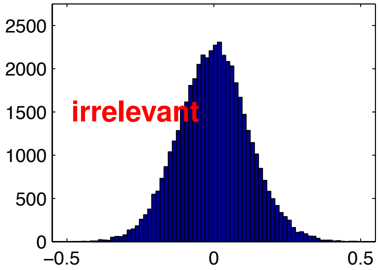
Example: $f(x) = x^2$ or $f(x) = |x - \bar{x}|$

Over-identification Condition: Any $\mathcal{S} \supset$ **endogenous** var.

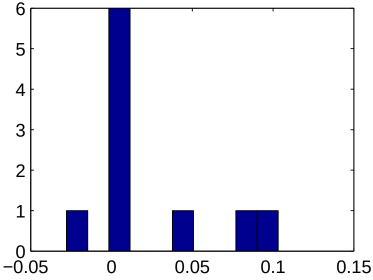
$$\min_{\beta_{\mathcal{S}}} \| \underbrace{E(Y - \mathbf{X}_{\mathcal{S}}^T \beta_{\mathcal{S}}) \mathbf{X}_{\mathcal{S}}}_{|\mathcal{S}| \text{ equations}} \|^2 + \| \underbrace{E(Y - \mathbf{X}_{\mathcal{S}}^T \beta_{\mathcal{S}}) f(\mathbf{X}_{\mathcal{S}}^2)}_{|\mathcal{S}| \text{ equations}} \|^2 \geq c.$$

Example: Hap Map Data

$\mathrm{corr}(X_j, \hat{\varepsilon}), \forall j$



$\{\mathrm{corr}(X_{S_0}, \hat{\varepsilon}), \mathrm{corr}(X^2_{S_0}, \hat{\varepsilon})\}$



FGMM fit using $EX_{S_0}\varepsilon = 0, EX^2_{S_0}\varepsilon = 0$. 5 genes selected.

Comparison of models

	No Fitting	Lasso	FGMM
# of parameters	1	23+1	5+1
AIC	**-2.289**	-2.883	**-2.807**
BIC	**-2.261**	-2.216	**-2.640**
RIC	**-2.070**	**2.324**	**-1.503**

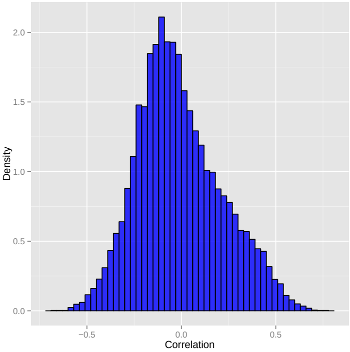
■ RIC (penalty = $2 \log p$) (*Foster and George, 94*) favors even more to the FGMM fit.

Data: 148 microarrays from GEO database and ArrayExpress.

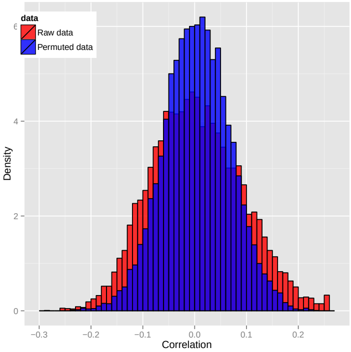
Response: expressions of gene *DDR1* (encodes receptor tyrosine kinases, related to the prostate cancer)

Covariates: remaining 12,718 genes

(a) Distribution of $\widehat{\mathrm{Corr}}(Y, X_j)$



(b) Distribution of $\widehat{\mathrm{Corr}}(X_j, \hat{\varepsilon})$

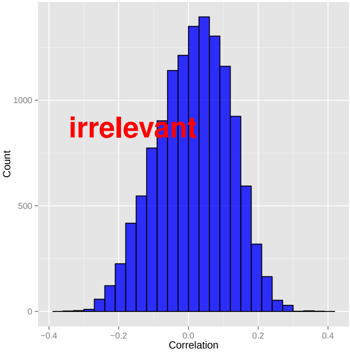


FGMM fit and diagnostics

Fitting: **FGMM** based on $EX_{S_0}\varepsilon = 0$, $EX^2_{S_0}\varepsilon = 0$.

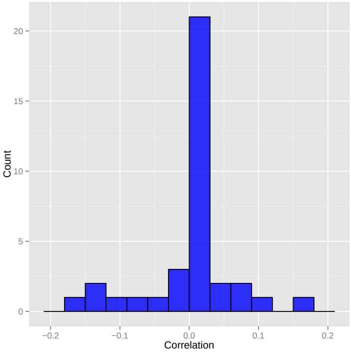
$\mathrm{corr}(X_j, \hat{\varepsilon}), \forall j$

(a) Distribution of residuals and genes



$\{\mathrm{corr}(X_{S_0}, \hat{\varepsilon}), \mathrm{corr}(X^2_{S_0}, \hat{\varepsilon})\}$

(b) Distribution of residuals and selected genes



Conclusion

★ High dimensionality is a source of endogeneity.

★ Endogeneity results in model selection inconsistency and parameter un-identifiability.

★ Exog. cond in high-dim is unrealistic and needs validation.

★ Exogeneity assumption should **NOT** be made on "unimportant variables".

★ FGMM can deliver model selection consistency under more realistic and verifiable assumptions.

Conclusion

- ★ High dimensionality is a source of endogeneity.
- ★ Endogeneity results in model selection inconsistency and parameter un-identifiability.
- ★ Exog. cond in high-dim is unrealistic and needs validation.
- ★ Exogeneity assumption should **NOT** be made on “unimportant variables”.
- ★ FGMM can deliver model selection consistency under more realistic and verifiable assumptions.

Conclusion

★ High dimensionality is a source of endogeneity.

★ Endogeneity results in model selection inconsistency and parameter un-identifiability.

★ Exog. cond in high-dim is unrealistic and needs validation.

★ Exogeneity assumption should **NOT** be made on "unimportant variables".

★ FGMM can deliver model selection consistency under more realistic and verifiable assumptions.

Conclusion

- ★ High dimensionality is a source of endogeneity.
- ★ Endogeneity results in model selection inconsistency and parameter un-identifiability.
- ★ Exog. cond in high-dim is unrealistic and needs validation.
- ★ Exogeneity assumption should **NOT** be made on “unimportant variables”.
- ★ FGMM can deliver model selection consistency under more realistic and verifiable assumptions.

The End

Thank

You

FDR Control under Dependency

Jianqing Fan

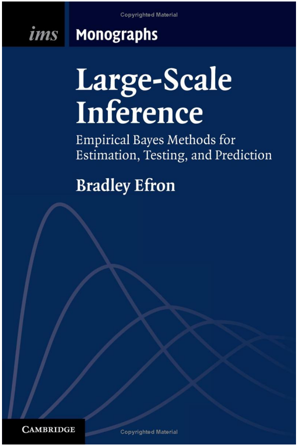
Princeton University

With **Xu Han**

May 28, 2014

Outline

1. Background
2. Principal Factor Approximation
3. FDP with Unknown Covariance
4. Numerical properties



Background

Large-Scale Multiple Testing

★ **Biology, Medicine, Genetics, Neuroscience**:

- analysis of high throughput data: genes, proteins, copy No.
- genome-wide association studies— SNPs w/ phenotype (e.g. weight, diseases, QTL) or gene expression (eQTL).
- detecting activated voxels after stimulii.

★ **Finance, Economics**: Find fund managers who have winning ability (*Barras, Scaillet & Wermers, 10*).

★ **Network and graphical models**: Detecting zero-corr patterns.

Statement of Problems

Problem: Given test statistics $Z_i \sim N(\mu_i, 1)$, wish to test

$$H_{0i} : \mu_i = 0 \quad vs \quad H_{1i} : \mu_i \neq 0, \qquad i = 1, \cdots, p.$$

★large p and sparse μ.

Dependence: $\mathbf{Z} \sim N_p(\mu, \Sigma)$, **unknown** Σ

Aim 1: ★**Consistent** estimation of False Discovery Proportion (FDP)

Aim 2: ★Improve the **power**.

Statement of Problems

Problem: Given test statistics $Z_i \sim N(\mu_i, 1)$, wish to test

$$H_{0i} : \mu_i = 0 \quad vs \quad H_{1i} : \mu_i \neq 0, \qquad i = 1, \cdots, p.$$

★large p and sparse μ.

Dependence: $\mathbf{Z} \sim N_p(\mu, \Sigma)$, **unknown** Σ

Aim 1: ★**Consistent** estimation of False Discovery Proportion (FDP)

Aim 2: ★Improve the **power**.

Dependent and Independence Tests

Discoveries: $\{j : |Z_j| > t\}$ for a critical value t. **Total** $= R(t)$**.**

False Discoveries: $\mathbf{V(t)} = \#$ of true nulls with $|Z_j| > t$.

Proportion: **FDP**$(t) = V(t)/R(t)$, $V(t)$ **unobservable** r.v.

Indep tests: $\mathrm{FDP}(t) \approx p_0 G(t)/R(t)$, a.s. $\star G(t) = P(|Z_i| > t)$.

Dep tests: $\mathrm{FDP}(t)$ varies from data to data. (*Owen, 05, Efron, 07, 10, Fan et al, 12*)

Dependent and Independence Tests

Discoveries: $\{j : |Z_j| > t\}$ for a critical value t. **Total** $= R(t)$.

False Discoveries: $\mathbf{V(t)}$ = # of true nulls with $|Z_j| > t$.

Proportion: **FDP**$(t) = V(t)/R(t)$, $V(t)$ **unobservable** r.v.

Indep tests: FDP$(t) \approx p_0 G(t)/R(t)$, a.s. ★$G(t) = P(|Z_i| > t)$.

