High-dimensional Feature Selection

Jianqing Fan

Princeton University

http://www.princeton.edu/∼jqfan

May 3, 2010

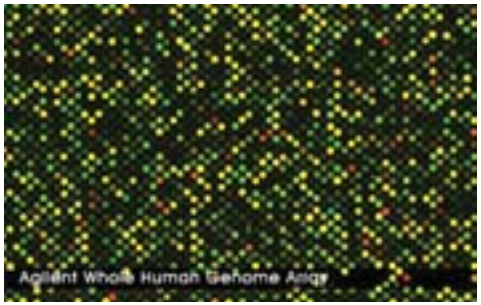
Outline

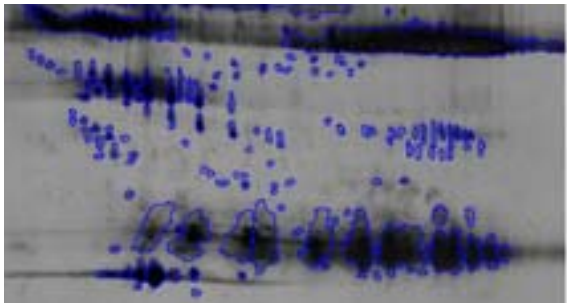
Rise of high-dimensionality

Examples: Biological Sciences

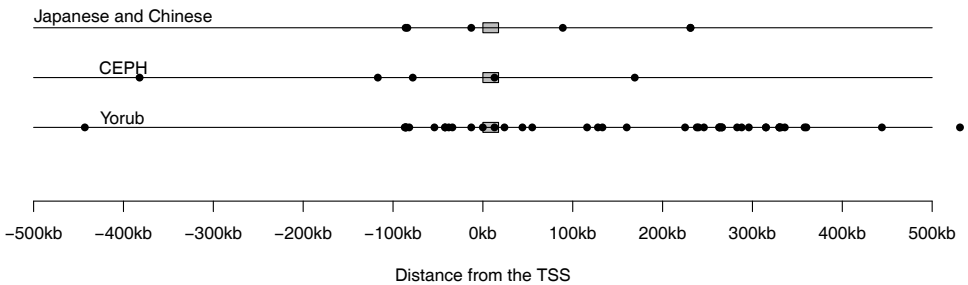
High-dim variable selection characterizes many contemporary statistical problems.

- Bioinformatic: disease classification / predicting clinical outcomes using microarray, proteomics, fMRI data;





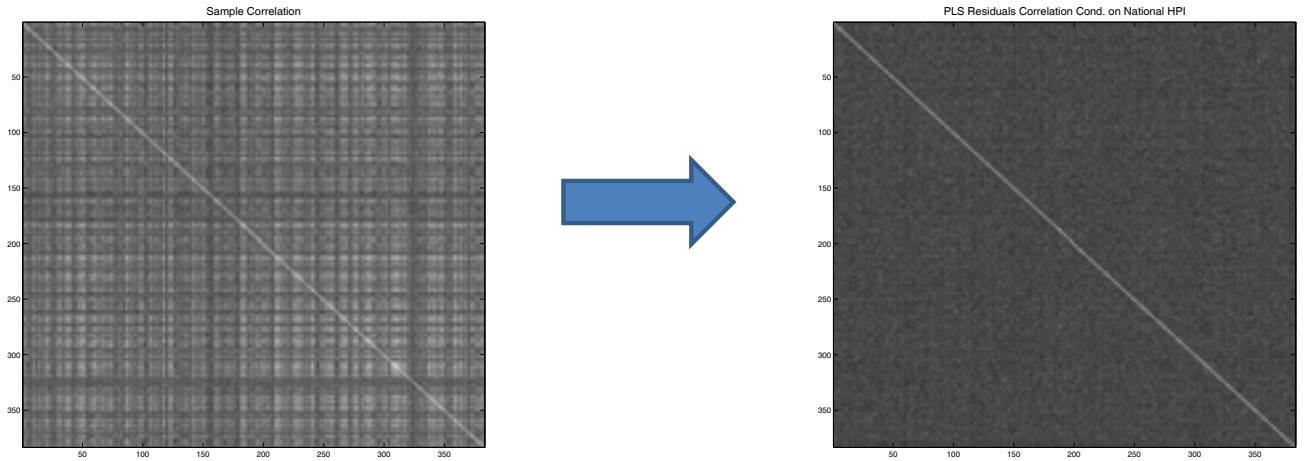
- Association studies between phenotypes and SNPs.



- eQTL

Example: Economics, Finance, Marketing

- HPA / drug sales collected in many regions
- Local correlation makes dimensionality growths quickly.

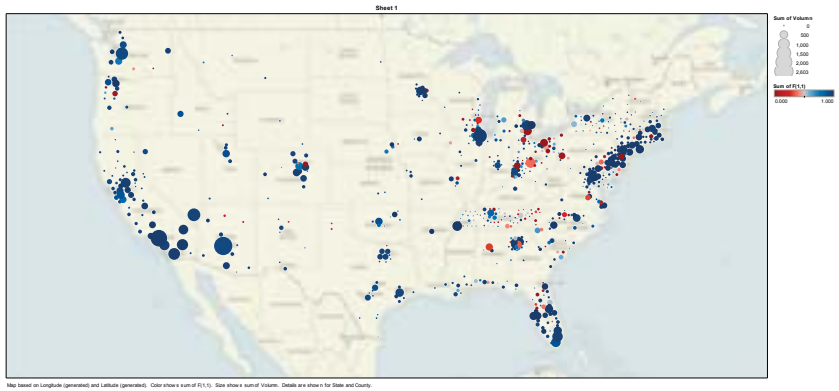


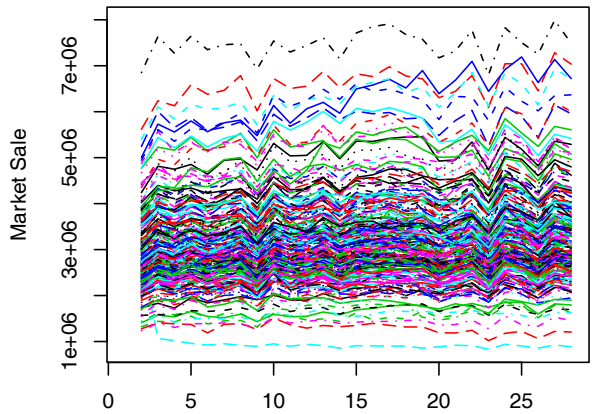
- 1000 neighborhoods requires 1 m parameters.
- Managing 2K stocks involves 2m elements in covariance.

Example: Spatial temporal data

Meteorology & Earth Sciences & Ecology

- Temperatures and other attributes (precipitation, population size) are collected over time and over many regions.





- Forecasting large panel data over a short time horizon poses more challenges.

Example: Machine Learning

- Document or text classification: E-mail spam.
  - Feature extractions: Frequency counting
  - Word-document information: For document x and word y, define

$$I_{x,y} = \log\left(\frac{nc_{x,y}}{\sum_\xi c_{\xi,y} \sum_\xi c_{x,\xi}}\right),$$

    where $c_{x,y} =$ No. of word y in doc x.

    ★Each word is summarized by $(I_{1,y}, \cdots, I_{p,y})$

    ★Each document summarized by $(I_{x,1}, \cdots, I_{x,q})$

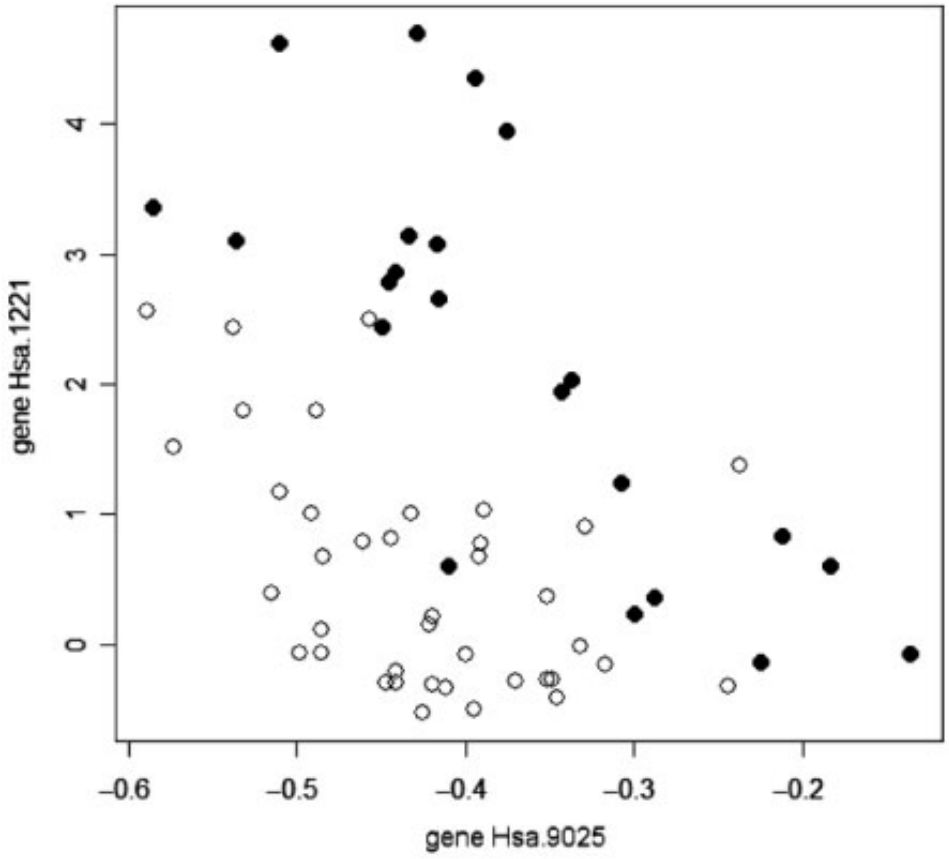
- Computer vision.

Growth of Dimensionality

- Dimen. grows rapidly w/ interactions: 5000 $\Longrightarrow$ 12.5m.

Synergy of Two Genes: colon cancer in Hanczar et al (2007).

e.g., $Y = I(X_1 + X_2 > 3)$ **and** $Y \perp X_1$.



white – patients; black – normal

Aims of High-dimensional Regression and Classification

- To construct as effective a method as possible to predict future observations.
- To gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction method.

Bickel (2008)

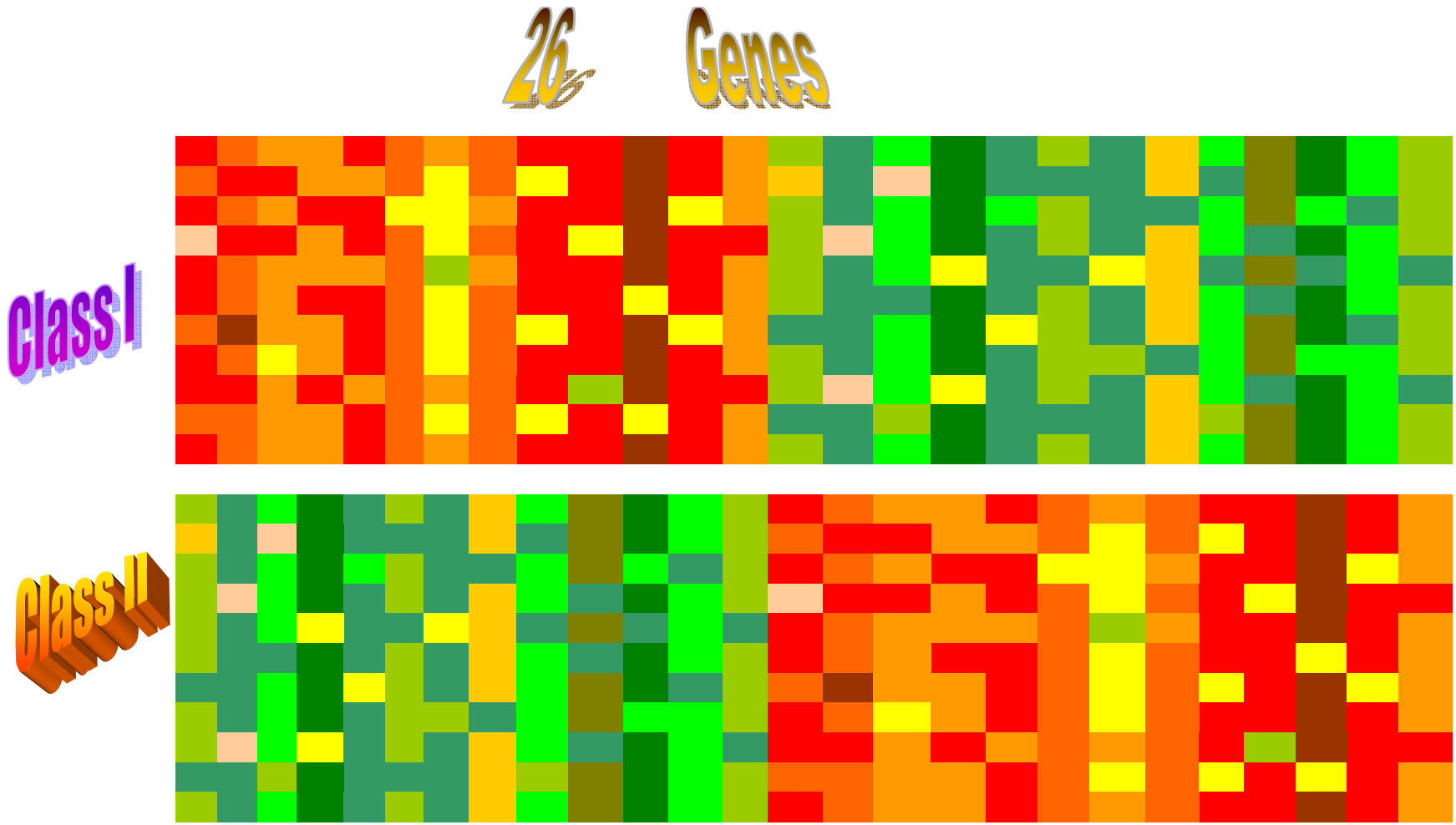
Aims of High-dimensional Regression and Classification

- To construct as effective a method as possible to predict future observations.