Dep tests: FDP(t) varies from data to data. (*Owen, 05, Efron, 07, 10, Fan et al, 12*)

An illustrative example

Equi-corr: $Z_i = \mu_i + \sqrt{\rho}W + \sqrt{1-\rho}\varepsilon_i,$ $\qquad W, \varepsilon_i \sim_{indep} N(0,1)$

Number of FD: $V(t) = \sum_{i=1}^{p_0} I(Z_i > t)$ $\qquad$ (one-sided tests)

Indep: $V(t) \approx p_0 \Phi(-t) = 22.8,$ $\qquad$ if $p_0 = 1000$, $t = 2$

Dependence: $\rho = 0.64$: $\qquad$ F-adj

$$V(t) = \sum_{i \in \text{null}} I(0.8W + 0.6\varepsilon_i > t) \approx p_0 \Phi\left(-\frac{t - 0.8W}{0.6}\right)$$

An illustrative example

Equi-corr: $Z_i = \mu_i + \sqrt{\rho} W + \sqrt{1-\rho}\varepsilon_i,$ $W, \varepsilon_i \sim_{indep} N(0,1)$

Number of FD: $V(t) = \sum_{i=1}^{p_0} I(Z_i > t)$ (one-sided tests)

Indep: $V(t) \approx p_0 \Phi(-t) = \mathbf{22.8},$ if $p_0 = 1000$, $t = 2$

Dependence: $\rho = 0.64$: F-adj

$$V(t) = \sum_{i \in \text{null}} I(0.8W + 0.6\varepsilon_i > t) \approx p_0 \Phi\left(-\frac{t - 0.8W}{0.6}\right)$$

Equiv-correlation (continued)

Number of False Discoveries:

1. $W = 0 \Longrightarrow V(t) \approx \mathbf{0.43}$ $\qquad W = 1 \Longrightarrow V(t) \approx \mathbf{22.8}$.
2. $W = 2 \Longrightarrow V(t) \approx \mathbf{252.5}$ $\qquad W = 3 \Longrightarrow V(t) \approx \mathbf{747.5}$.

★ Depends **sensitively** on realization of W;

★ **Consistently estimable**: $W = \bar{Z}/.8 + O_p(1/\sqrt{p})$ and

$$p_0 \Phi\left(-\frac{t - 0.8\hat{W}}{0.6}\right)/R(t), \qquad \hat{W} = \bar{Z}/.8$$

fdpa

Equiv-correlation (continued)

Number of False Discoveries:

1. $W = 0 \Longrightarrow V(t) \approx \mathbf{0.43}$ $\qquad W = 1 \Longrightarrow V(t) \approx \mathbf{22.8}$.
2. $W = 2 \Longrightarrow V(t) \approx \mathbf{252.5}$ $\qquad W = 3 \Longrightarrow V(t) \approx \mathbf{747.5}$.

★ Depends **sensitively** on realization of W;

★ **Consistently estimable**: $W = \bar{Z}/.8 + O_p(1/\sqrt{p})$ and

$$p_0 \Phi\left(-\frac{t - 0.8\hat{W}}{0.6}\right)/R(t), \qquad \hat{W} = \bar{Z}/.8$$

fdpa

Related Literature

★ **Weak Dependence**: Benjamini & Hochberg (95), Storey (02), Storey, Taylor & Siegmund (04); Genovese & Wasserman (02, 06), vande Laan, 04; Lehmann and Romano, 05; Romano and Wolf (07), ...

★ **Applicable to Dependence**: Benjamini & Yekutieli (01), Clarke and Hall (2009), Sun & Cai (2009), Liu and Shao (12)...

★ **Use of Dependence**: Efron (07, 10), Leek & Storey (08), Friguet, Kloareg & Causeur (09), Schwartzman (10), Fan, Han, and Gu, 12,...

■ **Not** necessarily a consistent estimate of FDP.

Principal Factor Approximation

Known Dependence

Fan, Han and Gu (2012, JASA)

Estimating Principal Factor

Test Statistics: $\mathbf{Z} \sim N(\mu, \Sigma)$, $\mathrm{diag}(\Sigma) = 1$.

SVD: $\Sigma = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T = \mathbf{B}\mathbf{B}^\mathbf{T} + \mathbf{A}$. Σ known.

★$\mathbf{B} = (\sqrt{\lambda_1}\gamma_1, \cdots, \sqrt{\lambda_k}\gamma_k)$, $\mathbf{A} =$ residual matrix.

Decomposition: $\mathbf{Z} = \mu + \mathbf{B}\mathbf{W} + \mathbf{K}$ $\mathbf{W} \sim N(0, I_k)$ and $\mathbf{K} \sim N(0, \mathbf{A})$.

Realized Principal Factors: $\min_{\mu,w} \|\mathbf{Z} - \mu - \mathbf{B}\mathbf{W}\|^2 + \lambda\|\mu\|_1$

(same as Huber-ψ) or simply L_1-fit: $\min_w \|\mathbf{Z} - \mathbf{B}\mathbf{W}\|_1$.

Estimating Principal Factor

Test Statistics: $\mathbf{Z} \sim N(\mu, \Sigma)$, $\quad$ $\mathrm{diag}(\Sigma) = 1$.

SVD: $\Sigma = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T = \mathbf{B}\mathbf{B}^{\mathbf{T}} + \mathbf{A}$. $\quad$ Σ known.

★ $\mathbf{B} = (\sqrt{\lambda_1}\gamma_1, \cdots, \sqrt{\lambda_k}\gamma_k)$, $\quad$ $\mathbf{A}$ = residual matrix.

Decomposition: $\mathbf{Z} = \mu + \mathbf{B}\mathbf{W} + \mathbf{K}$ $\quad$ $\mathbf{W} \sim N(0, I_k)$ and $\mathbf{K} \sim N(0, \mathbf{A})$.

Realized Principal Factors: $\min_{\mu,w} \|\mathbf{Z} - \mu - \mathbf{B}\mathbf{W}\|^2 + \lambda\|\mu\|_1$

(same as Huber-ψ) or simply L_1-fit: $\min_w \|\mathbf{Z} - \mathbf{B}\mathbf{W}\|_1$.

Estimation of FDP

Input: test statistics $\mathbf{Z} \sim N(\mu, \Sigma)$ **Available in R**

1. **SVD:** $\Sigma = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T = \mathbf{B}\mathbf{B}^\mathbf{T} + \mathbf{A}$

2. **Estimating factors**: $\min_w \|\mathbf{Z} - \mathbf{B}\mathbf{W}\|_1$

3. **Estimation of FDP**: $\widehat{\mathrm{FDP}}(t) = \frac{\sum_{j=1}^{p} P(\hat{\eta}_i, t)}{R(t)}$. exam

★ $P(\eta_i, t) = P_{null}\{|Z_i| > t | \mathbf{W}\}$

- $= \Phi(a_i(z_{t/2} + \eta_i)) + \Phi(a_i(z_{t/2} - \eta_i))$,
- $\eta_i = \mathbf{b}_i^T \mathbf{W}$, $\mathbf{b}_i = i^{th}$ row of $\mathbf{B}$ $a_i = (1 - \|\mathbf{b}_i\|^2)^{-1/2}$.

Estimation of FDP

Input: test statistics $\mathbf{Z} \sim N(\mu, \Sigma)$ **Available in R**

1. **SVD:** $\Sigma = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^T = \mathbf{B}\mathbf{B}^{\mathbf{T}} + \mathbf{A}$
2. **Estimating factors**: $\min_w \|\mathbf{Z} - \mathbf{B}\mathbf{W}\|_1$
3. **Estimation of FDP**: $\widehat{\mathrm{FDP}}(t) = \frac{\sum_{j=1}^{p} P(\hat{\eta}_i, t)}{R(t)}$. exam

★ $P(\eta_i, t) = P_{null}\{|Z_i| > t|\mathbf{W}\}$

- $= \Phi(a_i(z_{t/2} + \eta_i)) + \Phi(a_i(z_{t/2} - \eta_i))$,
- $\eta_i = \mathbf{b}_i^T \mathbf{W}$, $\mathbf{b}_i = i^{th}$ row of $\mathbf{B}$ $a_i = (1 - \|\mathbf{b}_i\|^2)^{-1/2}$.

Related to Efron (2010)

- **Gram-Charlier**: $V(t) = \phi(t) - \sum_{j=1}^{\infty} (-1)^j \frac{A_j}{j!} \phi^{(j-1)}(t)$
  $A_j \sim ID(0, \alpha_j)$ with $\alpha_j = \sum_{i \neq \imath'} \mathrm{cor}(Z_i, Z_i')^j$ (*Schwartzman, 10*)
- Efron takes $j = 2$ in computing $E(V(t)|A)$.
- Basis function (Hermit polynomial) expansion vs singular value decomposition.
- Different methods in estimating A's and W's

Consistency and Rate of Convergence

False discoveries: $V(t) = \sum_{i \in \text{true null}} P(\eta_i, t) + o(p)$

Theorem: $\mathrm{FDP}(t) - \mathrm{FDP}_A(t) = o_p(1)$, $\mathrm{FDP}_A(t) = \frac{\sum_{j=1}^{p} P(\eta_i, t)}{R(t)}$,

if $p^{-1}(\lambda_{k+1}^2 + \cdots + \lambda_p^2)^{1/2} \longrightarrow 0$.

- If $\lambda_{\max} = o(p^{1/2})$, we can take $k = 0$ $\Longrightarrow$ **independence**
- Convergence rate: $o_p(p^{-\delta/2})$ if $p^{-1}(\lambda_{k+1}^2 + \cdots + \lambda_p^2)^{1/2} = p^{-\delta}$.

Accuracy: $|\widehat{\mathrm{FDP}}(t) - \mathrm{FDP}_A(t)| = O_p\left(\|\hat{\mathbf{W}} - \mathbf{W}\|\right)$.

Consistency and Rate of Convergence

False discoveries: $V(t) = \sum_{i \in \text{true null}} P(\eta_i, t) + o(p)$

Theorem: $\mathrm{FDP}(t) - \mathrm{FDP}_A(t) = o_p(1)$, $\quad \mathrm{FDP}_A(t) = \frac{\sum_{j=1}^p P(\eta_i, t)}{R(t)}$,

if $p^{-1}(\lambda_{k+1}^2 + \cdots + \lambda_p^2)^{1/2} \longrightarrow 0$.

- If $\lambda_{\max} = o(p^{1/2})$, we can take $k = 0$ $\Longrightarrow$ **independence**
- Convergence rate: $o_p(p^{-\delta/2})$ if $p^{-1}(\lambda_{k+1}^2 + \cdots + \lambda_p^2)^{1/2} = p^{-\delta}$.

Accuracy: $|\widehat{\mathrm{FDP}}(t) - \mathrm{FDP}_A(t)| = O_p\left(\|\hat{\mathbf{W}} - \mathbf{W}\|\right)$.

Estimated vs true FDP (Simulation results)

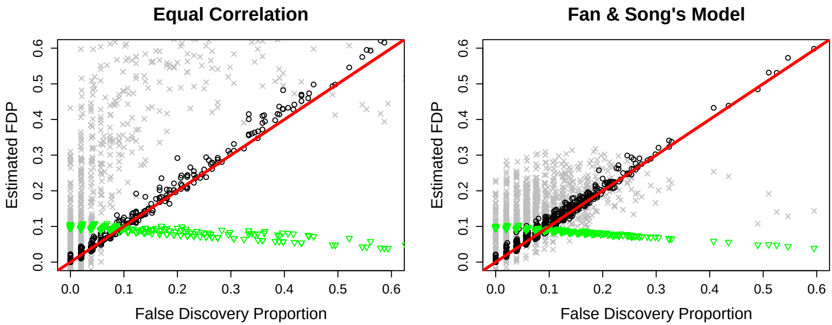


Figure: $p = 1000$, $p_1 = 50$, $n = 100$, $t = 2.8$, nonzero $\beta_i = 1$, $N_{sim} = 1000$.

★cross = Efron's approach; ★circle = PFA

★green = Storey's (2002) estimate $pt/R(t)$

Additional simulation results

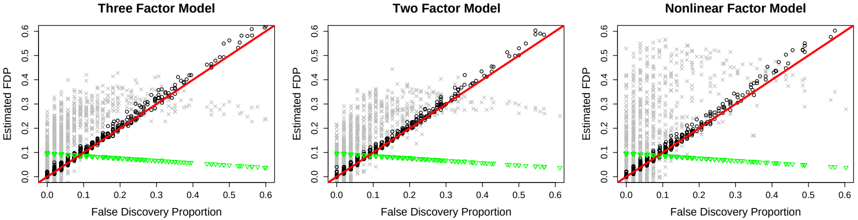


Figure: $p = 1000$, $p_1 = 50$, $n = 100$, $t = 2.8$, nonzero $\beta_i = 1$, $N_{sim} = 1000$.

Factor adjusted method

Conventional methods: Rank determined by $|Z_i|$, not ideal for dependent data. Note that

$$Z_i - \mathbf{b}_i^T \mathbf{W} \sim N(\mu_i, 1 - \|\mathbf{b}_i\|^2),$$

Factor-adjusted method: Use the new test statistics

$$Y_i = a_i(Z_i - \mathbf{b}_i^T \widehat{\mathbf{W}}) \sim N(a_i\mu_i, 1)$$ exam

- Increase signal-noise ratio $a_i = (1 - \|\mathbf{b}_i\|^2)^{-1/2} \geq 1$
- Rank determined by $|Y_i|$, **NOT** $|Z_i|$.

Factor adjusted method

Conventional methods: Rank determined by $|Z_i|$, not ideal for dependent data. Note that

$$Z_i - \mathbf{b}_i^T \mathbf{W} \sim N(\mu_i, 1 - \|\mathbf{b}_i\|^2),$$

Factor-adjusted method: Use the new test statistics

$$Y_i = a_i(Z_i - \mathbf{b}_i^T \widehat{\mathbf{W}}) \sim N(a_i\mu_i, 1)$$

exam

- Increase signal-noise ratio $a_i = (1 - \|\mathbf{b}_i\|^2)^{-1/2} \geq 1$
- Rank determined by $|Y_i|$, **NOT** $|Z_i|$.

FDP with Unknown Dependence

Two Questions

- What accuracy of $\hat{\Sigma}$ needed for the plug-in method to work?
- What structures of Σ lead to such an accuracy?

Aim: Investigate the required eigen properties.

Two Questions

- What accuracy of $\hat{\Sigma}$ needed for the plug-in method to work?
- What structures of Σ lead to such an accuracy?

Aim: Investigate the required eigen properties.

Estimate FDP(t) under Unknown Dependence

0. **Estimating Σ**: Obtain an estimate $\hat{\Sigma}$.

1. **SVD**: $\hat{\Sigma} = \widehat{\mathbf{B}}\widehat{\mathbf{B}}^T + \widehat{\mathbf{A}}$.

   Recall $\mathbf{Z} = \mu + \mathbf{B}\mathbf{W} + K$. Run OLS ignore μ

2. **Estimate factor**: $\hat{\mathbf{W}} = (\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}'\mathbf{Z} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_k)^{-1}\widehat{\mathbf{B}}'\mathbf{Z}$.

3. **Estimated FDP**: Compute

$$\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) = \sum_{i=1}^{p}[\Phi(\widehat{a}_i(z_{t/2} + \widehat{\eta}_i)) + \Phi(\widehat{a}_i(z_{t/2} - \widehat{\eta}_i))]/R(t)$$

with $\widehat{a}_i = (1 - \|\widehat{\mathbf{b}}_i\|^2)^{-1/2}$ and $\widehat{\eta}_i = \widehat{\mathbf{b}}_i^T \widehat{\mathbf{w}}$.

Estimate FDP(t) under Unknown Dependence

0. **Estimating Σ**: Obtain an estimate $\hat{\Sigma}$.

1. **SVD**: $\hat{\Sigma} = \widehat{\mathbf{B}}\widehat{\mathbf{B}}^T + \widehat{\mathbf{A}}$.

   Recall $\mathbf{Z} = \mu + \mathbf{BW} + K$. Run OLS ignore μ

2. **Estimate factor**: $\hat{\mathbf{W}} = (\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}'\mathbf{Z} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_k)^{-1}\widehat{\mathbf{B}}'\mathbf{Z}$.

3. **Estimated FDP**: Compute

$$\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) = \sum_{i=1}^{p}[\Phi(\widehat{a}_i(z_{t/2} + \widehat{\eta}_i)) + \Phi(\widehat{a}_i(z_{t/2} - \widehat{\eta}_i))]/R(t)$$

with $\widehat{a}_i = (1 - \|\widehat{\mathbf{b}}_i\|^2)^{-1/2}$ and $\widehat{\eta}_i = \widehat{\mathbf{b}}_i^T \widehat{\mathbf{w}}$.

Accuracy of FDP(t) Estimation

Theorem 1: Under Conditions C1–C4, we have

$$|\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p(p^{-\delta} + kp^{-\kappa} + k\|\mu\|_2 p^{-1/2}).$$

(C1) $R(t)/p > H$ for some $H > 0$ as $p \to \infty$.

(C2) $\max_{i \le k} \|\widehat{\gamma}_i - \gamma_i\| = O_p(p^{-\kappa})$ for some $\kappa > 0$.

(C3) $\sum_{i=1}^{k} |\widehat{\lambda}_i - \lambda_i| = o_p(p^{1-\delta})$.

- $\sum_{i=1}^{k} |\widehat{\lambda}_i - \lambda_i| = \sum_{i=1}^{k} \lambda_i |\widehat{\lambda}_i/\lambda_i - 1| \le p \max_{i \le k} |\widehat{\lambda}_i/\lambda_i - 1|$.

Accuracy of FDP(t) Estimation

Theorem 1: Under Conditions C1–C4, we have

$$|\widehat{\mathrm{FDP}}_{\mathrm{U}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p(p^{-\delta} + kp^{-\kappa} + k\|\mu\|_2 p^{-1/2}).$$

(C1) $R(t)/p > H$ for some $H > 0$ as $p \to \infty$.

(C2) $\max_{i \le k} \|\widehat{\gamma}_i - \gamma_i\| = O_p(p^{-\kappa})$ for some $\kappa > 0$.

(C3) $\sum_{i=1}^k |\widehat{\lambda}_i - \lambda_i| = o_p(p^{1-\delta})$.

■ $\sum_{i=1}^k |\widehat{\lambda}_i - \lambda_i| = \sum_{i=1}^k \lambda_i |\widehat{\lambda}_i/\lambda_i - 1| \le p \max_{i \le k} |\widehat{\lambda}_i/\lambda_i - 1|$.

Case I: Sparse Covariance Matrix

Conditions (C2) and (C3) hold if $\|\widehat{\Sigma}-\Sigma\| = O_p(p^{-\kappa})$ and $\lambda_i - \lambda_{i+1} \geq d > 0$ for $i \leq k$. (*Weyl theorem & Davis and Kahan theorem*)

- ★ Operator norm consistency is generally obtained under **sparse** structures (*Bickel and Levina, 08; Lam and Fan, 09; Cai and Liu, 11*).
- ★ No operator norm consistency for strong dependence (e.g. factor model).

Case II: Approximate Factor Model

Model: $\mathbf{y}_i = \mu + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i, \qquad i = 1, \cdots, n,$ $\qquad \Sigma_u$ **sparse**.