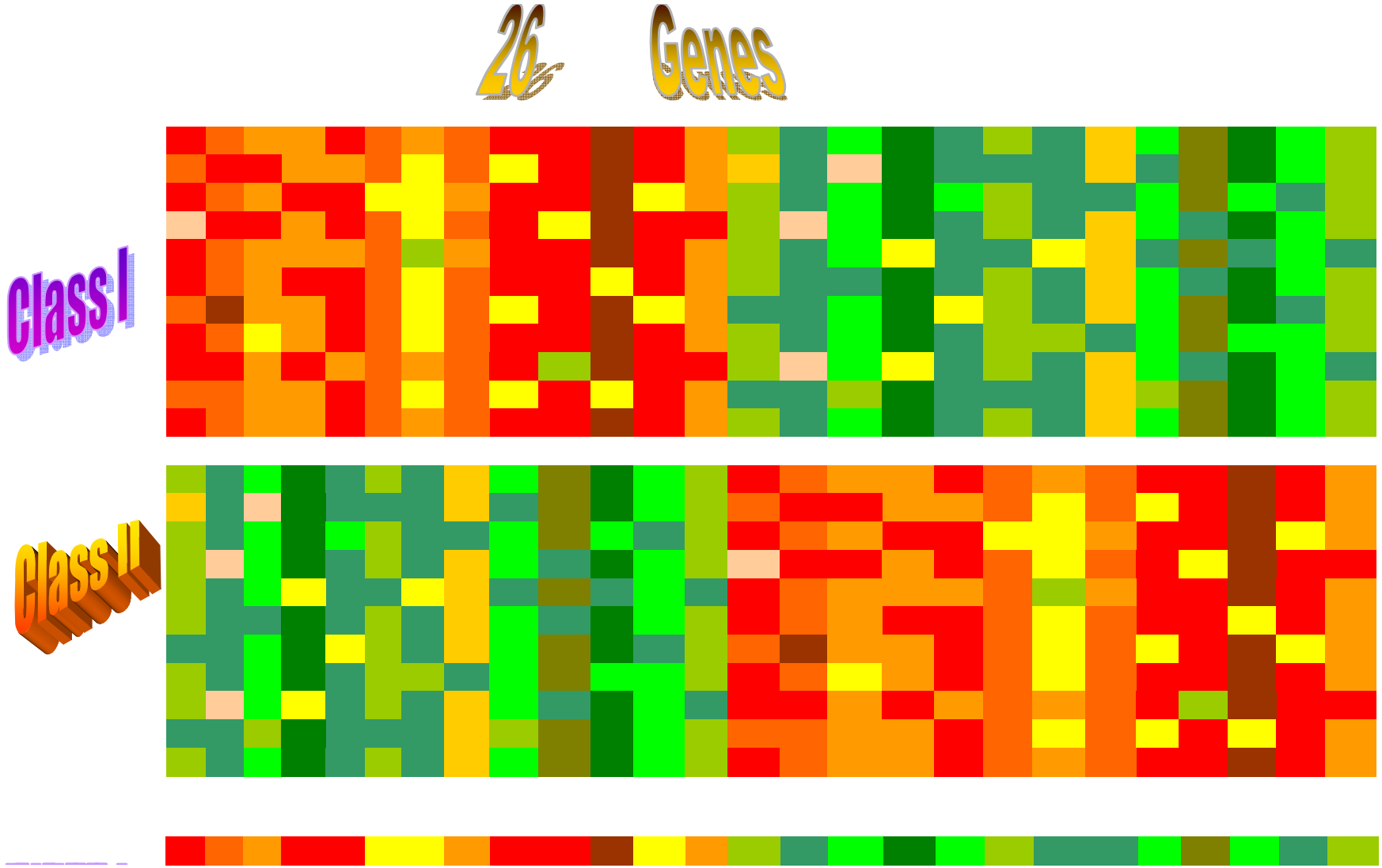
- To gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction method.

- Bickel (2008)

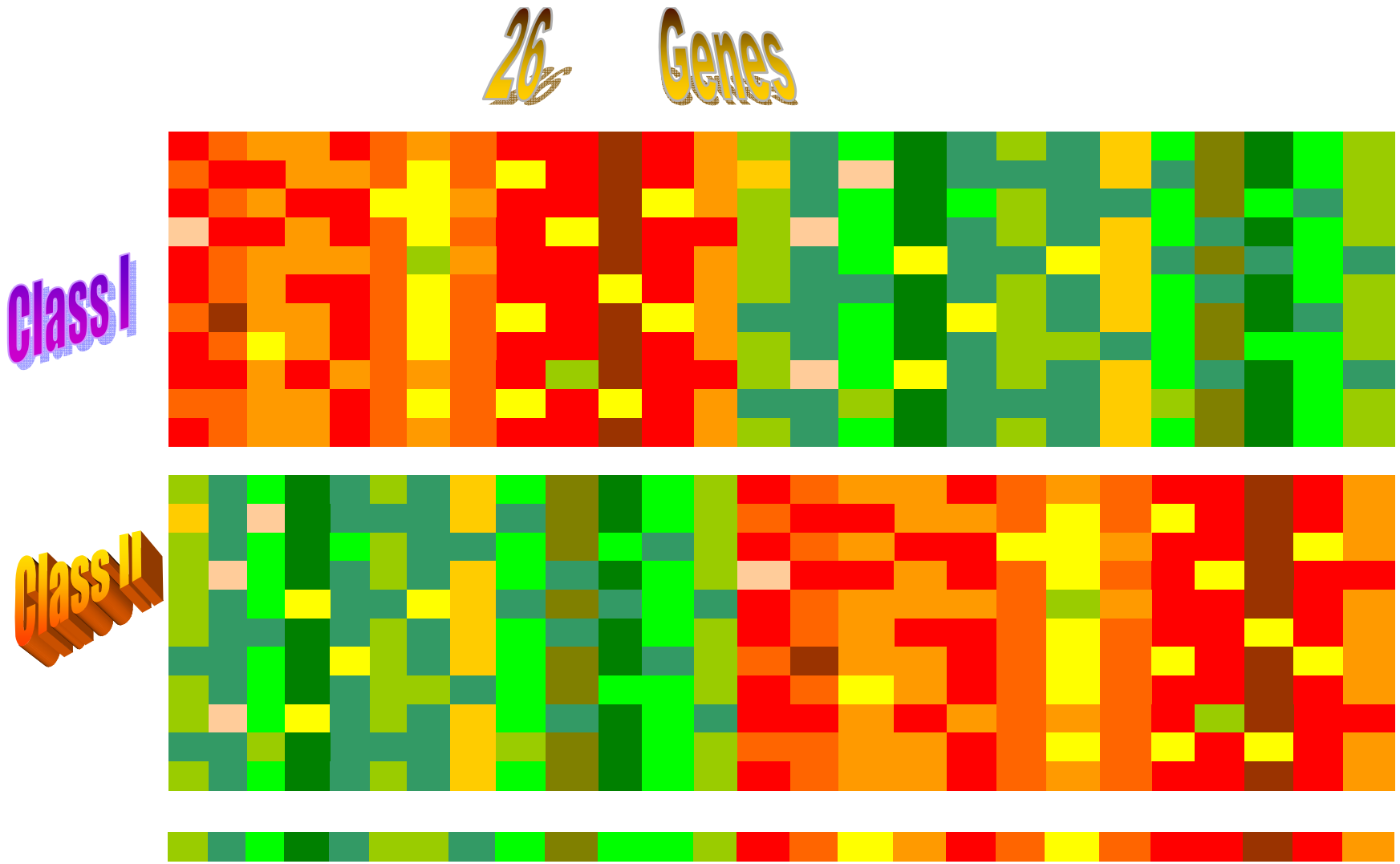
Goal of Feature Selection



Goal of Feature Selection



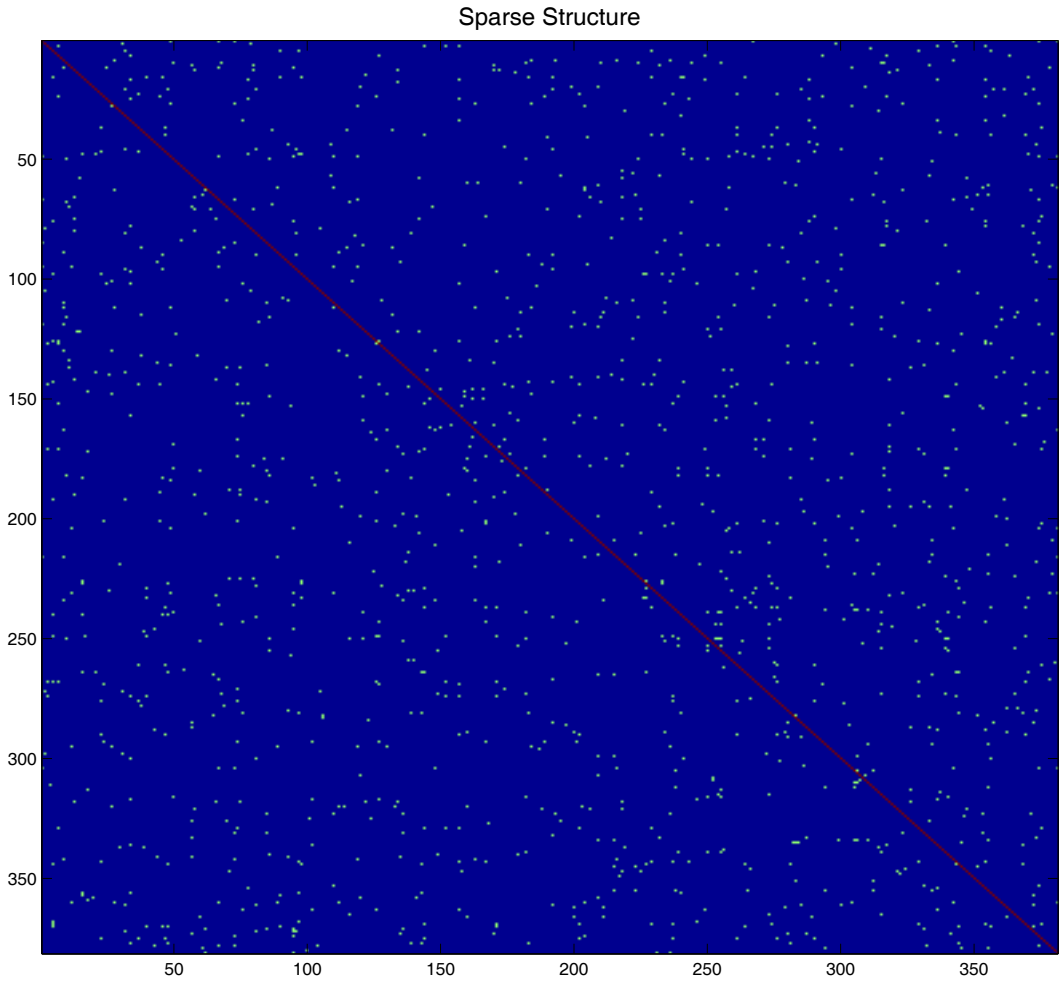
Goal of Feature Selection



Questions: How to select? How many?

Popular Assumption: Sparsity

Dimen: $\log p = O(n^a)$ **Intrinsic dim**: $s \ll n$. (Sparsity)



Essential Assumption: homogeneity

Sparsity: $\theta_i \sim_{i.i.d.} (1-p_1)\delta_0 + p_1 F_1, \qquad p_1 \approx 0.$

- two mixtures with known atom **0**

Homogeneity: $\theta_i \sim_{i.i.d.} p_1 F_1 + p_2 F_2 + p_3 F_3 + \cdots + p_k F_k$

e.g. $F_1 = \delta_0$, $F_2 = \delta_\mu$ (**unknown**).

Example: Projecting housing prices.

- Local regions have "$\approx$" regression coefficients
- Time lag dependence have "$\approx$" homogeneous.

Example: Counting "+" in lab tests, collapsing categorical variables.

Essential Assumption: homogeneity

Sparsity: $\theta_i \sim_{i.i.d.} (1-p_1)\delta_0 + p_1 F_1, \qquad p_1 \approx 0.$

- two mixtures with known atom **0**

Homogeneity: $\theta_i \sim_{i.i.d.} p_1 F_1 + p_2 F_2 + p_3 F_3 + \cdots + p_k F_k$

e.g. $F_1 = \delta_0$, $F_2 = \delta_\mu$ (**unknown**).

Example: Projecting housing prices.

- Local regions have "$\approx$" regression coefficients
- Time lag dependence have "$\approx$" homogeneous.

Example: Counting "+" in lab tests, collapsing categorical variables.

Essential Assumption: homogeneity

Sparsity: $\theta_i \sim_{i.i.d.} (1-p_1)\delta_0 + p_1 F_1, \qquad p_1 \approx 0.$

- two mixtures with known atom **0**

Homogeneity: $\theta_i \sim_{i.i.d.} p_1 F_1 + p_2 F_2 + p_3 F_3 + \cdots + p_k F_k$

e.g. $F_1 = \delta_0$, $F_2 = \delta_\mu$ (**unknown**).

Example: Projecting housing prices.

- Local regions have "$\approx$" regression coefficients
- Time lag dependence have "$\approx$" homogeneous.

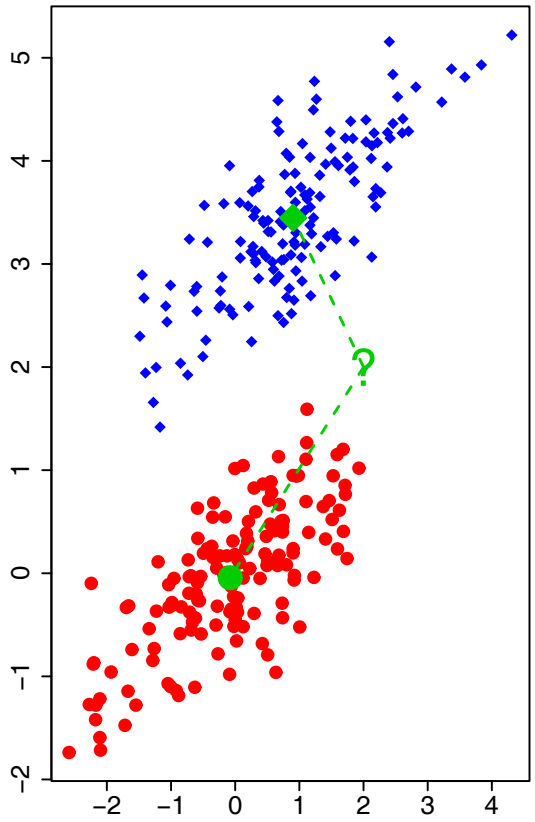
Example: Counting "$+$" in lab tests, collapsing categorical variables.

Impact of Dimensionality

1. Noise accumulation

Regression:

- **Not** directly implementable if $p > n$.
- Prediction error is $(1 + \frac{p}{n})\sigma^2$, if $p \leq n$.



Classification: No implementation problems, but **error rates**

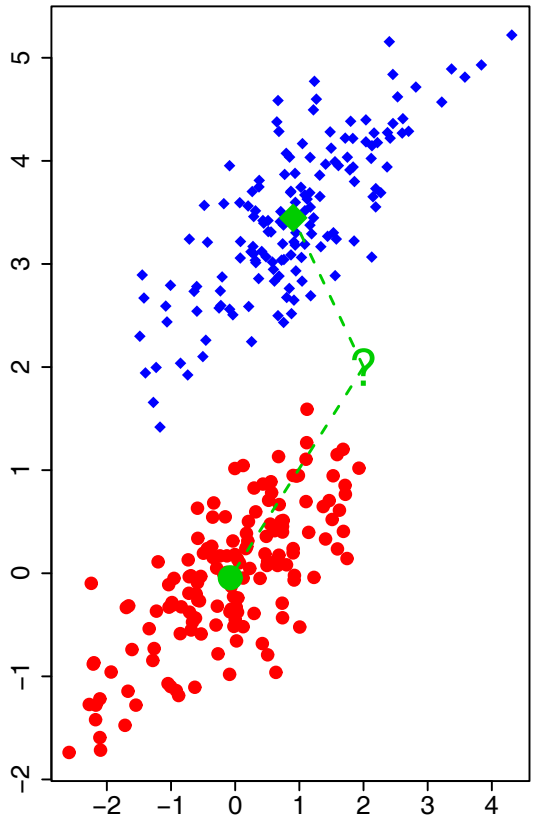
—depend on $C_p^2/\sqrt{p}$ (*Fan & Fan 08*), C_p is **distance**.

—**perfectly classifiable** if $C_p^2/\sqrt{p} \to \infty$ (*Hall, Pittelkow & Ghosh,08*).

1. Noise accumulation

Regression:

- **Not** directly implementable if $p > n$.
- Prediction error is $(1 + \frac{p}{n})\sigma^2$, if $p \leq n$.