1. Run singular value decomposition: $\mathbf{S}_n = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T$.
2. Compute $\hat{\mathbf{R}} = \sum_{j=k+1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T$.
3. Apply (adaptive) thresholding:

$$\widehat{\mathbf{R}}^{\mathcal{T}} = (\hat{r}_{ij}^{\mathcal{T}}), \quad \hat{r}_{ij}^{\mathcal{T}} = \hat{r}_{ij} I(|\hat{r}_{ij}| \geq \tau_{ij})$$

4. Compute $\hat{\Sigma} = \sum_{j=1}^{k} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \widehat{\mathbf{R}}^{\mathcal{T}}$. (*POET, Fan, Liao, Mincheva, 13*)

Choice of k: Smallest k such that $\lambda_k > \varepsilon / \sqrt{p}$

Case II: Approximate Factor Model

Model: $\mathbf{y}_i = \mu + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i, \qquad i = 1, \cdots, n,$ $\qquad \Sigma_u$ **sparse**.

1. Run singular value decomposition: $\mathbf{S}_n = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T$.
2. Compute $\hat{\mathbf{R}} = \sum_{j=k+1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T$.
3. Apply (adaptive) thresholding:

$$\widehat{\mathbf{R}}^{\mathcal{T}} = (\hat{r}_{ij}^{\mathcal{T}}), \quad \hat{r}_{ij}^{\mathcal{T}} = \hat{r}_{ij} I(|\hat{r}_{ij}| \geq \tau_{ij})$$

4. Compute $\hat{\Sigma} = \sum_{j=1}^{k} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \widehat{\mathbf{R}}^{\mathcal{T}}$. (*POET, Fan, Liao, Mincheva, 13*)

■ **Choice of k**: Smallest k such that $\lambda_k > \varepsilon/\sqrt{p}$

Strong Dependence

Theorem 3: For approximate factor model, we have

$$|\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p(\delta_n) + O(k\|\mu\|_2 p^{-1/2}),$$

where $\delta_n = \sqrt{\frac{\log p}{n}} + \frac{1}{\sqrt{p}} + \sqrt{\frac{m_p}{p}} + \frac{p_1}{p}$, when k is finite.

- POET is accuracy enough for FPA.
- Obtained by an application of Fan, Liao and Mincheva (2013).

Strong Dependence

Theorem 3: For approximate factor model, we have

$$|\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t) - \mathrm{FDP}_{\mathrm{A}}(t)| = O_p(\delta_n) + O(k\|\mu\|_2 p^{-1/2}),$$

where $\delta_n = \sqrt{\frac{\log p}{n}} + \frac{1}{\sqrt{p}} + \sqrt{\frac{m_p}{p}} + \frac{p_1}{p}$, when k is finite.

- POET is accuracy enough for FPA.
- Obtained by an application of Fan, Liao and Mincheva (2013).

Simulation Studies

Simulation Setup

- **Model**: $\mathbf{y}_i = \mu + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i$ for $i = 1, \cdots, n$.
- **Components**: $\mathbf{f}_i \sim N_3(0, \mathbf{I}_3)$, $\mathbf{u}_i \sim N_p(0, \mathbf{I}_p)$, $\{\mathbf{u}_i\}_{t\geq 1}$ and $\{\mathbf{f}_i\}_{t\geq 1}$ indep.
- **Loadings**: $\mathbf{B}_{ij} \sim$ *i.i.d.* $U(-1, 1)$, then fixed.
- **Parameters**: $p = 1000$, $n = 500$, $p_1 = 50$, $t = 2.576$, nonzero $\mu_i = 1$ and $N_{sim} = 200$.
- **Purposes**: Compare $\widehat{\mathrm{FDP}}_{\mathrm{A}}(t)$ vs $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$.

Estimating FDP: $\widehat{\mathrm{FDP}}_{\mathrm{A}}(t)$ vs $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$

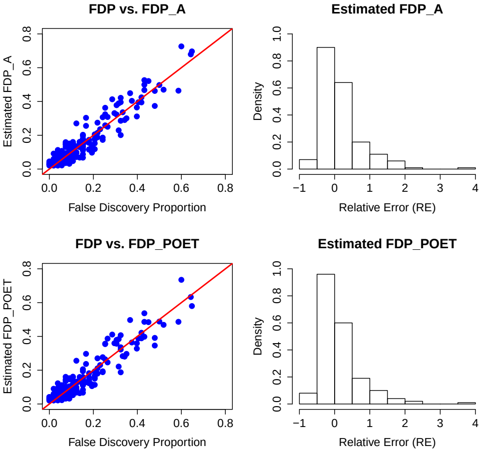


Figure: $\widehat{\mathrm{FDP}}_{\mathrm{A}}(t)$ is based on **known** Σ, $p = 1000$, $n = 500$, $p_1 = 50$, $t = 2.576$, $k = 3$, nonzero $\mu_i = 1$ and $N_{sim} = 200$. RE$= (\widehat{\mathrm{FDP}}(t) - \mathrm{FDP}(t))/\mathrm{FDP}(t)$.

Estimating FDP: LAD vs LS vs SCAD

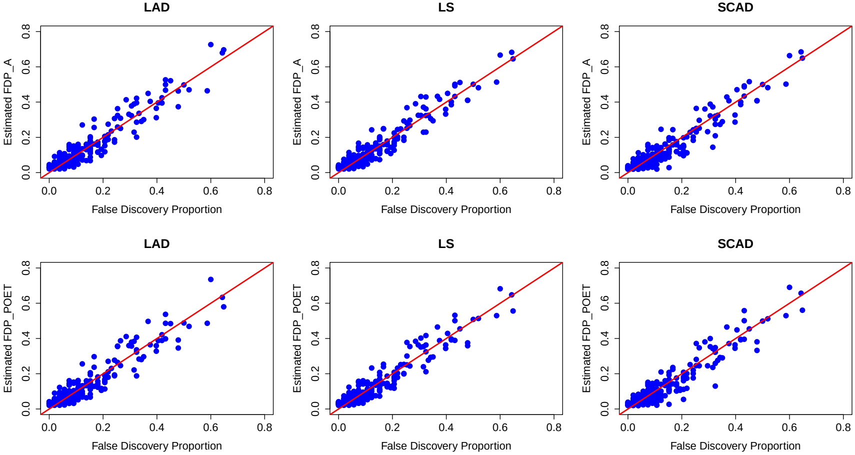


Figure: LAD (L_1), LS (L_2), SCAD (penalized L_2)

Accuracy of Estimating FDP

Table: Relative error between true $\mathrm{FDP}(t)$ and the estimators $\widehat{\mathrm{FDP}}_{\mathrm{A}}(t)$ and $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$ obtained by LAD, LS and SCAD.

	mean(RE_{A})	SD(RE_{A})	mean(RE_{P})	SD(RE_{P})
LAD	**0.1818**	0.5810	**0.1583**	0.5797
LS	0.1645	0.5398	0.1444	0.5413
SCAD	**0.0700**	0.5306	**0.0431**	0.5223

- RE_{A} and RE_{P} are the relative errors of $\widehat{\mathrm{FDP}}_{\mathrm{A}}(t)$ and $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$.

Estimating FDP: Nonnormality

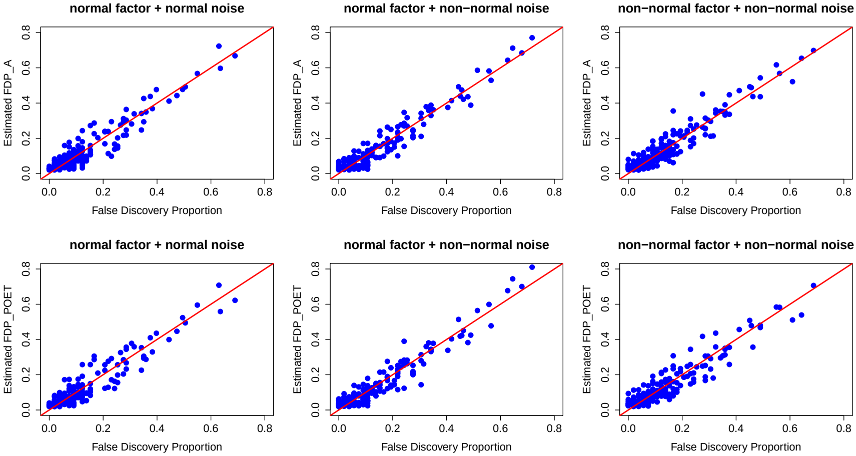


Figure: The non-normal distribution is *i.i.d.* standardized Student-*t* with DoF= 5.

Accuracy of Estimating FDP

Table: Relative error between true $\mathrm{FDP}(t)$ and the estimators $\widehat{\mathrm{FDP}}_{\mathrm{A}}(t)$ and $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$ under nonnormality.

	mean(RE_{A})	SD(RE_{A})	mean(RE_{P})	SD(RE_{P})
N-**f** + N-**u**	0.1708	0.6364	0.1660	0.6414
N-**f** + t-**u**	0.1146	0.5867	0.0908	0.5705
t-**f** + t-**u**	0.1637	0.6376	0.1388	0.6549

- RE_{A} and RE_{P} are the relative errors of $\widehat{\mathrm{FDP}}_{\mathrm{A}}(t)$ and $\widehat{\mathrm{FDP}}_{\mathrm{POET}}(t)$.

Real Data Analysis

Breast Cancer Study (Hedenfalk et al., 2001)

- ★ Two genetic mutations known to increase breast cancer risk: BRCA1 & BRCA2.
- ★ $n = 7$ BRCA1 women, $\mathbf{X}_1, \cdots, \mathbf{X}_n \sim N_p(\mu^X, \Sigma)$;
  $m = 8$ BRCA2 women, $\mathbf{Y}_1, \cdots, \mathbf{Y}_m \sim N_p(\mu^Y, \Sigma)$.
- ★ Microarray of expression levels on $p = 3226$ genes.

Two sample comparison: **BRCA1 $\equiv$ BRCA2?**

Test statistics: $\mathbf{Z} = \sqrt{nm/(n+m)}(\overline{\mathbf{X}} - \overline{\mathbf{Y}}) \sim N_p(\mu, \Sigma)$, with

$$\mu = \sqrt{nm/(n+m)}(\mu^X - \mu^Y).$$

Multiple hypothesis test:

$$H_{0j} : \mu_j = 0 \qquad \text{vs} \qquad H_{1j} : \mu_j \neq 0 \qquad j = 1, \cdots, p.$$

Gene Expression Heatmap: BRCA1 vs BRCA2

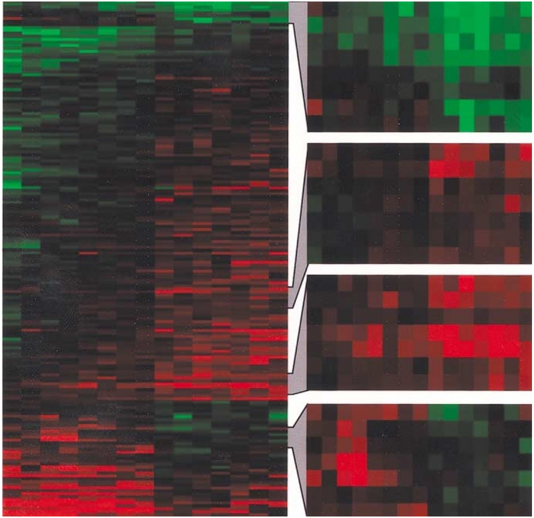


Figure: **Red** color means overexpression, while **green** color means underexpression.

$R(t)$, $\widehat{V}(t)$ and $\widehat{\text{FDP}}_{\text{POET}}(t)$

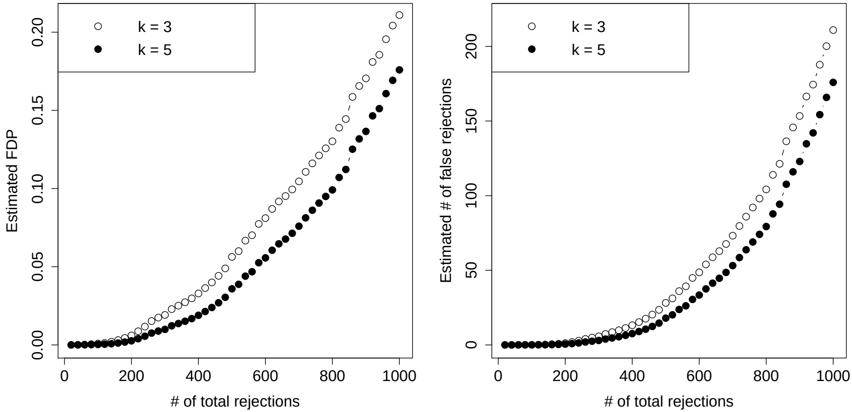


Figure: $\widehat{\text{FDP}}_{\text{POET}}(t)$ and $\widehat{V}(t)$ as functions of $R(t)$ for $p = 3226$ genes.

Summary

★ Derive asymptotic expression for FDP under arbitrary dependence;

★ Propose PFA to consistently estimate FDP when Σ unknown;

★ Establish asymptotic theory for the method;

★ Improve power properties by factor-adjustment;

★ Evaluate finite sample performance by extensive simulation studies.

Summary

★ Derive asymptotic expression for FDP under arbitrary dependence;

★ Propose PFA to consistently estimate FDP when Σ unknown;

★ Establish asymptotic theory for the method;

★ Improve power properties by factor-adjustment;

★ Evaluate finite sample performance by extensive simulation studies.

Acknowledgement

Thank  *You*

Robust Sparse Quadratic Discriminantion

Jianqing Fan

Princeton University

with **Tracy Ke, Han Liu and Lucy Xia**

May 26, 2014

Outline

1. Introduction
2. Rayleigh Quotient for sparse QDA
3. Optimization Algorithm
4. Application to Classification
5. Theoretical Results
6. Numerical Studies

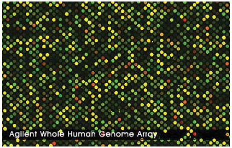
Introduction

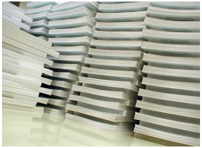
High Dimensional Classification

High-dimensional Classification

■ pervades all facets of machine learning and Big Data

- **Biomedicine**: disease classification / predicting clinical outcomes / biological process using microarray or proteomics data.



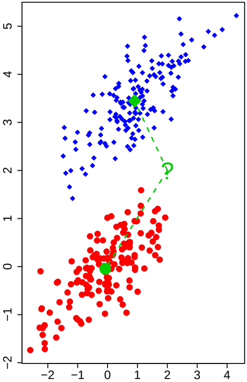



- **Machine learning**: Document/text classification, image classification

- **Social Networks**: Community detection

Classification

Training data: $\{\mathbf{X}_{i1}\}_{i=1}^{n_1}$ and $\{\mathbf{X}_{i2}\}_{i=1}^{n_2}$ for classes 1 and 2.



Aim: Classify a new data $\mathbf{X}$ by $I\{f(\mathbf{X}) < c\} + 1$

- Family of functions f: linear, quadratic
- Criterion for selecting f: logistic, hinge

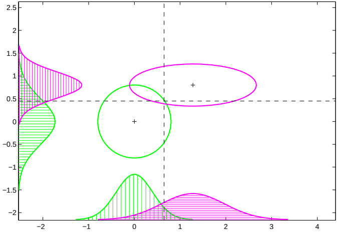
Convex surrogate

A popular approach

Sparse linear classifiers: Minimize classification errors (*Bickel& Levina, 04, Fan & Fan, 08; Shao et al. 11; Cai & Liu, 11; Fan, et al, 12*).

★Works well with **Gaussian** data with **equal** variance.

★Powerless if centroids are the same; no interaction considered



■Heteroscadestic variance? Non-Gaussian distributions?

Other popular approaches

- Plug-in quadratic discriminant.

  ★needs $\mathbf{\Sigma}_1^{-1}$, $\mathbf{\Sigma}_2^{-1}$; ★Gaussianity.

- Kernel SVM, logistic regression.

  ★inadequate use of dist.; ★few results; ★interactions

- Minimizing classification error:

  ★non-convex; not easily computable.

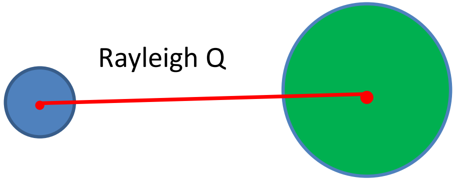
What new today?

1. Find a quadratic rule that max. Rayleigh Quotient.
2. Non-equal covariance matrices;
3. Fourth cross-moments avoided using elliptical distributions
4. Uniform estimation of means and variance for heavy-tails.

Rayleigh Quotient Optimization

Rayleigh Quotient

$$\mathrm{Rq}(f) = \frac{\text{between-class-var}}{\text{within-class-var}} \propto \frac{[\mathbb{E}_1 f(\mathbf{X}) - \mathbb{E}_2 f(\mathbf{X})]^2}{\pi \mathrm{var}_1[f(\mathbf{X})] + (1-\pi)\mathrm{var}_2[f(\mathbf{X})]}$$



- In the "classical" setting, $\mathrm{Rq}(f)$ is equiv. to $\mathrm{Err}(f)$
- In "broader" setting, it is a surrogate of classification error.
- Of independent scientific interest.

Rayleigh quotient for quadratic loss

Quadratic projection: $Q_{\Omega,\delta}(\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{X} - 2\delta^\top \mathbf{X}$.

With $\pi = \mathbb{P}(Y=1)$ and $\kappa = \frac{1-\pi}{\pi}$, we have

$$\mathrm{Rq}(Q) \propto \frac{[\mathbf{D}(\boldsymbol{\Omega},\delta)]^2}{\mathbf{V_1}(\boldsymbol{\Omega},\delta) + \kappa \mathbf{V_2}(\boldsymbol{\Omega},\delta)} = \mathrm{R}(\boldsymbol{\Omega},\delta),$$

- $D(\boldsymbol{\Omega},\delta) = \mathbb{E}_1 Q(\mathbf{X}) - \mathbb{E}_2 Q(\mathbf{X})$.
- $V_k(\boldsymbol{\Omega},\delta) = \mathrm{var}_k(Q(\mathbf{X})),\ k = 1, 2$.
- Reduce to **ROAD** (*Fan, Feng, Tong, 12*) when linear.

Challenge and Solution

Challenge: involve all fourth cross moments.

Solution: Consider the elliptical family.

$$\mathbf{X} = \mu + \xi \mathbf{\Sigma}^{1/2}\mathbf{U}, \qquad E\xi^2 = d, \quad \mathbf{X} \sim \mathcal{E}(\mu, \mathbf{\Sigma}, g)$$

Variance of Quadratic Form

$$\mathrm{var}(Q(\mathbf{X})) = 2(1+\gamma)\mathrm{tr}(\mathbf{\Omega\Sigma\Omega\Sigma}) + \gamma[\mathrm{tr}(\mathbf{\Omega\Sigma})]^2$$
$$+4(\mathbf{\Omega}\mu - \delta)^\top \mathbf{\Sigma}(\mathbf{\Omega}\mu - \delta), \qquad \textbf{quadratic in } \mathbf{\Omega}, \delta,$$

where $\gamma = \frac{E(\xi^4)}{d(d+2)} - 1$ *is the kurtosis parameter.*

Rayleigh Quotient under elliptical family

Semiparametric model: Two classes: $\mathcal{E}(\mu_1, \boldsymbol{\Sigma}_1, g)$ and $\mathcal{E}(\mu_2, \boldsymbol{\Sigma}_2, g)$.