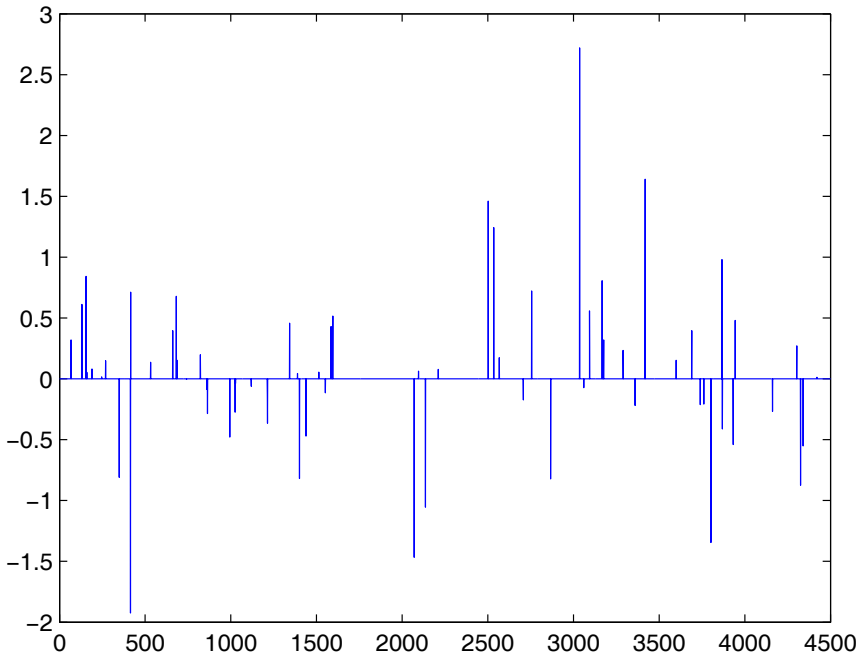
Classification: No implementation problems, but **error rates**

—depend on $C_p^2/\sqrt{p}$ (*Fan & Fan 08*), C_p is **distance**.

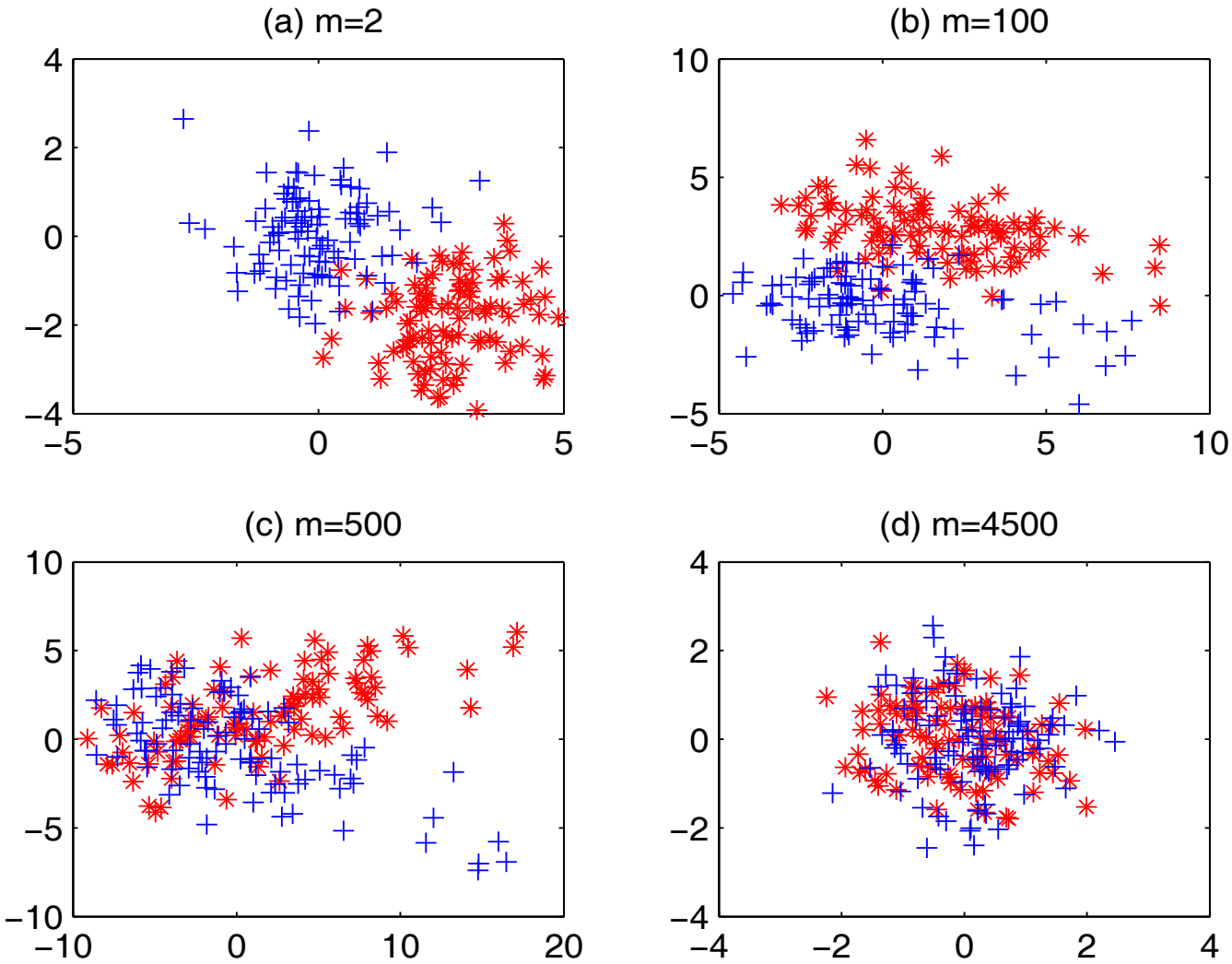
—**perfectly classifiable** if $C_p^2/\sqrt{p} \to \infty$ (*Hall, Pittelkow & Ghosh,08*).

An illustration

- **dimensionality**: $p = 4500, \qquad n = 200$
- **Signals**: $\mu_1 = 0.98\delta_0 + 0.02\text{DE}, \qquad \mu_2 = 0$



Impact of Dimensionality on classification

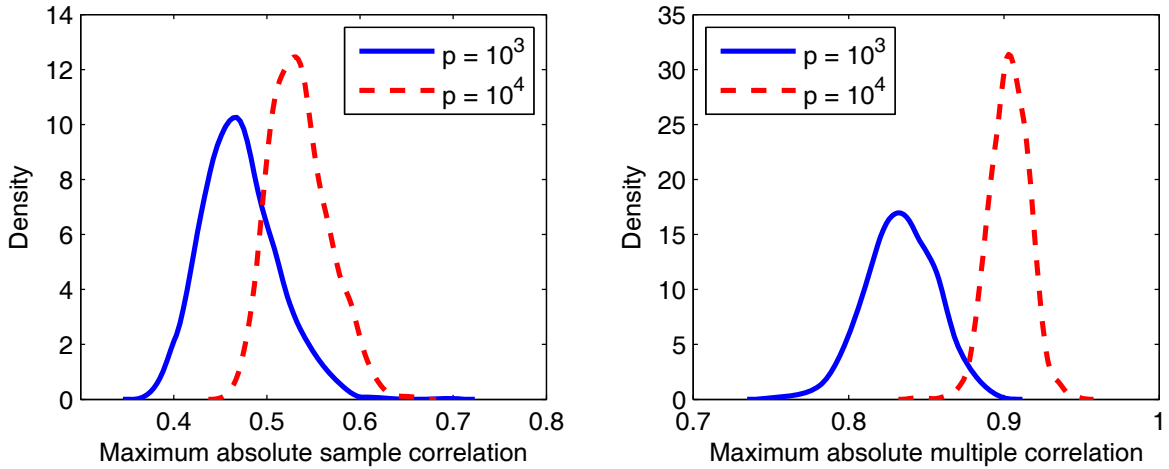


- Classification power depends on $\sum_{i=1}^{d} \alpha_i^2 / \sqrt{d}$.

2. Spurious correlations

An experiment: Generate $n = 50$ $Z_1, \cdots, Z_p \sim_{i.i.d.} N(0,1)$;

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



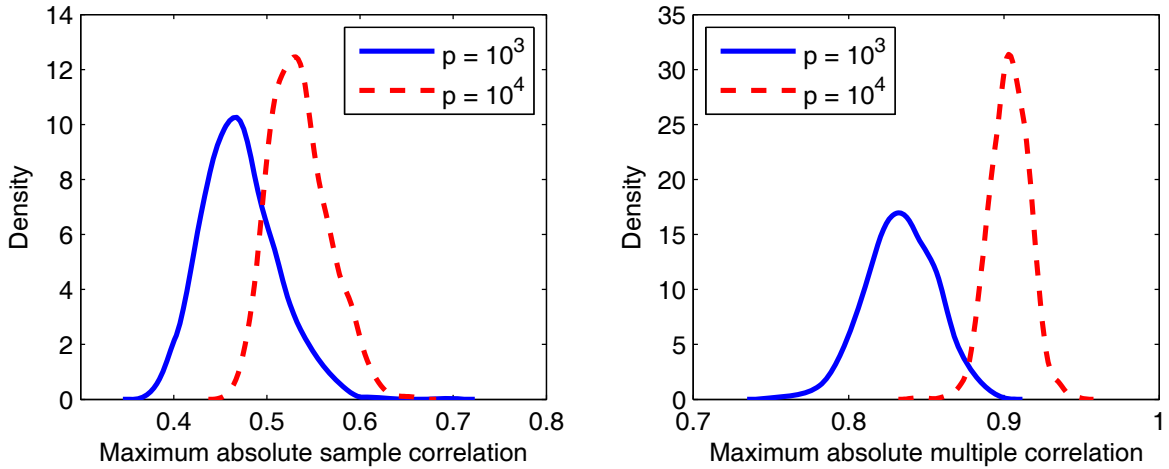
- compute maximum multiple correlation:

$R = \max_{|S|=5} \text{corr}(Z_1, \mathbf{Z}_S)$.

2. Spurious correlations

An experiment: Generate $n = 50$ $Z_1, \cdots, Z_p \sim_{i.i.d.} N(0,1)$;

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



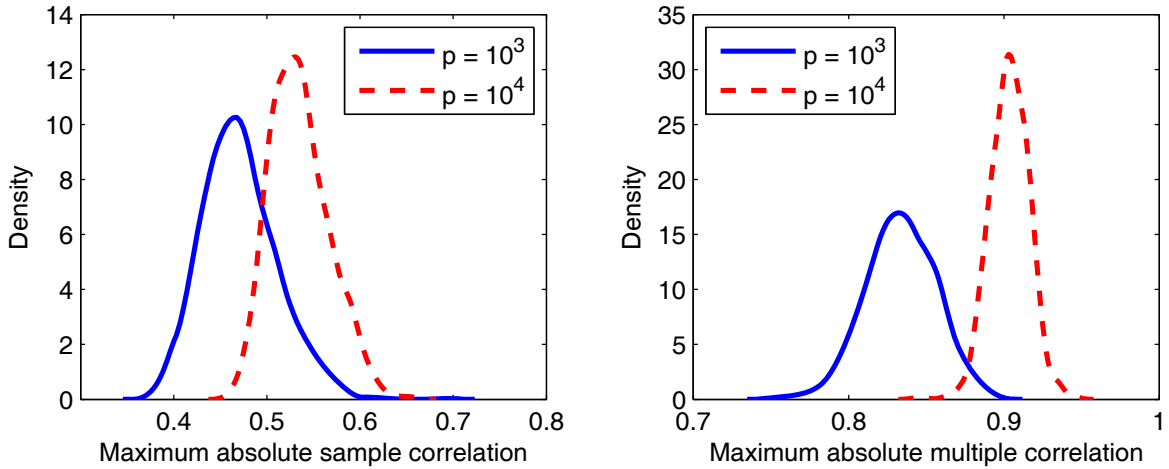
- compute maximum multiple correlation:

$R = \max_{|S|=5} \text{corr}(Z_1, \mathbf{Z}_S)$.

2. Spurious correlations

An experiment: Generate $n = 50$ $Z_1, \cdots, Z_p \sim_{i.i.d.} N(0,1)$;

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



- compute maximum multiple correlation:

$R = \max_{|S|=5} \text{corr}(Z_1, \mathbf{Z}_S)$.

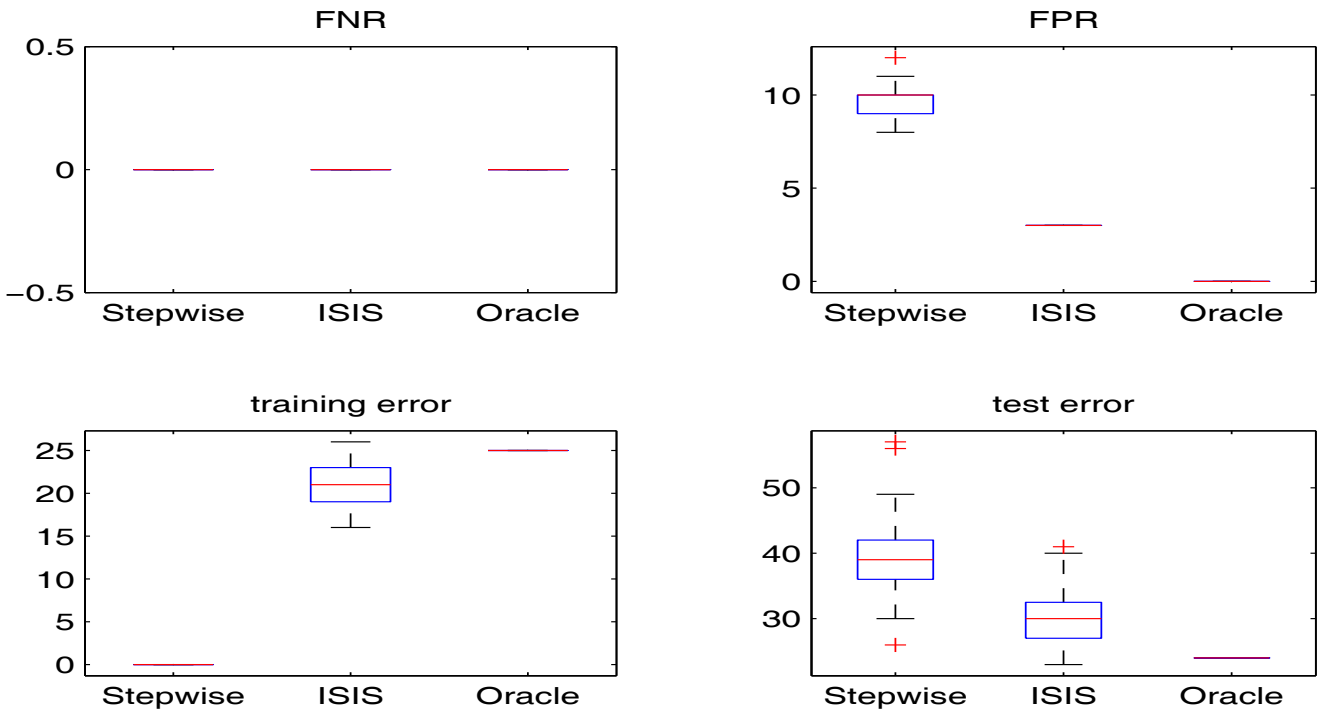
False scientific discoveries

If Z_1 is responsible for breast cancer, but we can also discover other 5 genes, **indep of outcome**!

- $Y = 1$ and 0, whether a neuroblastoma child has 3-y EFS.

  $n = 125$: 25 "+" and 100 "−", testing $= 114$

- X's are independent normal, **simulating** gene expressions.