D*, *V_1* and *V_2: involve only μ_1, μ_2, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ **and** γ

Examples of γ:

	Gaussian	t_ν	Contaminated Gaussian(ω, τ)	Compound Gaussian $U(1,2)$
γ	0	$\frac{2}{\nu-2}$	$\frac{1+\omega(\tau^4-1)}{(1+\omega(\tau^2-1))^2} - 1$	$\frac{1}{6}$

Sparse quadratic solution

Simplification: Using homogeneity,

$$\operatorname*{argmax}_{\Omega,\delta} \frac{[D(\Omega,\delta)]^2}{V_1(\Omega,\delta)+\kappa V_2(\Omega,\delta)} \propto \operatorname*{argmin}_{D(\Omega,\delta)=1} \underbrace{V_1(\Omega,\delta)+\kappa V_2(\Omega,\delta)}_{\mathbf{V}(\Omega,\delta)}$$

Sparsified version: $\Omega \in \mathbb{R}^{d\times d}, \delta \in \mathbb{R}^d$

$$\operatorname*{argmin}_{(\Omega,\delta):D(\Omega,\delta)=1} V(\Omega,\delta)+\lambda_1|\Omega|_1+\lambda_2|\delta|_1.$$

- Applicable to linear discriminant $\Longrightarrow$ ROAD

Robust Estimation and Optimization Algorithm

Robust Estimation of Mean

Problems: Elliptical distributions can have heavy tails.

Challenges: ★Sample median $\not\approx$ mean when skew (e.g. EX^2)
★Need uniform conv. for exponentially many σ_{ii}^2.

How to estimate mean with exponential concentration for heavy tails?

Robust Estimation of Mean

Problems: Elliptical distributions can have heavy tails.

Challenges: ★Sample median $\not\approx$ mean when skew (e.g. EX^2)
★Need uniform conv. for exponentially many σ_{ii}^2.

How to estimate mean with exponential concentration for heavy tails?

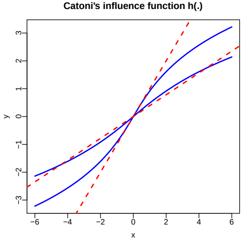
Catoni's M-estimator $\widehat{\mu}$

$$\sum_{i=1}^{n} h(\alpha_{n,d}(x_{ij} - \widehat{\mu}_j)) = 0, \qquad \alpha_{n,d} \to 0.$$

1. h strictly increasing: $\log(1-y+y^2/2) \le h(y) \le \log(1+y+y^2/2)$.
2. $\alpha_{n,d} = \left\{ \frac{4\log(n\vee d)}{n[v + \frac{4v\log(n\vee d)}{n-4\log(n\vee d)}]} \right\}^{1/2}$ with $v \ge \max_j \sigma_{jj}^2$.

$$|\widehat{\mu}_j - \mu_j|_\infty = O_p\left(\sqrt{\frac{\log d}{n}}\right)$$

needs bounded 2^{nd} moment



Robust Estimation of Σ_k

1. $\widehat{\eta}_j = \widehat{EX_j^2}$, Catoni's M-estimator using $\{x_{1j}^2, \cdots, x_{nj}^2\}$.

2. **variance estimation**: for a small δ_0,

$$\widehat{\sigma}_j^2 = \widehat{\Sigma}_{jj} = \max\{\widehat{\eta}_j - \widehat{\mu}_j^2, \delta_0\}.$$

3. **Off-diagonal elements**:

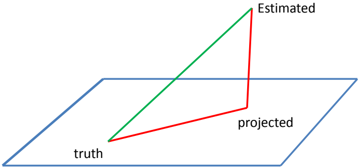
$$\widehat{\Sigma}_{jk} = \widehat{\sigma}_j \widehat{\sigma}_k \underbrace{\sin(\pi \widehat{\tau}_{jk}/2)}_{\textbf{robust corr}}$$

$\widehat{\tau}_{jk}$: Kendall's tau correlation (*Liu, et al, 12; Zou & Xue, 12*).

Projection into nonnegative matrix

■ $\widehat{\Sigma}$ is **indefinite**: **sup-norm projection**:

$$\widetilde{\Sigma} = \underset{\mathbf{A} \geq 0}{\operatorname{argmin}} \left\{ |\mathbf{A} - \widehat{\boldsymbol{\Sigma}}|_\infty \right\}, \qquad \text{convex optimization}$$



Property: $|\widetilde{\Sigma} - \boldsymbol{\Sigma}|_\infty \leq 2|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}|_\infty$.

Robust Estimation of γ

Recall: $\gamma = \frac{1}{d(d+2)}\mathbb{E}(\xi^4) - 1$ and

$$\mathbb{E}(\xi^4) = \mathbb{E}\{[(\mathbf{X}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{X}-\mu)]^2\}.$$

Intuitive estimator: —also estimable for **subvectors**.

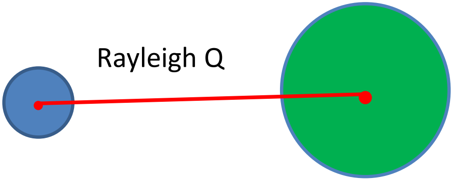
$$\widehat{\gamma} = \max\left\{\frac{1}{d(d+2)}\frac{1}{n}\sum_{i=1}^{n}[(\mathbf{X}_i-\widetilde{\mu})^\top \widetilde{\mathbf{\Omega}}(\mathbf{X}_i-\widetilde{\mu})]^2 - 1, \quad 0\right\},$$

★$\widetilde{\mu}$ and $\widetilde{\mathbf{\Omega}}$ are estimators of μ and $\mathbf{\Sigma}^{-1}$ (CLIME, *Cai, et al, 11*).

Properties: $|\widehat{\gamma}-\gamma| \le C\max\left\{|\widetilde{\mu}-\mu|_\infty, \ |\widetilde{\mathbf{\Omega}}-\mathbf{\Sigma}^{-1}|_\infty\right\}$.

Linearized Augmented Lagrangian

Target: $\min_{D(\Omega,\delta)=1} V(\Omega,\delta) + \lambda_1|\Omega|_1 + \lambda_2|\delta|_1.$

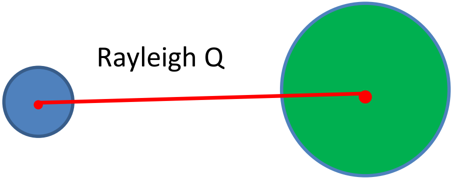


■ Let $F_\rho(\Omega,\delta,\nu) = \underbrace{V(\Omega,\delta) + \nu[\mathbf{D}(\Omega,\delta) - \mathbf{1}] + \rho[\mathbf{D}(\Omega,\delta) - \mathbf{1}]^{\mathbf{2}}}_{\text{quadratic in } \Omega \text{ and } \delta}$

$$\Omega^{(1)} \Rightarrow \delta^{(1)} \Rightarrow \nu^{(1)} \Longrightarrow \Omega^{(2)} \Rightarrow \delta^{(2)} \Rightarrow \nu^{(2)} \Longrightarrow \cdots$$

Linearized Augmented Lagrangian: Details

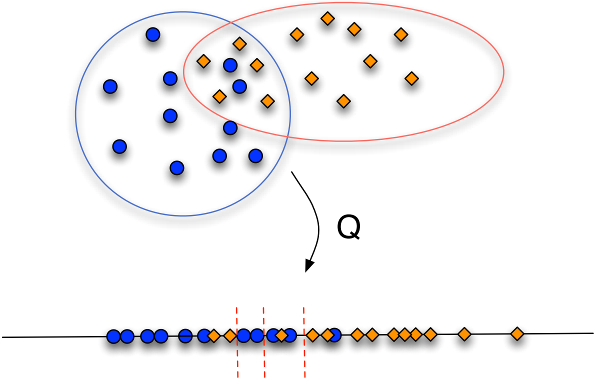
■ Minimize $\mathbf{F}_\rho(\boldsymbol{\Omega}, \delta, \nu) + \lambda_1 |\boldsymbol{\Omega}|_1 + \lambda_2 |\delta|_1$.



- $\boldsymbol{\Omega}^{(k)} = \mathrm{argmin}_{\boldsymbol{\Omega}} \left\{ F_\rho(\boldsymbol{\Omega}, \delta^{(k-1)}, \nu^{(k-1)}) + \lambda_1 |\boldsymbol{\Omega}|_1 \right\}$, **(soft-thresh.)**
- $\delta^{(k)} = \mathrm{argmin}_{\delta} \left\{ F_\rho(\boldsymbol{\Omega}^{(k)}, \delta, \nu^{(k-1)}) + \lambda_2 |\delta|_1 \right\}$, **(LASSO)**
- $\nu^{(k)} = \nu^{(k-1)} + 2\rho[D(\boldsymbol{\Omega}^{(k)}, \delta^{(k)}) - 1]$.

Application to Classification

Finding a Threshold



Where to Cut???