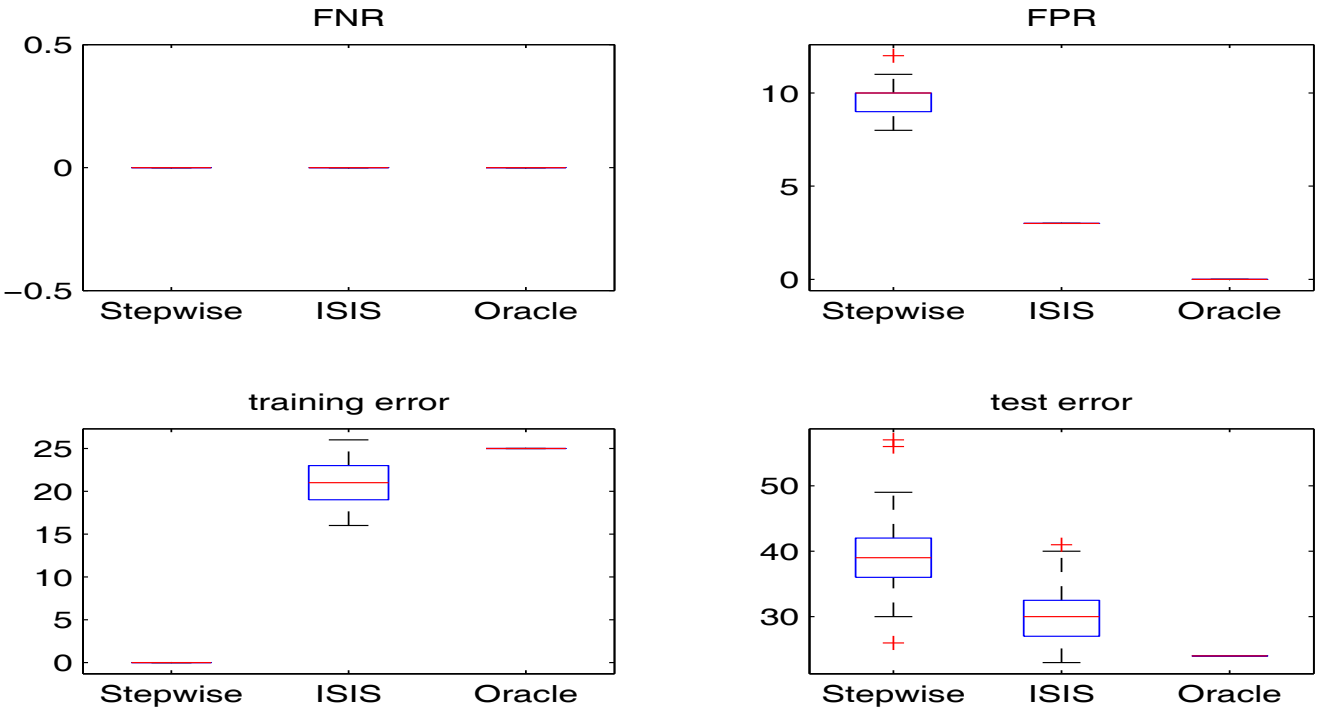
False scientific discoveries

If Z_1 is responsible for breast cancer, but we can also discover other 5 genes, **indep of outcome**!

- $Y = 1$ and 0, whether a neuroblastoma child has 3-y EFS.

  $n = 125$: 25 "+" and 100 "−", testing $= 114$

- X's are independent normal, **simulating** gene expressions.



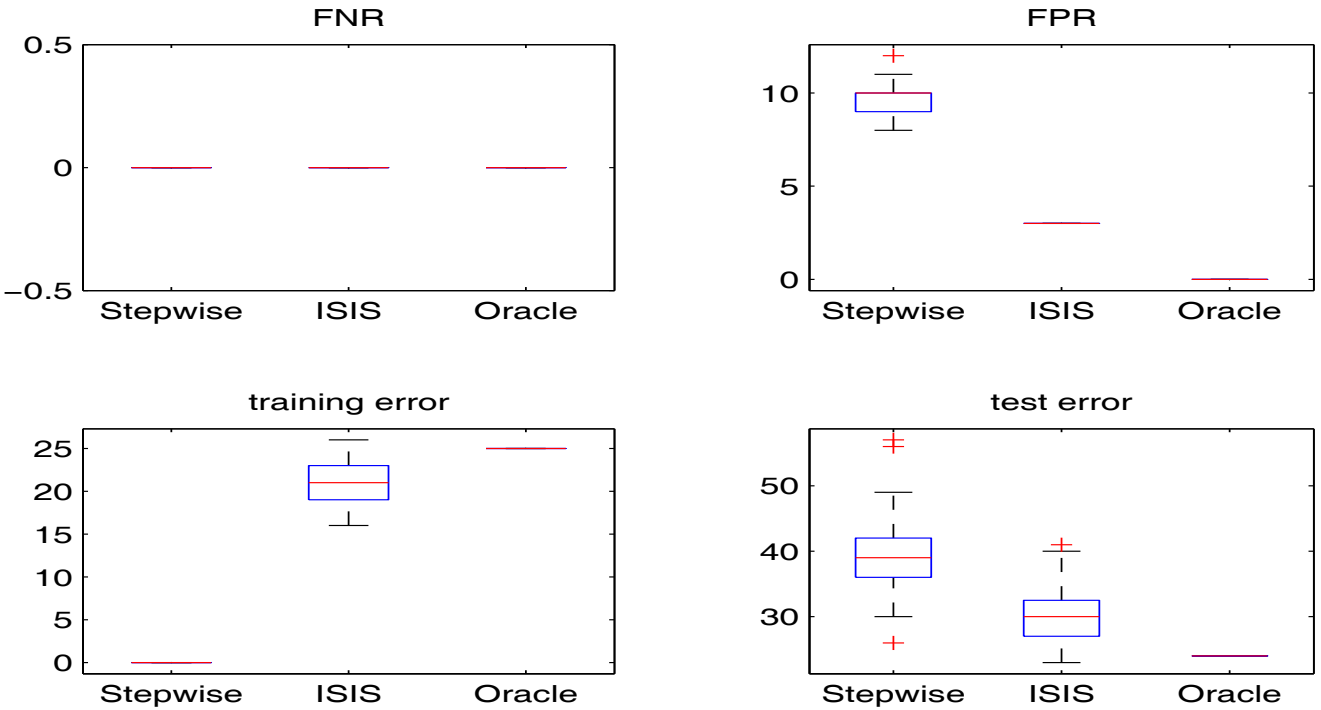
False scientific discoveries

If Z_1 is responsible for breast cancer, but we can also discover other 5 genes, **indep of outcome**!

- $Y = 1$ and 0, whether a neuroblastoma child has 3-y EFS.

  $n = 125$: 25 "+" and 100 "−", testing $= 114$

- X's are independent normal, **simulating** gene expressions.



Impact on Statistical Inference

False statistical inferences: If $Y = Z_1$ and fit

$$Y = \mathbf{X}_{\hat{M}}^T \beta + \varepsilon,$$

the residual variance

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T(\mathbf{I}_n - \mathbf{P}_{\hat{M}})\mathbf{y}}{n - \hat{s}} = (1 - \gamma_n^2)\frac{\|\varepsilon\|^2}{n - \hat{s}},$$

Fraction of bias: $\gamma_n^2 = \varepsilon^T \mathbf{P}_{\hat{M}} \varepsilon / \|\varepsilon\|^2 = O_P(\hat{\mathbf{s}} \log \mathbf{p}/\mathbf{n})$.

Naive two-stage: Use the **selected** model and refit the data.

Seriously underestimate the variance.

Impact on Statistical Inference

False statistical inferences: If $Y = Z_1$ and fit

$$Y = \mathbf{X}_{\hat{M}}^T \beta + \varepsilon,$$

the residual variance

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T(\mathbf{I}_n - \mathbf{P}_{\hat{M}})\mathbf{y}}{n - \hat{s}} = (1 - \gamma_{\mathbf{n}}^{\mathbf{2}})\frac{\|\varepsilon\|^2}{n - \hat{s}},$$

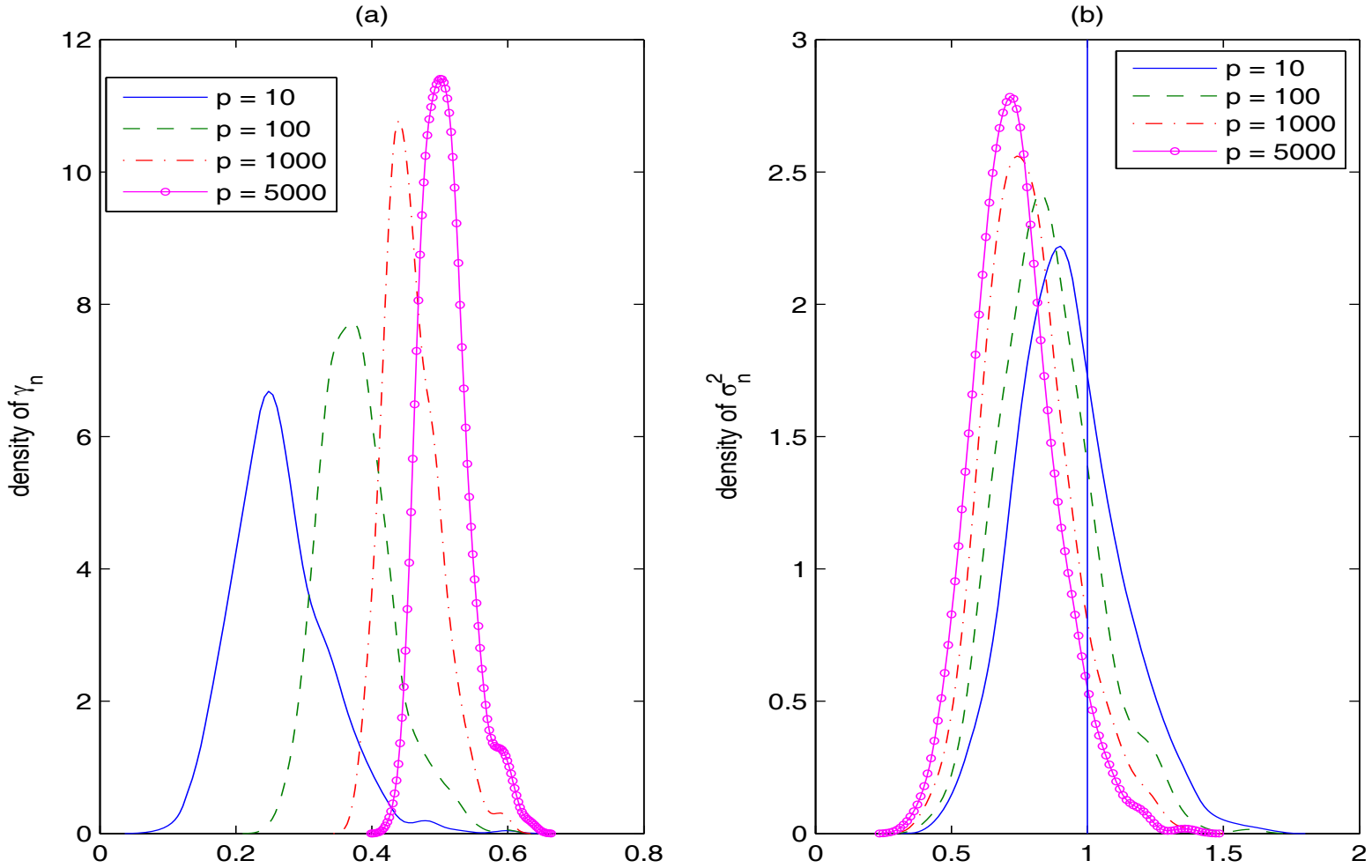
Fraction of bias: $\gamma_n^2 = \varepsilon^T \mathbf{P}_{\hat{M}} \varepsilon / \|\varepsilon\|^2 = O_P(\hat{\mathbf{s}} \log \mathbf{p}/\mathbf{n})$.

Naive two-stage: Use the **selected** model and refit the data.

Seriously underestimate the variance.

Impact of spurious correlation on variance est (I)

- $\hat{s} = 1$ with dimensionality p various.



Spurious variables predict realized noises

Data Generating Process: $Y = 2X_1 + 0.3X_2 + \varepsilon$

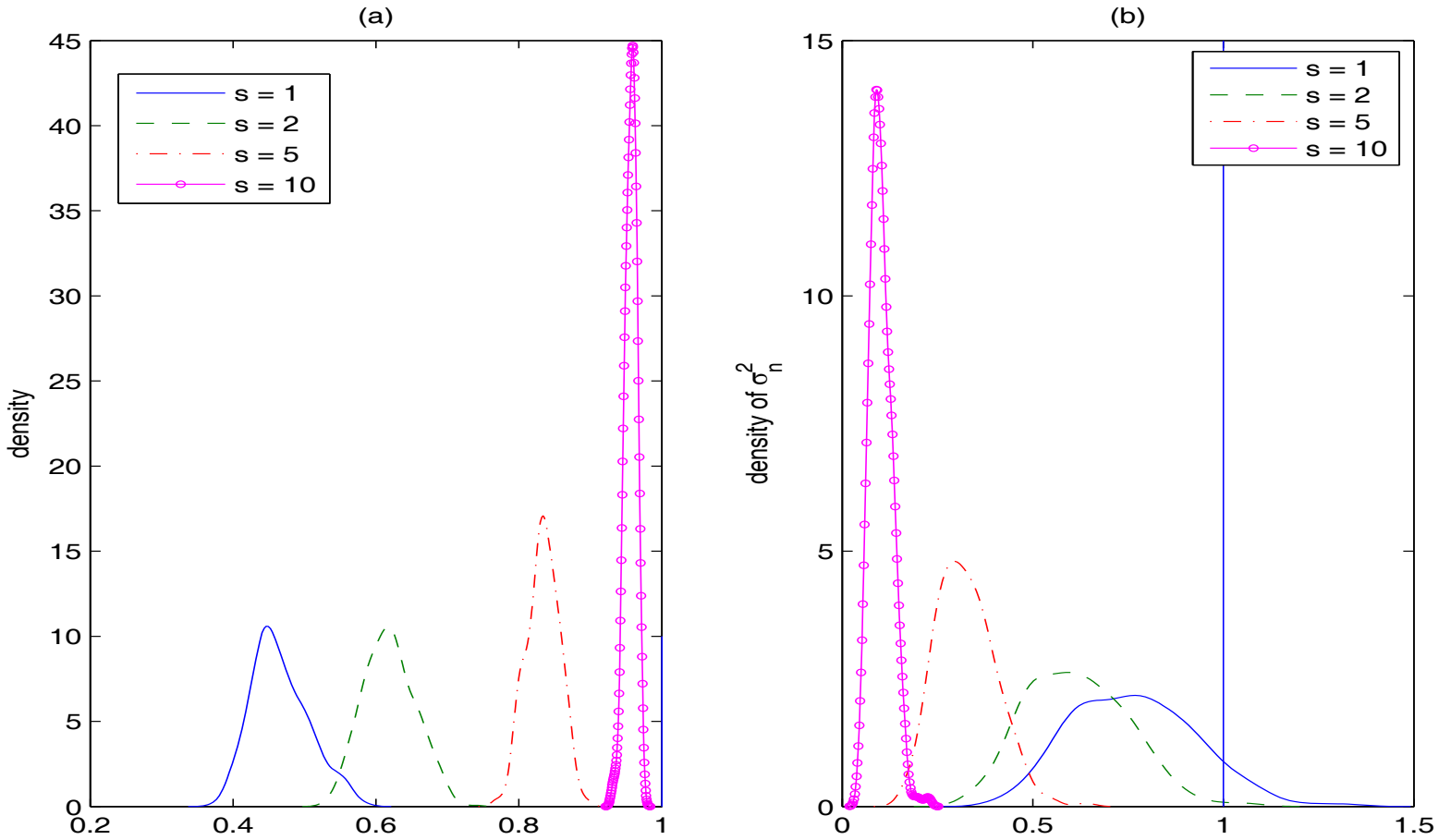
Spurious variables: selected to predict realized noise.

Stepwise addition: Selected **coordinated** variables to best predict ε.

- The more spurious variables, the better realized noises are predicted.

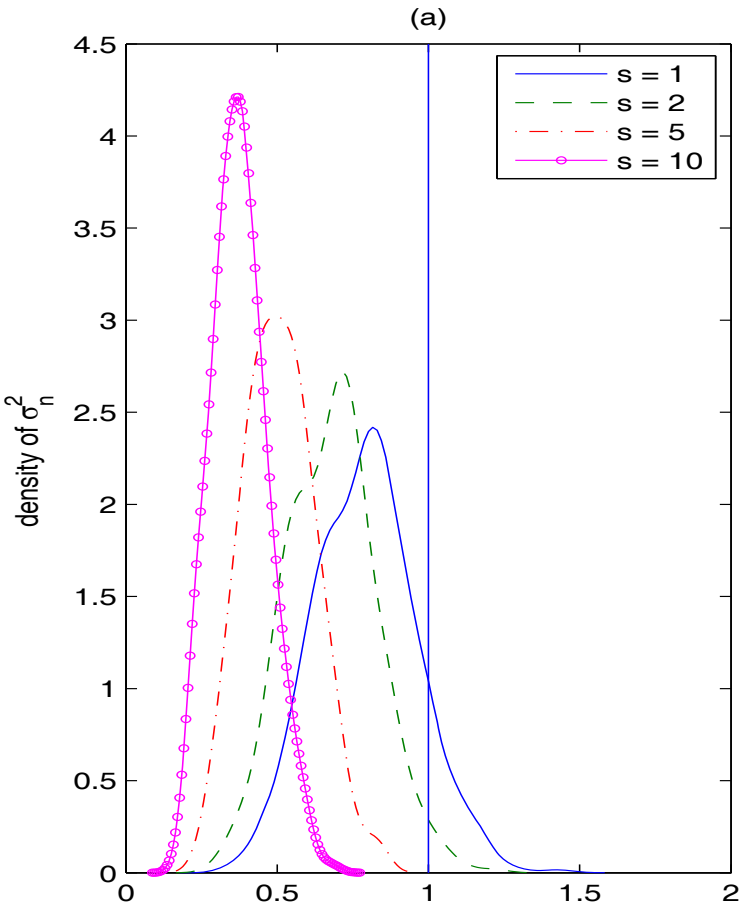
Impact of spurious correlation on variance est (II)

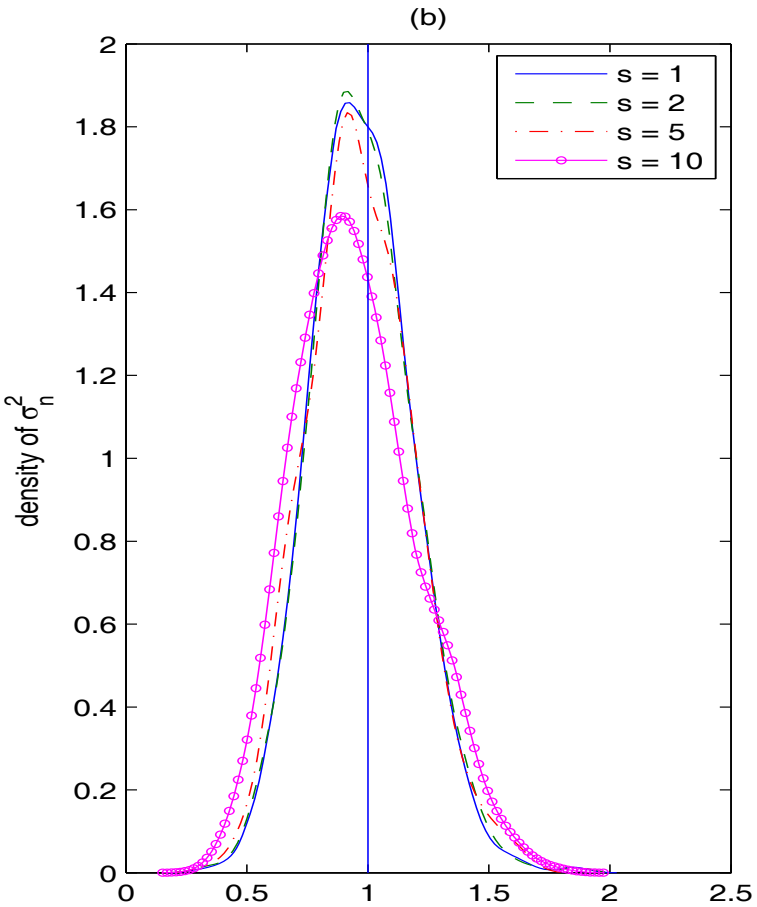
- $p = 1000, n = 50$ with various spurious variables $\hat{s}$.
- stepwise addition algorithm.



Naive two-stage and RCV

- $p = 1000, n = 50$ with various spurious variables $\hat{s}$.
- Correlation screening (SIS).





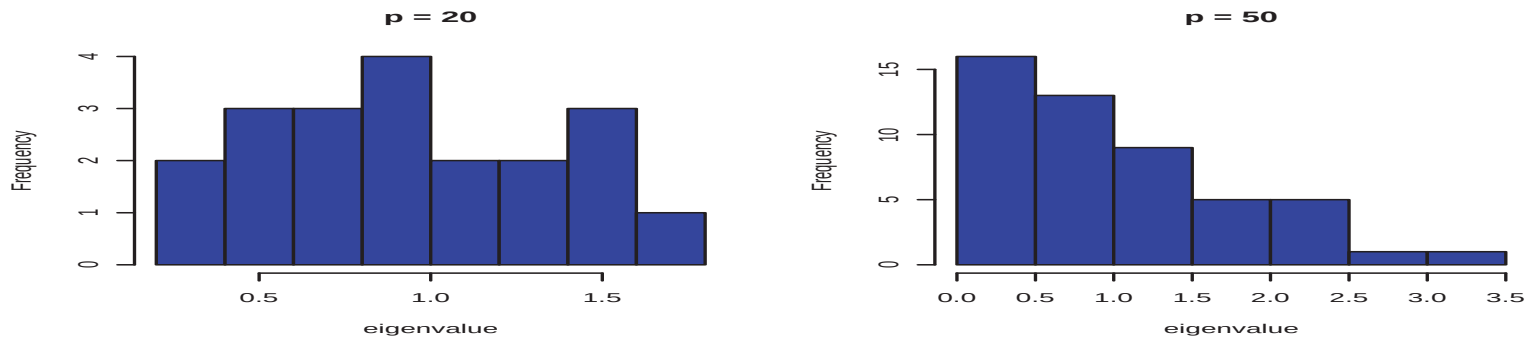
3. Estimated Covariance Matrices

Spectral distribution: of engenvalues $\lambda_1, \cdots, \lambda_p$ of Σ.

Identity matrix: $\Sigma = I_p$. $\lambda_1 = \cdots = \lambda_p = 1$.

Data: $\mathbf{X}_1, \cdots, \mathbf{X}_n \sim_{i.i.d.} N(0, I_p)$. Let $\hat{\Sigma}_n$ be the sample cov.

What is the spectral distribution of $\hat{\Sigma}$?



Low-dim	Moderate-dim	High-dim	Ultra-High
$p \ll n$	$p = cn, c < 1$	$p = cn, c > 1$	$p \gg n$
δ_1	Tracy-Wisdom Law	Mixture	δ_0

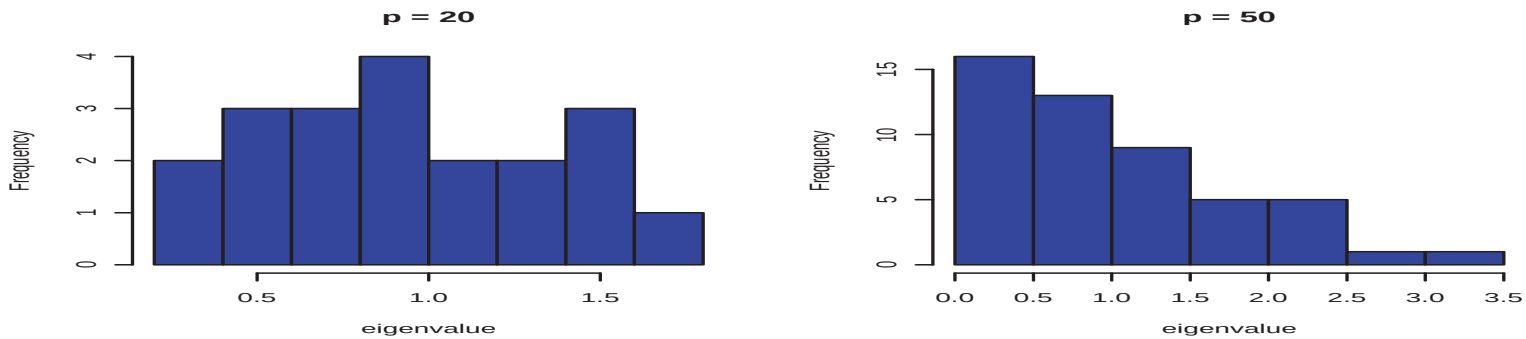
3. Estimated Covariance Matrices

Spectral distribution: of engenvalues $\lambda_1, \cdots, \lambda_p$ of Σ.

Identity matrix: $\Sigma = I_p$. $\lambda_1 = \cdots = \lambda_p = 1$.

Data: $\mathbf{X}_1, \cdots, \mathbf{X}_n \sim_{i.i.d.} N(0, I_p)$. Let $\hat{\Sigma}_n$ be the sample cov.

What is the spectral distribution of $\hat{\Sigma}$?



Low-dim	Moderate-dim	High-dim	Ultra-High
$p \ll n$	$p = cn, c < 1$	$p = cn, c > 1$	$p \gg n$
δ_1	Tracy-Wisdom Law	Mixture	δ_0

Curse of Ultrahigh Dimensionality

- Computational cost
- Stability
- Estimation accuracy: ★noise accumulation ★spurious corr

Key Idea: **Large-scale** screening + **moderate-scale** searching.

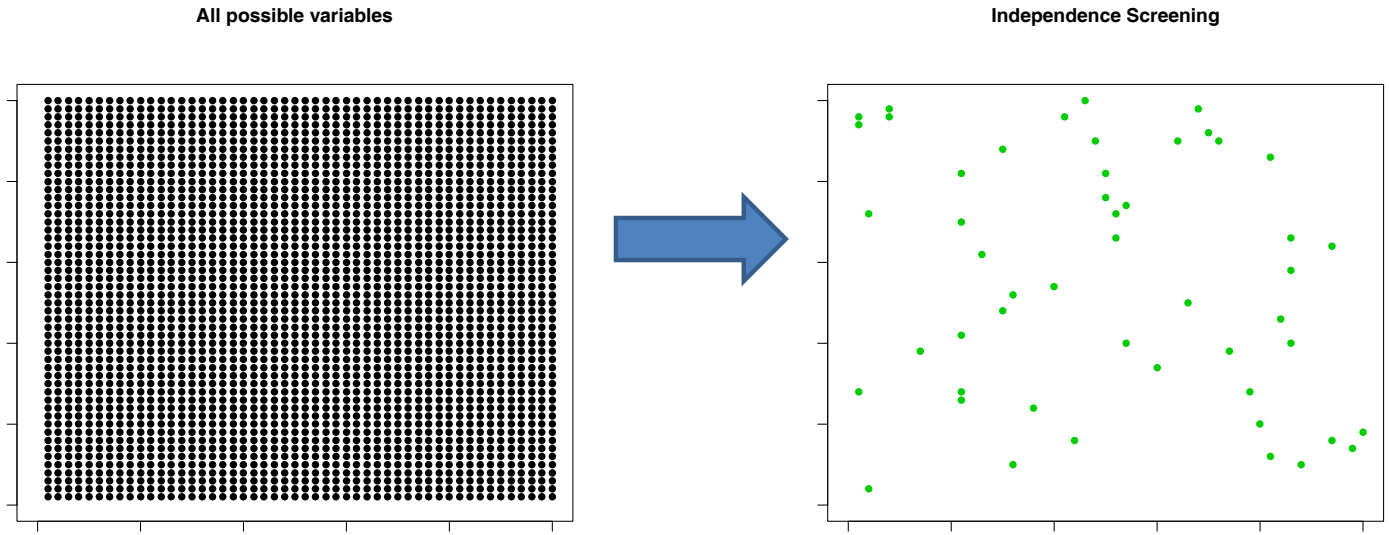
Penalized quasi-likelihood

a moderate-scale selection

Folded concave penalized quasi-likelihood

$$Q(\beta) = n^{-1} \sum_{i=1}^{n} L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \qquad \textit{(Fan \& Li, 01)}$$

- Simultaneously estimate coefs and choose variables.

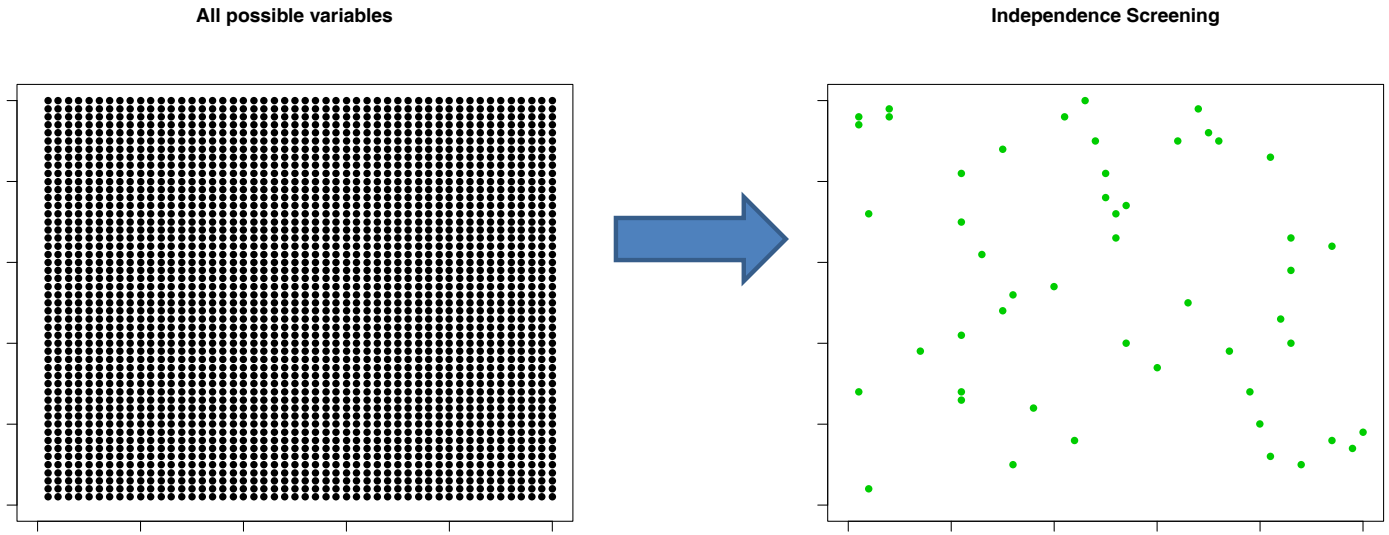


- What is the role of penalty functions?
- Popular choice L_1. Preferred choice: SCAD (**folded-concave**).
  - Better bias property and model selection consistency.