Finding a Threshold

Back to approx

- Classification rule: $I\{\mathbf{Z}^\top \boldsymbol{\Omega}\mathbf{Z} - 2\mathbf{Z}^\top \delta < c\} + 1$.
- Reparametrization: $c = tM_1(\boldsymbol{\Omega}, \delta) + (1-t)M_2(\boldsymbol{\Omega}, \delta)$.
- Minimizing wrt t an **approximated** classification error:

$$\overline{\mathrm{Err}}(t) \equiv \pi\bar{\Phi}\left(\frac{(1-t)D(\boldsymbol{\Omega},\delta)}{\sqrt{V_1(\boldsymbol{\Omega},\delta)}}\right) + (1-\pi)\bar{\Phi}\left(\frac{tD(\boldsymbol{\Omega},\delta)}{\sqrt{V_2(\boldsymbol{\Omega},\delta)}}\right),$$

Overview of Our Procedure

Raw Data

Robust M-estimator, and Kendall's tau correlation estimation

$\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_2, \widehat{\gamma}$

Rayleigh quotient optimization (a regularized convex programming)

$(\widehat{\boldsymbol{\Omega}}, \widehat{\delta})$

Find threshold of $c(t^*)$, where t^* is found by minimizing $\overline{\mathrm{Err}}\,(\widehat{\boldsymbol{\Omega}}, \widehat{\delta}, t)$

Quadratic Classification Rule:

$$f(\widehat{\boldsymbol{\Omega}}, \widehat{\delta}, c(t^*)) = I(\mathbf{Z}^\top \widehat{\boldsymbol{\Omega}} \mathbf{Z} - 2\mathbf{Z}^\top \widehat{\delta} < c(t^*))$$

Theoretical Results

Oracle Solutions

Oracle solution corresponding to λ_0:

$$(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0}) = \underset{D(\mathbf{\Omega},\delta)=1}{\operatorname{argmin}} \left\{ V(\mathbf{\Omega}, \delta) + \lambda_0 |\mathbf{\Omega}|_1 + \lambda_0 |\delta|_1 \right\}.$$

Special case w/ $\lambda_0 = 0$: $(\mathbf{\Omega}^*_0, \delta^*_0) = \operatorname{argmin}_{D(\mathbf{\Omega},\delta)=1} V(\mathbf{\Omega}, \delta)$.

Estimates from Quadro:

$$(\widehat{\mathbf{\Omega}}, \widehat{\delta}) = \underset{\widehat{D}(\mathbf{\Omega},\delta)=1}{\operatorname{argmin}} \left\{ \widehat{V}(\mathbf{\Omega}, \delta) + \lambda |\mathbf{\Omega}|_1 + \lambda |\delta|_1 \right\}$$

Executive Summary

Challenges: Constraints involve estimators, not unbiased.

1. Oracle performance in terms of Raleigh Quotient under RE.
2. Its generalization allows flexibility of sparsity.
3. $\overline{\mathrm{Err}}(t)$ provides a valid approximation.
4. Raleight Quotient provides a good surrogate for classification error.

Restricted Eigenvalue

■ But target is quadratic in $\boldsymbol{\Omega}$ and δ.

$$\mathbf{Q}_k = \begin{bmatrix} \left(2(1+\gamma)\boldsymbol{\Sigma}_k + 4\mu_k\mu_k^\top\right)\otimes\boldsymbol{\Sigma}_k + \gamma\,\mathrm{vec}(\boldsymbol{\Sigma}_k)\,\mathrm{vec}(\boldsymbol{\Sigma}_k)^\top & -4\mu_k\otimes\boldsymbol{\Sigma}_k \\ -4\mu_k^\top\otimes\boldsymbol{\Sigma}_k & 4\boldsymbol{\Sigma}_k \end{bmatrix}$$

RE on $\mathbf{Q} = \mathbf{Q}_1 + \kappa\mathbf{Q}_2$: For S and $\bar{c} \geq 0$, define its RE by

$$\Theta(S;\bar{c}) = \min_{\mathbf{v}:|\mathbf{v}_{S^c}|_1 \leq \bar{c}|\mathbf{v}_S|_1} \frac{\mathbf{v}^\top\mathbf{Q}\mathbf{v}}{|\mathbf{v}_S|^2}.$$

(Bickel et al, 09; van de Geer, 07; Candes and Tao, 05)

Oracle Inequality on Rayleigh Quotient

With $\lambda = C\eta \max\{s_0^{1/2}\Delta_n, k_0^{1/2}\lambda_0\}[R(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0})]^{-1/2}$,

$$\frac{R(\widehat{\mathbf{\Omega}}, \widehat{\delta})}{R(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0})} \geq 1 - A\eta^2 \max\left\{s_0\Delta_n, s_0^{1/2}k_0^{1/2}\lambda_0\right\}.$$

Estimation error: $\Delta_n = \max_{k=1,2}\{|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty, |\widehat{\mu}_k - \mu_k|_\infty\}$.

Sparsity: $S = \mathrm{supp}[\mathrm{vec}(\mathbf{\Omega}^*_{\lambda_0})^\top, (\delta^*_{\lambda_0})^\top]^\top$, $s_0 = |S|$ and $k_0 = \max\{s_0, \mathrm{R}(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0})\}$.

- For some $a_0, c_0, u_0 > 0$, $\Theta(S,0) \geq c_0$, $\Theta(S,3) \geq a_0$, and $R(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0}) \geq u_0$.
- $\max\{s_0\Delta_n, s_0^{1/2}k_0^{1/2}\lambda_0\} < 1, \qquad 4s_0\Delta_n^2 < a_0c_0$.

Oracle Inequality on Rayleigh Quotient

Oracle Inequality on Rayleigh Quotient

With $\lambda = C\eta \max\{s_0^{1/2}\Delta_n, k_0^{1/2}\lambda_0\}[R(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0})]^{-1/2}$,

$$\frac{R(\widehat{\mathbf{\Omega}}, \widehat{\delta})}{R(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0})} \geq 1 - A\eta^2 \max\left\{s_0\Delta_n, s_0^{1/2}k_0^{1/2}\lambda_0\right\}.$$

Estimation error: $\Delta_n = \max_{k=1,2}\{|\widehat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}_k|_\infty, |\widehat{\mu}_k - \mu_k|_\infty\}$.

Sparsity: $S = \text{supp}[\text{vec}(\mathbf{\Omega}^*_{\lambda_0})^\top, (\delta^*_{\lambda_0})^\top]^\top$, $s_0 = |S|$ and $k_0 = \max\{s_0, \mathrm{R}(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0})\}$.

- For some $a_0, c_0, u_0 > 0$, $\Theta(S,0) \geq c_0$, $\Theta(S,3) \geq a_0$, and $R(\mathbf{\Omega}^*_{\lambda_0}, \delta^*_{\lambda_0}) \geq u_0$.
- $\max\{s_0\Delta_n, s_0^{1/2}k_0^{1/2}\lambda_0\} < 1, \qquad 4s_0\Delta_n^2 < a_0c_0$.

Oracle Inequality: Corollaries

Corrolary 2 ($\lambda_0 = 0$): With our robust est, when

$$\lambda > C s_0^{1/2} R_{\max}^{-1/2} \sqrt{\log(d)/n},$$

with prob $\geq 1 - (n \vee d)^{-1}$,

$$R(\widehat{\boldsymbol{\Omega}}, \widehat{\delta}) \geq \left(1 - A s_0 \sqrt{\log(d)/n}\right) R_{\max},$$

★ $R_{\max} = \mathrm{R}(\boldsymbol{\Omega}_0^*, \delta_0^*)$,

Approximate of Classification Error

To definition

Under normality & mild conditions, as $d \to \infty$,

$$\left|\mathrm{Err}(\boldsymbol{\Omega}, \delta, t) - \overline{\mathrm{Err}}(\boldsymbol{\Omega}, \delta, t)\right| = \frac{\mathrm{rank}(\boldsymbol{\Omega}) + \mathbf{o}(\mathbf{d})}{[\min\{\mathbf{V_1}(\boldsymbol{\Omega}, \delta), \mathbf{V_2}(\boldsymbol{\Omega}, \delta)\}]^{3/2}}.$$

★ If $\mathrm{var}_k(Q(\mathbf{X})) > c_0 d^\theta$ for $\theta > 2/3$, then $|\mathrm{Err} - \overline{\mathrm{Err}}| = o(1)$.

★ $t^* = \operatorname*{argmin}_t \overline{\mathrm{Err}}(\boldsymbol{\Omega}, \delta, t)$ is reasonable.

Rayleigh Quotient versus $\overline{\mathrm{Err}}(\Omega, \delta, t)$: Notation

- $H(x) = \bar{\Phi}(1/\sqrt{x})$, where $\bar{\Phi} = 1 - \Phi$.
- $R^{(t)} = R(\boldsymbol{\Omega}, \delta)$ w/ weight $\kappa(t) \equiv \frac{1-\pi}{\pi} \frac{(1-t)^2}{t^2}$.
- $R_k = R_k(\boldsymbol{\Omega}, \delta) = [D(\boldsymbol{\Omega}, \delta)]^2 / V_k(\boldsymbol{\Omega}, \delta)$, for $k = 1, 2$.
- $U_1 = U_1(\boldsymbol{\Omega}, \delta, t) = \min\left\{(1-t)^2 R_1,\ \frac{1}{(1-t)^2 R_1}\right\}$.
- $U_2 = U_2(\boldsymbol{\Omega}, \delta, t) = \min\left\{t^2 R_2,\ \frac{1}{t^2 R_2}\right\}$.
- $U = U(\boldsymbol{\Omega}, \delta, t) = \max\{U_1/U_2,\ U_2/U_1\}$.
- $R_0 = \max\{\min\{R_1, 1/R_1\}, \min\{R_2, 1/R_2\}\}$ & $\Delta R = |R_1 - R_2|$.

Rayleigh Quotient versus $\overline{\mathrm{Err}}(\mathbf{\Omega}, \delta, t)$

Distance between $\overline{\mathrm{Err}}(\mathbf{\Omega}, \delta, t)$ and monotone transform of $\mathrm{R}(\mathbf{\Omega}, \delta)$

There exists a constant $C > 0$ such that

$$\left|\overline{\mathrm{Err}}(\mathbf{\Omega}, \delta, t) - H\left(\frac{\pi}{(1-t)^2 R^{(t)}(\mathbf{\Omega}, \delta)}\right)\right| \le C\left[\max\{U_1, U_2\}\right]^{1/2} \cdot |U-1|^2.$$

In particular, when $t = 1/2$,

$$\left|\overline{\mathrm{Err}}(\mathbf{\Omega}, \delta, t) - H\left(\frac{4\pi}{R^{(t)}(\mathbf{\Omega}, \delta)}\right)\right| \le C R_0^{1/2} \cdot \left(\frac{\Delta R}{R_0}\right)^2.$$

★Remarks:

- $|V_1 - V_2| \ll \min\{V_1, V_2\}$, then $\Delta R \ll R_0$.
- $R_0 \le 1$ always. $R_0 \to 0$ when $R_1, R_2 \to \infty$, or $R_1, R_2 \to 0$, or $R_1 \to 0, R_2 \to \infty$.
- Under mild conditions, a monotone transform of $\mathrm{R}(\mathbf{\Omega}, \delta)$ approximates $\overline{\mathrm{Err}}$, and hence approximates the true error $\mathrm{Err}(\mathbf{\Omega}, \delta)$.