Folded concave penalized quasi-likelihood

$$Q(\beta) = n^{-1} \sum_{i=1}^{n} L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \qquad \text{(Fan \& Li, 01)}$$

- Simultaneously estimate coefs and choose variables.

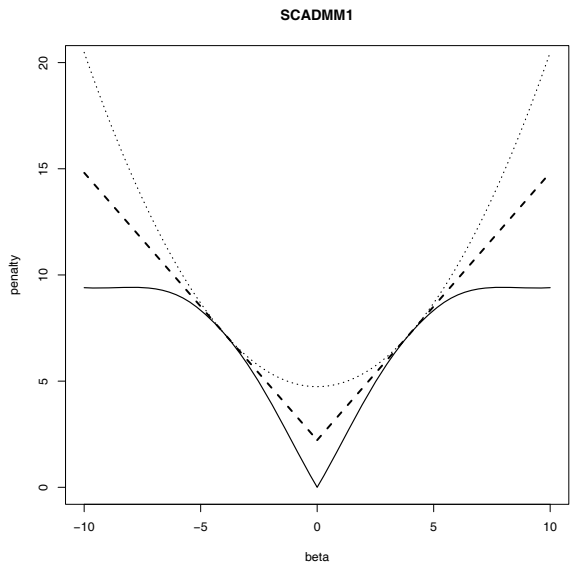


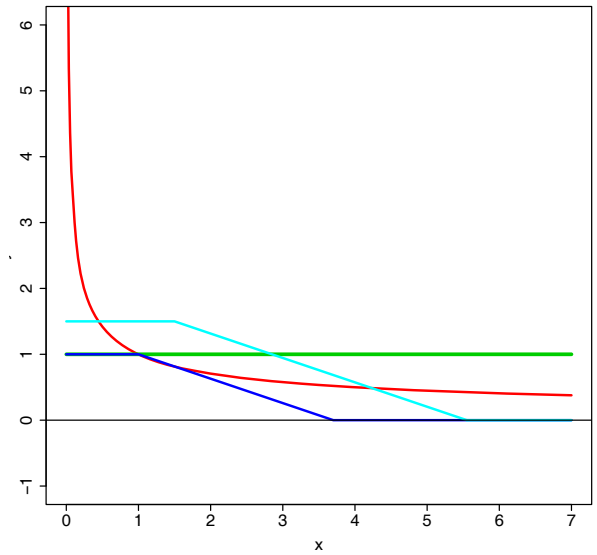
- What is the role of penalty functions?
- Popular choice L_1. Preferred choice: SCAD (**folded-concave**).
  - Better bias property and model selection consistency.

Iterated reweighted LASSO

$$Q(\beta) \approx n^{-1} \sum_{i=1}^{n} L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^{p} \left\{ p_\lambda(|\beta_j^{(k)}|) + \mathbf{p}'_\lambda(|\beta_\mathbf{j}^{(\mathbf{k})}|)(|\beta_\mathbf{j}| - |\beta_j^{(k)}|) \right\}.$$

$$Q^{\mathbf{app}}(\beta) = n^{-1} \sum_{i=1}^{n} L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^{p} w_j |\beta|_j, \qquad w_j = p'_\lambda(|\beta_j^{(k)}|)$$





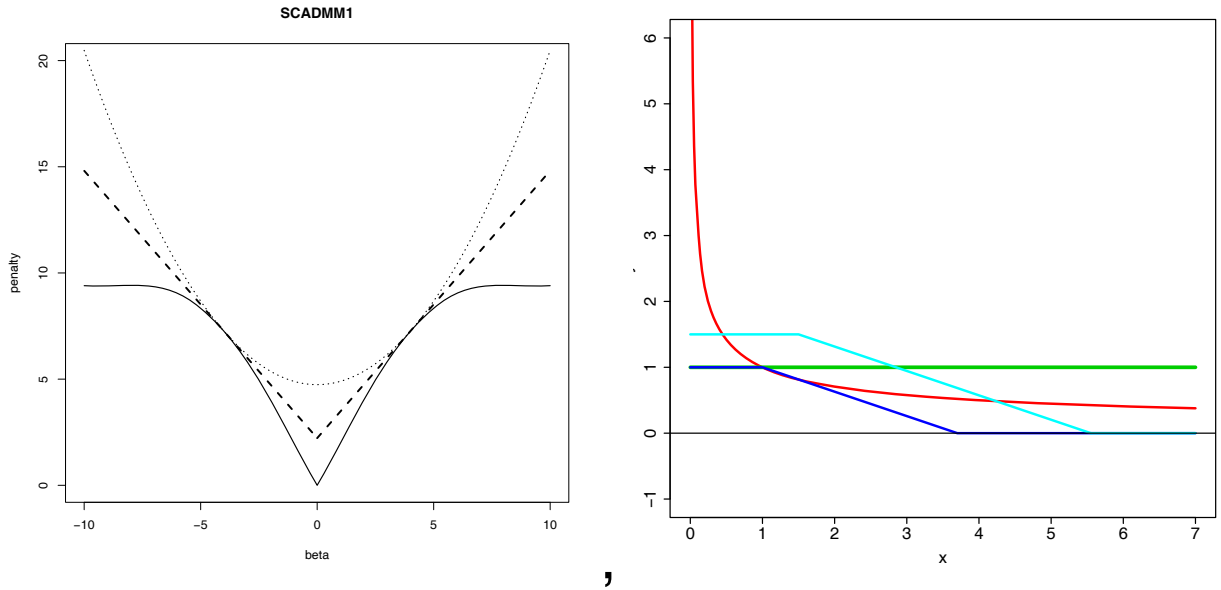
,

- $\beta^{(0)} = 0 \Longrightarrow$ LASSO.
- Iteration reduces the bias
- Zero is a non-absorbing state (comparing $w_j = 1/|\beta_j^{(k)}|^\gamma$).

Iterated reweighted LASSO

$$Q(\beta) \approx n^{-1} \sum_{i=1}^{n} L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^{p} \left\{ p_\lambda(|\beta_j^{(k)}|) + \mathbf{p}'_\lambda(|\beta_\mathbf{j}^{(\mathbf{k})}|)(|\beta_\mathbf{j}| - |\beta_j^{(k)}|) \right\}.$$

$$Q^{\mathbf{app}}(\beta) = n^{-1} \sum_{i=1}^{n} L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^{p} w_j |\beta|_j, \qquad w_j = p'_\lambda(|\beta_j^{(k)}|)$$



- $\beta^{(0)} = 0 \Longrightarrow$ LASSO.
- Iteration reduces the bias
- Zero is a non-absorbing state (comparing $w_j = 1/|\beta_j^{(k)}|^\gamma$).

Remarks

Other algorithms: **LQA** (*Fan & Li, 01*); **LLA** (*Zou & Li, 08*); **PLUS** (*Zhang, 09*); **Coordinate optimization** (*Fu & Jiang, 99*).

Capacity: handle NP-dimensionality with wider capacity.

possesses an oracle property (*Fan & Lv, 09*), reducing the bias of LASSO.

Remarks

Other algorithms: **LQA** (*Fan & Li, 01*); **LLA** (*Zou & Li, 08*); **PLUS** (*Zhang, 09*); **Coordinate optimization** (*Fu & Jiang, 99*).

Capacity: handle NP-dimensionality with wider capacity.

possesses an oracle property (*Fan & Lv, 09*),
reducing the bias of LASSO.

Remarks

Other algorithms: **LQA** (*Fan & Li, 01*); **LLA** (*Zou & Li, 08*); **PLUS** (*Zhang, 09*); **Coordinate optimization** (*Fu & Jiang, 99*).

Capacity: handle NP-dimensionality with wider capacity.

- possesses an oracle property (*Fan & Lv, 09*), reducing the bias of LASSO.

The ISIS Method

a two-scale framework

A two-scale method

Key Idea: **Large-scale** screening + **moderate-scale** searching.

Illustration of ISIS

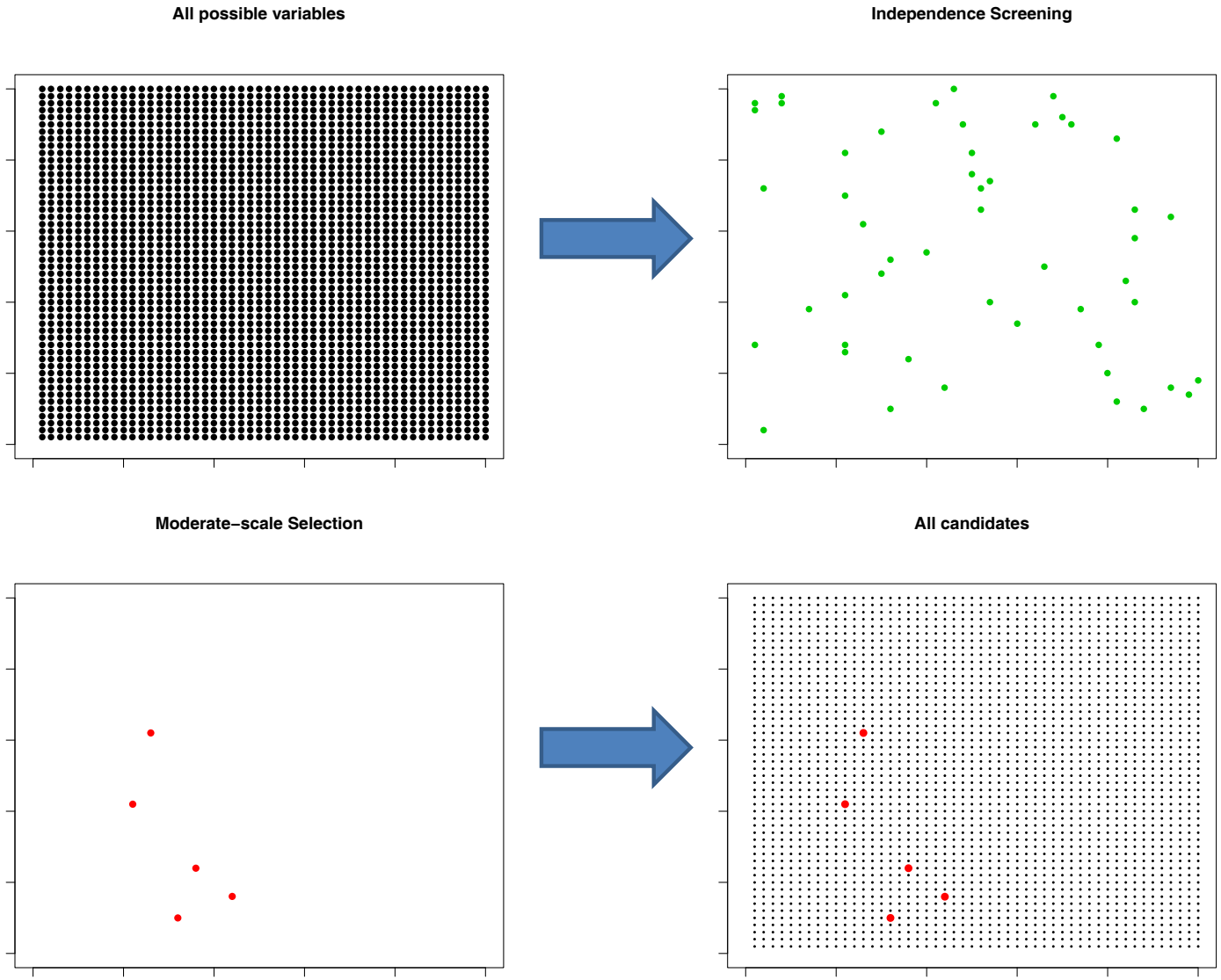
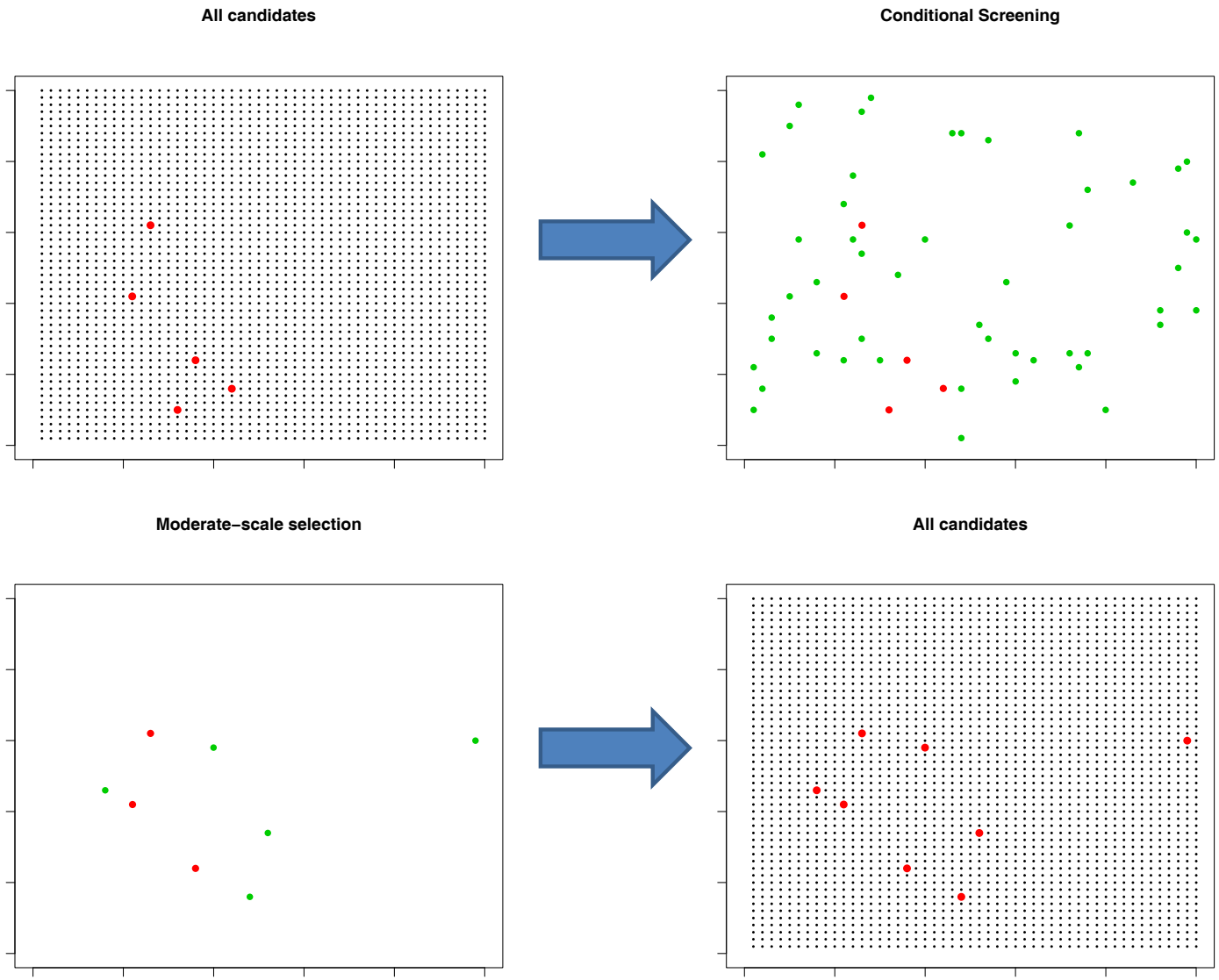
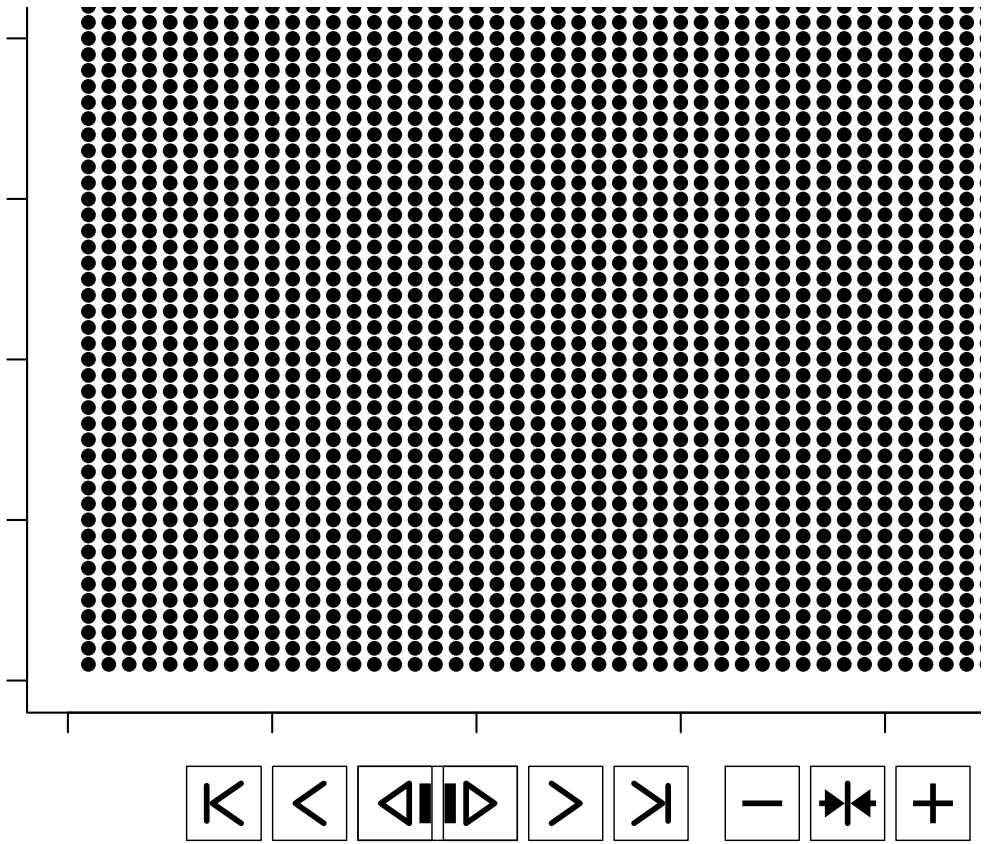


Illustration of ISIS



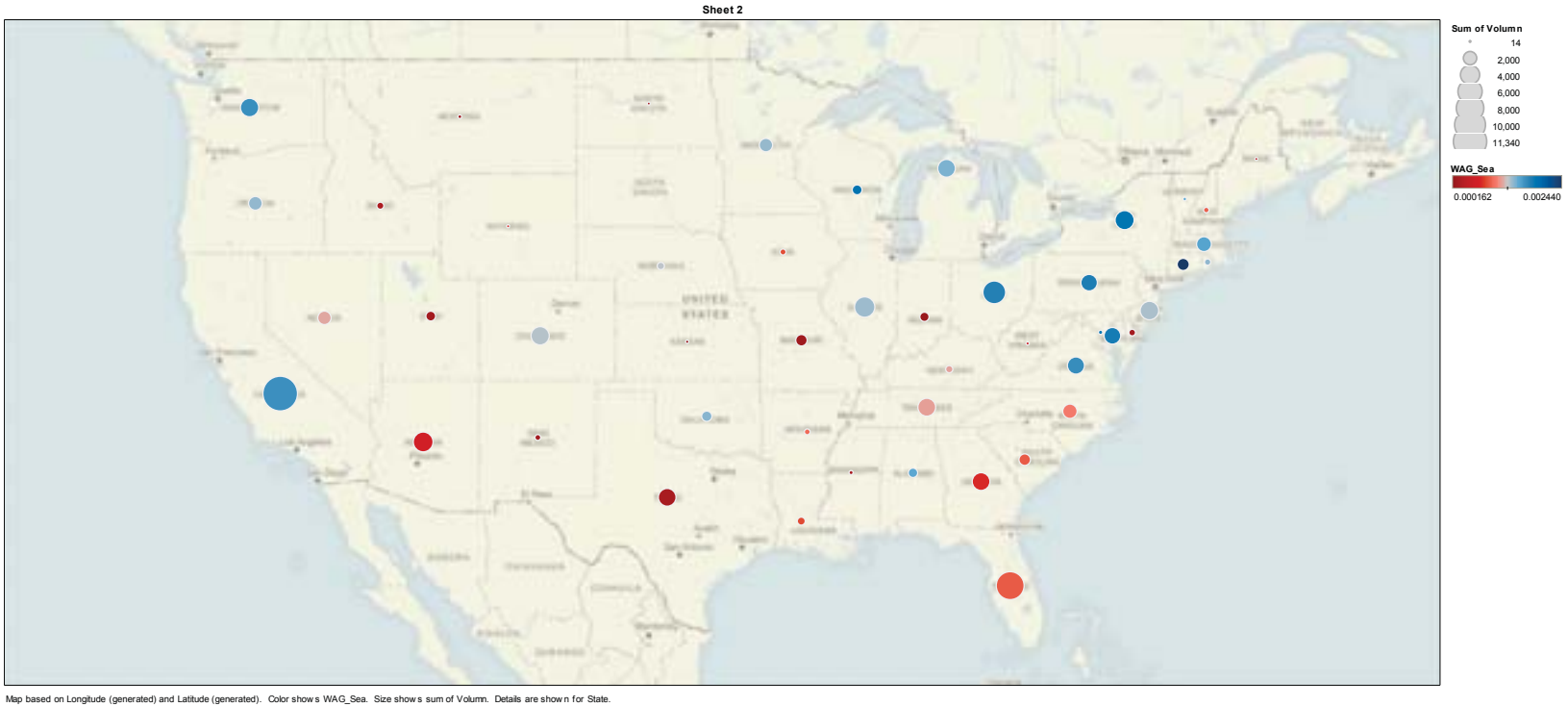


A case study

Forecasting home price appreciation

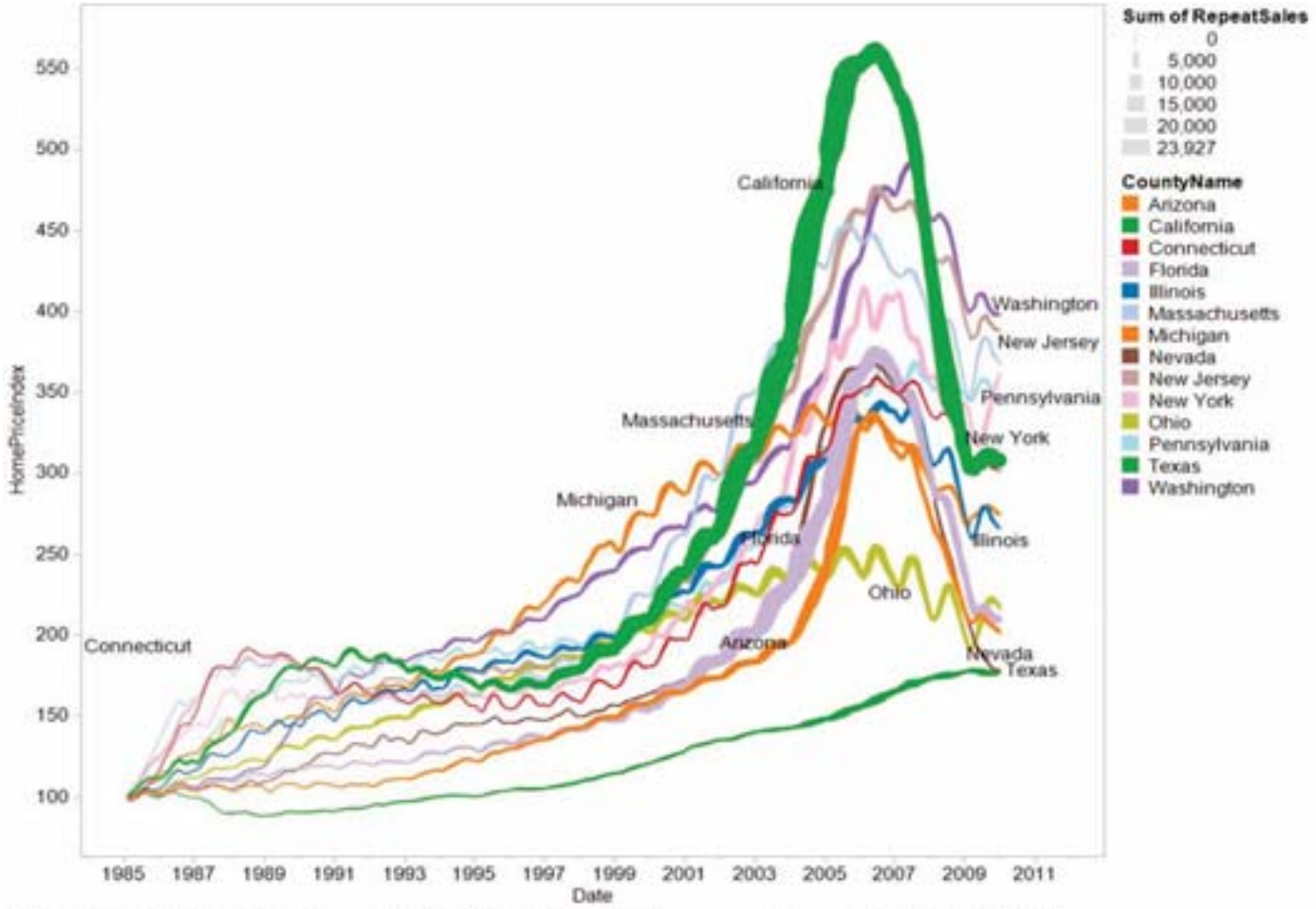
The Data and Objective

Data: HPA collected at ”≈” 1000 CBSA.



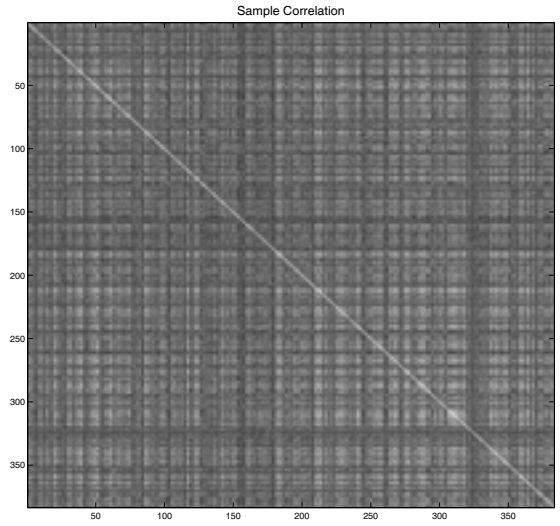
Objective: To project HPA over 30-40 years for approx 1000 CBSAs based on national assumption.

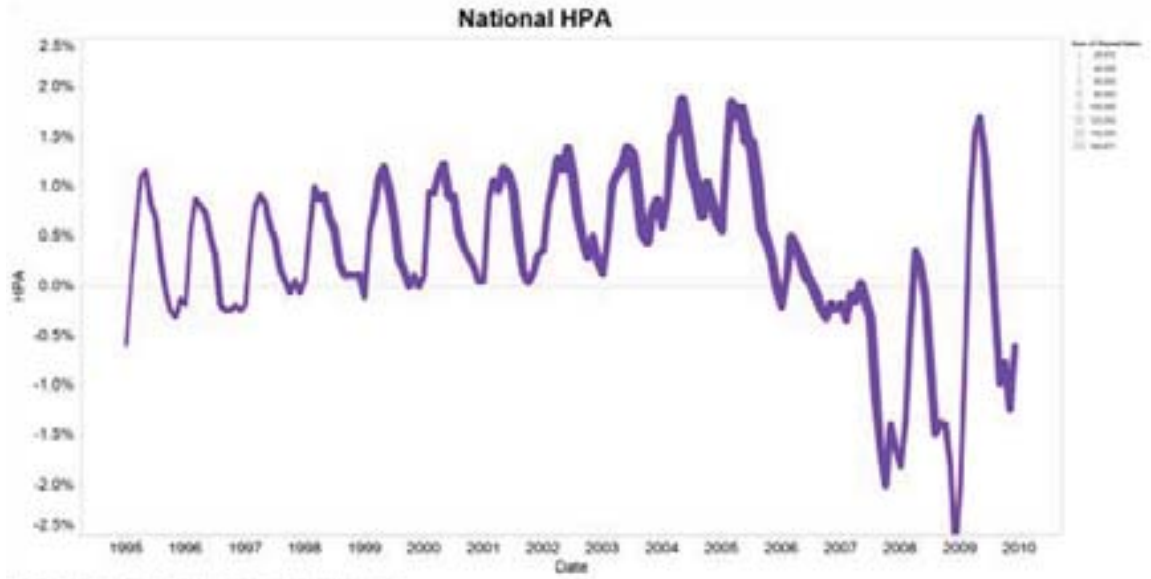
Some Examples

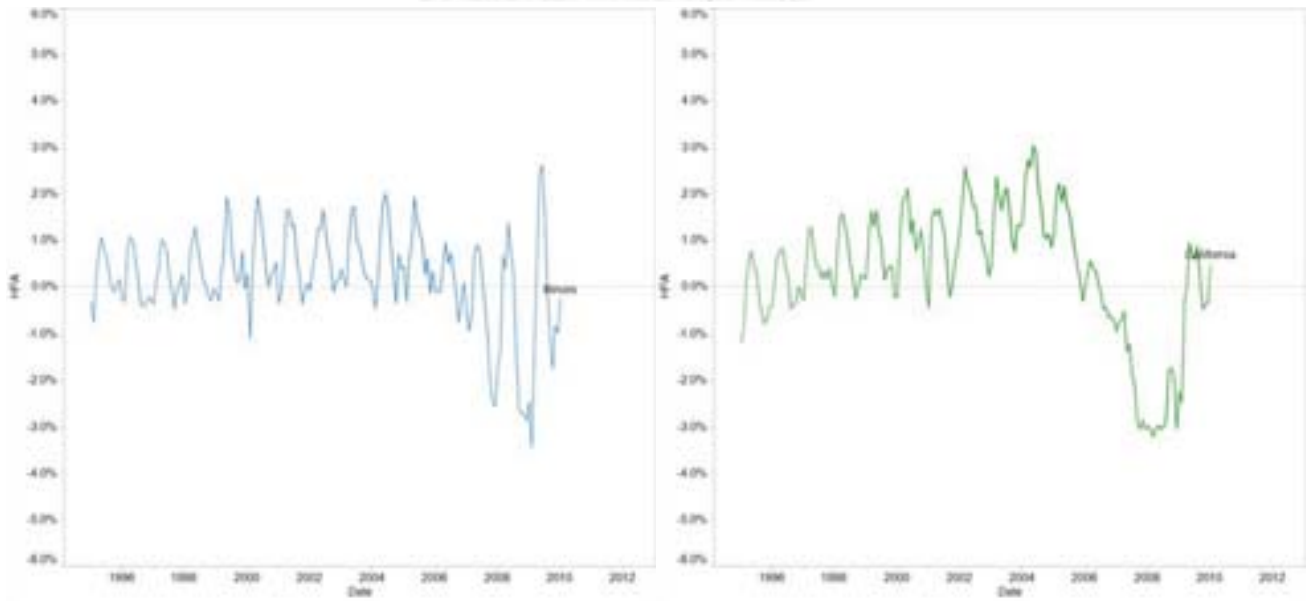


Date vs. HomePriceIndex. Color shows details about CountyName. Size shows sum of RepeatSales. The view is filtered on CountyName and Date. The CountyName filter keeps 14 members. The Date filter ranges from 1/22/1985 to 1/1/2010. The marks are labeled by CountyName.