Numerical Studies

Simulation Setup

- $d = 40, n_1 = n_2 = 50$, testing: $N_1 = N_2 = 4000$.
- Repeat 100 times.
- Augmented Lagrangian parameters:

$$\rho = 0.5, \nu^0 = 0, \delta^0 = \mathbf{0}.$$

- (λ_1, λ_2) are chosen by optimal tuning.

Simulation: Gaussian Settings ($\mu_1 = \mathbf{0}$)

★ Model 1: $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = \text{diag}(\mathbf{1.3}_{10}, \mathbf{1}_{30})$, $\mu_2 = (\mathbf{0.7}_{10}^\top, \mathbf{0}_{30}^\top)^\top$.

★ Model 2: $\mathbf{\Sigma}_1 = \mathrm{diag}(\mathbf{A}, \mathbf{I}_{20})$, with $\mathbf{A}$ equi-corr $\rho = 0.4$.
$\mathbf{\Sigma}_2 = (\mathbf{\Sigma}_1^{-1} + \mathbf{I})^{-1}$. $\mu_2 = \mathbf{0}_d$.

★ Model 3: $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ as Model 2 and μ_2 as Model 1.

Methods: ★Sparse Logistic Reg with interactions (SLR)
★Linear-SLR ★ROAD ★Quadro-0 (non-robust)

Design of Simulation: t-Distribution Settings

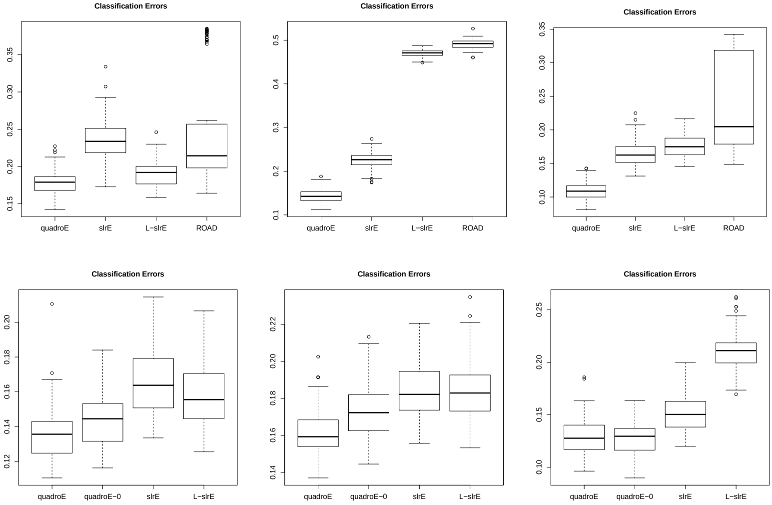
Multivariate t-dist.: $t_\nu(\mu_1, \boldsymbol{\Sigma}_1)$ and $t_\nu(\mu_2, \boldsymbol{\Sigma}_2)$, with $\nu = 5$.

★ Model 4: Same as Model 1.

★ Model 5: Same as Model 1, but $\boldsymbol{\Sigma}_2$ fractional WN w/ $l = 0.2$, i.e. $|\Sigma_2(i,j)| = O(|i-j|^{1-2l})$.

★ Model 6: Same as Model 1, but $\boldsymbol{\Sigma}_2 = (0.6^{|j-k|})$ —AR(1).

Results — Classification errors

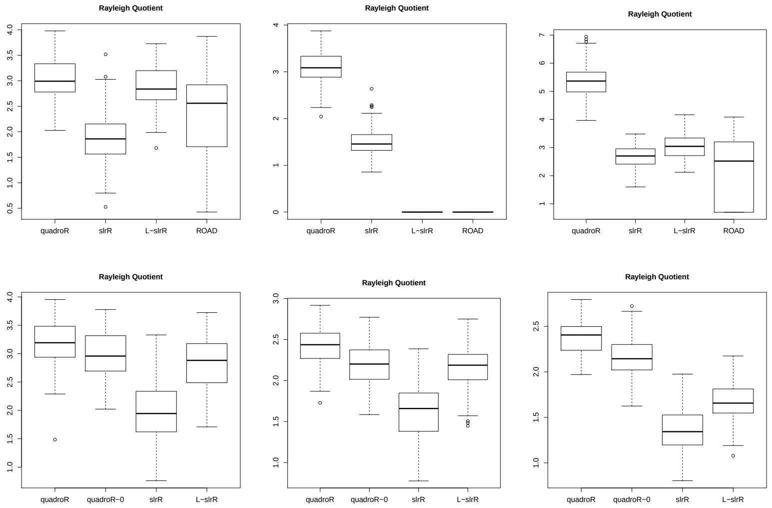


Results — Classification errors

	QUADRO	SLR	L-SLR	ROAD
Model 1	**0.179**	0.235	0.191	0.246
Model 2	**0.144**	0.224	0.470	0.491
Model 3	**0.109**	0.164	0.176	0.235

	QUADRO	QUADRO-0	SLR	L-SLR
Model 4	**0.136**	0.144	0.167	0.157
Model 5	**0.161**	0.173	0.184	0.184
Model 6	**0.130**	**0.129**	0.152	0.211

Results — Rayleigh Quotients



Results — Rayleigh Quotients

	QUADRO	SLR	L-SLR	ROAD
Model 1	**3.016**	1.874	2.897	2.193
Model 2	**3.081**	1.508	0	0
Model 3	**5.377**	2.681	3.027	2.184

	QUADRO	QUADRO-0	SLR	L-SLR
Model 4	**3.179**	2.975	1.984	2.846
Model 5	**2.415**	2.191	1.625	2.166
Model 6	**2.374**	2.160	1.363	1.669

Empirical Study: Breast Tumor Data

GPL96 data: $d = 12679$ genes, $n_1 = 1142$ (breast tumor) and $n_2 = 6982$ (non-breast tumor).

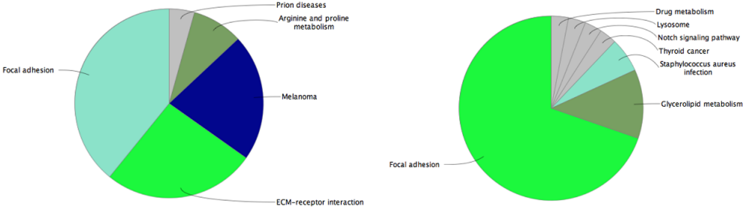
Testing and training: 200 and 942 samples from each class.

★Repeat 100 times

Tuning parameters: Half used to estimate $(\delta, \mathbf{\Sigma})$; half selecting regularization parameters.

Classification errors on testing set		
QUADRO	SLR	L-SLR
0.014	0.025	0.025
(0.007)	(0.007)	(0.009)

Pathway Enrichment



Quadro pathways (139) **SLR pathways (128)**

Figure: From KEGG database, genes selected by Quadro belong to 5 of the pathways that contain more than two genes; correspondingly, genes selected by SLR belong to 7 pathways.

★ QUADRO provides fewer, but more enriched pathways.

★ *ECM-receptor* is highly related to breast cancer.

Gene Ontology (GO) Enrichment Analysis

GO ID	GO attribute	No. of Genes	p-value
0048856	anatomical structure development	58	3.7E-12
0032502	developmental process	62	2.9E-10
0048731	system development	52	3.1E-10
0007275	multicellular organismal development	55	1.8E-8
0001501	skeletal system development	15	1.3E-6
0032501	multicellular organismal process	66	1.4E-6
0048513	organ development	37	1.4E-6
0009653	anatomical structure morphogenesis	28	8.7E-6
0048869	cellular developmental process	34	1.9E-5
0030154	cell differentiation	33	2.1E-5
0007155	cell adhesion	18	2.4E-4
0022610	biological adhesion	18	2.2E-4
0042127	regulation of cell proliferation	19	2.9E-4
0009888	tissue development	17	3.7E-4
0007398	ectoderm development	9	4.8E-4
0048518	positive regulation of biological process	34	5.6E-4
0009605	response to external stimulus	20	6.3E-4
0043062	extracellular structure organization	8	7.4E-4
0007399	nervous system development	22	8.4E-4

- ★ Selected biological processes are related to previously enriched pathways.
- ★ *Cell adhesion* is known to be highly related to *cell communication pathways*, including *focal adhesion* and *ECM-receptor interaction*.

Summary

★ Propose Rayleigh Quotient for quadratic classification.

★ Use elliptical dist to avoid fourth cross-moments.

★ Adopt Catoni's M-est and Kendall's tau for robust est.

★ Convex optimization solved by augmented Lagrangian.

★ Explore its applications to classification.

★ Oracle inequalities, Rayleigh quotient and class. error.

Summary

★ Propose Rayleigh Quotient for quadratic classification.

★ Use elliptical dist to avoid fourth cross-moments.

★ Adopt Catoni's M-est and Kendall's tau for robust est.

★ Convex optimization solved by augmented Lagrangian.

★ Explore its applications to classification.

★ Oracle inequalities, Rayleigh quotient and class. error.

The End

Thank  *You*