Location Correlation and Seasonality







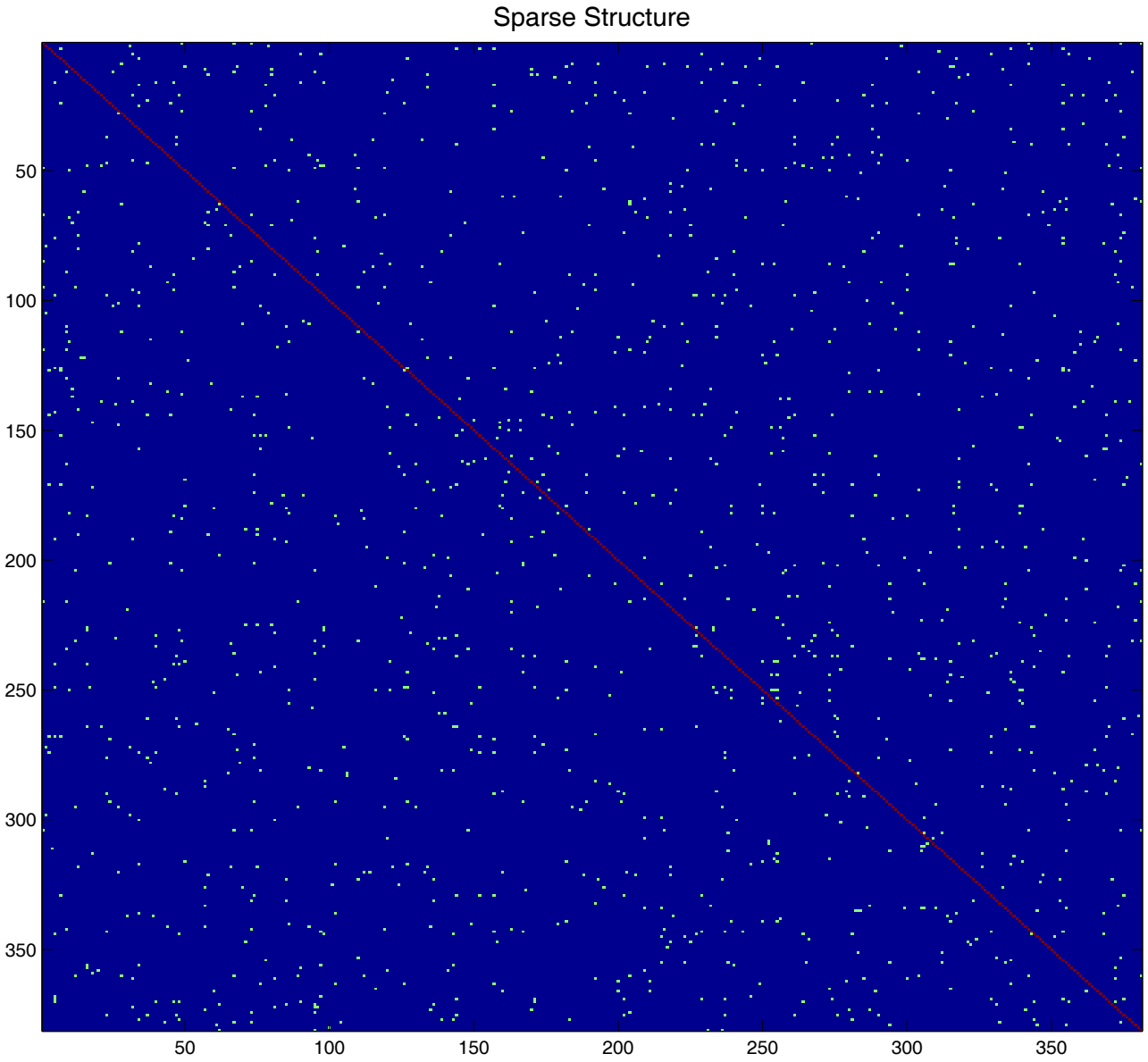
Conditional sparsity

Model: Y_{t+1} is the HPA in one CBSA:

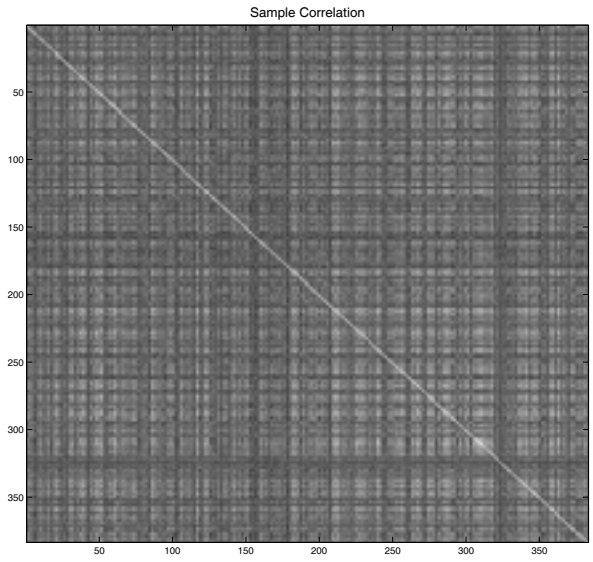
$$Y_{t+1} = \beta_0 + \beta_1 \mathbf{X_{N,t}} + \sum_{j=1}^{381} \beta_j X_{t,j} + \varepsilon_t$$

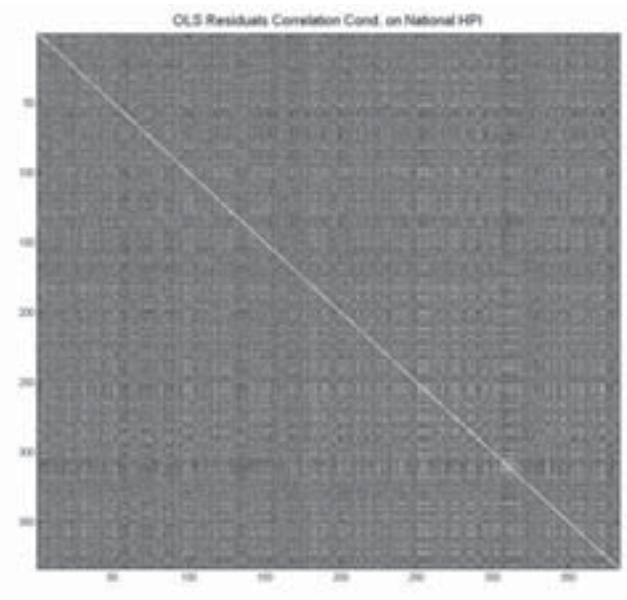
- $\{\beta_j\}_{j=2}^{381}$ are sparse
- Explored by penalized least-squares with SCAD and LLA
- Results 30% more accurate than the simple time series modeling

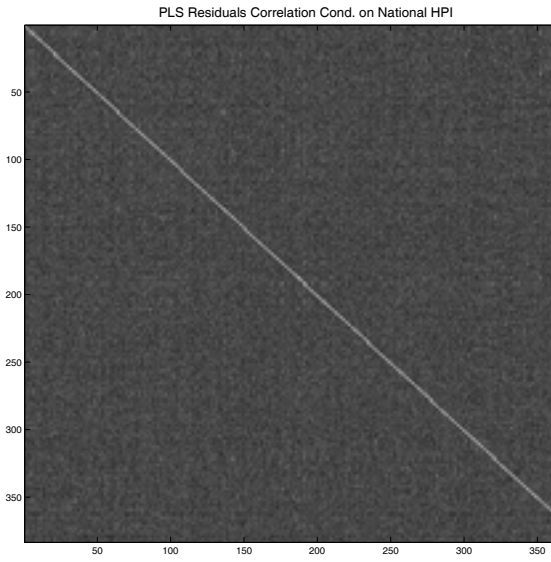
Local neighborhood selection

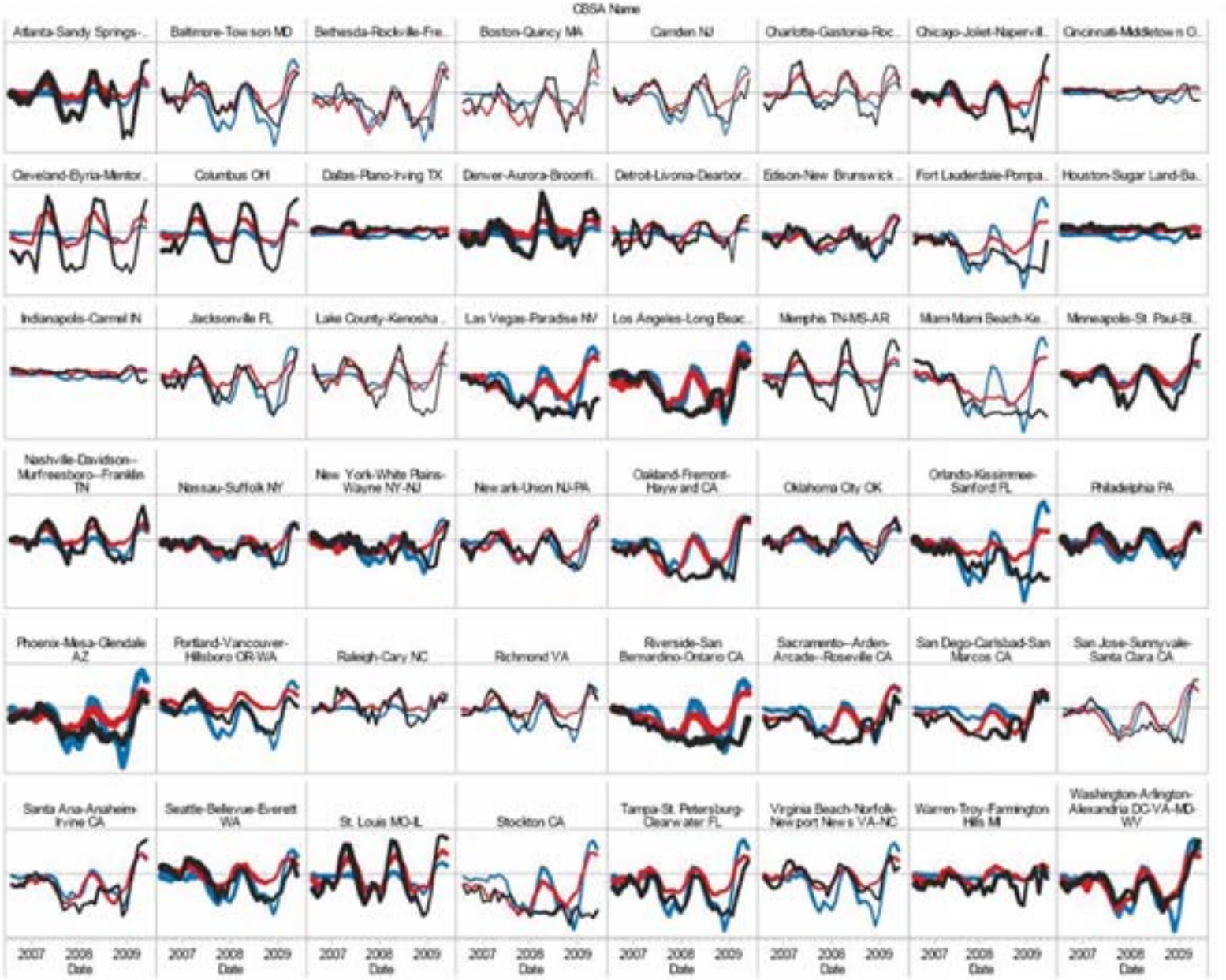


Effectiveness of sparse modeling








CBSA Name
Atlanta-Sandy Springs-...
Baltimore-Towson MD
Bethesda-Rockville-Fre...
Boston-Quincy MA
Camden NJ
Charlotte-Gastonia-Roc...
Chicago-Joliet-Naperville...
Cincinnati-Middletown O...
Cleveland-Elyria-Mentor...
Columbus OH
Dallas-Plano-Irving TX
Denver-Aurora-Broomfi...
Detroit-Livonia-Dearbor...
Edison-New Brunswick ...
Fort Lauderdale-Pompa...
Houston-Sugar Land-Ba...
Indianapolis-Carmel IN
Jacksonville FL
Lake County-Kenosha ...
Las Vegas-Paradise NV
Los Angeles-Long Beac...
Memphis TN-MS-AR
Miami-Miami Beach-Ke...
Minneapolis-St. Paul-Bl...
Nashville-Davidson--Murfreesboro--Franklin TN
Nassau-Suffolk NY
New York-White Plains-Wayne NY-NJ
Newark-Union NJ-PA
Oakland-Fremont-Hayward CA
Oklahoma City OK
Orlando-Kissimmee-Sanford FL
Philadelphia PA
Phoenix-Mesa-Glendale AZ
Portland-Vancouver-Hillsboro OR-WA
Raleigh-Cary NC
Richmond VA
Riverside-San Bernardino-Ontario CA
Sacramento--Arden-Arcade--Roseville CA
San Diego-Carlsbad-San Marcos CA
San Jose-Sunnyvale-Santa Clara CA
Santa Ana-Anaheim-Irvine CA
Seattle-Bellevue-Everett WA
St. Louis MO-IL
Stockton CA
Tampa-St. Petersburg-Clearwater FL
Virginia Beach-Norfolk-Newport News VA-NC
Warren-Troy-Farmington Hills MI
Washington-Arlington-Alexandria DC-VA-MD-WV
2007
2008
2009
Date

Summary

- High-dimensionality and massive data collection characterize many contemporary statistical problems from frontiers of science, engineering and humanities.
- Impact of dimensionality: ★noise accumulation; ★spurious correlation; ★intensive computation
- Massive data collections and new scientific research have strong impact on mathematical thinking, methodological development, scientific computing and theoretical studies:

Summary

- High-dimensionality and massive data collection characterize many contemporary statistical problems from frontiers of science, engineering and humanities.
- Impact of dimensionality: ★noise accumulation; ★spurious correlation; ★intensive computation
- Massive data collections and new scientific research have strong impact on mathematical thinking, methodological development, scientific computing and theoretical studies:

Summary

- High-dimensionality and massive data collection characterize many contemporary statistical problems from frontiers of science, engineering and humanities.
- Impact of dimensionality: ★noise accumulation; ★spurious correlation; ★intensive computation
- Massive data collections and new scientific research have strong impact on mathematical thinking, methodological development, scientific computing and theoretical studies:

Summary

- High-dimensionality and massive data collection characterize many contemporary statistical problems from frontiers of science, engineering and humanities.
- Impact of dimensionality: ★noise accumulation; ★spurious correlation; ★intensive computation
- Massive data collections and new scientific research have strong impact on mathematical thinking, methodological development, scientific computing and theoretical studies:

Conclusion

- The exciting developments in frontiers of science and technology clearly represent the golden opportunities for mathematical sciences with significant challenges.

- Mathematical sciences will grow stronger when they confront the problems of high societal impacts while providing fundamental understanding to these problems and their associated methods that push theory, methods, computation and science forward.

Acknowledgement

Thank

You