Transelliptical Component Analysis

Fang Han
Johns Hopkins University

Han Liu
Princeton University

Neural Information Processing Systems (NIPS)

Lake Tahoe, Nevada, 2012

General Framework

- Using semiparametric model
- Obtaining nonparametric modeling flexibility
- Achieving nearly optimal parametric rates
- Simple and computational efficient

Outline

- Sparse Principal Component Analysis
- Transelliptical Component Analysis
- Concluding Remarks

Sparse Principal Component Analysis

Leading Eigenvectors Estimation

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n i.i.d d-variate random vectors with covariance matrix $\boldsymbol{\Sigma}$.

Goal: Estimate the top m leading eigenvectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ of $\boldsymbol{\Sigma}$ based on the samples $\{\boldsymbol{X}_i\}$.

Applications

- **Large-scale genomic data**
- **Brain imaging data**
- **Equity data**
- ...

Sparse Leading Eigenvectors

Let $\{\boldsymbol{X}_i : i = 1, \ldots, n\}$ be n random samples from a $d-$variate **Gaussian** or **sub-Gaussian** distribution with covariance matrix $\boldsymbol{\Sigma}$.

Goal: Estimate $\boldsymbol{\theta}_1$ based on $\{\boldsymbol{X}_i\}$, while $\boldsymbol{\theta}_1$ is sparse.

Assumption and Estimation

Assumption 1: Assume that $||\boldsymbol{\theta}_1||_0 = s < n$.

Estimation:

$$\widehat{\boldsymbol{\theta}}_1^* = \underset{\boldsymbol{v} \in \mathbb{R}^d}{\arg\max}\, \boldsymbol{v}^T \boldsymbol{\Sigma}_n \boldsymbol{v}, \quad \text{subject to } \boldsymbol{v} \in \mathbb{S}^{d-1} \cap \mathbb{B}_0(s),$$

where $\boldsymbol{\Sigma}_n$ is the sample covariance matrix, $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$ and $\mathbb{B}_0(s) := \{\boldsymbol{v} \in \mathbb{R}^d : ||\boldsymbol{v}||_0 \leq s\}$.

Rates of Convergence

Assumption 2: Assume that $\boldsymbol{X}$ is Gaussian or sub-Gaussian distributed.

Theorem. Let $||\boldsymbol{\theta}_1||_0 = s$. Then the parametric rate of convergence under the ℓ_2 norm is

$$O_P\left(\frac{1}{\lambda_1 - \lambda_2}\sqrt{\frac{s\log d}{n}}\right),$$

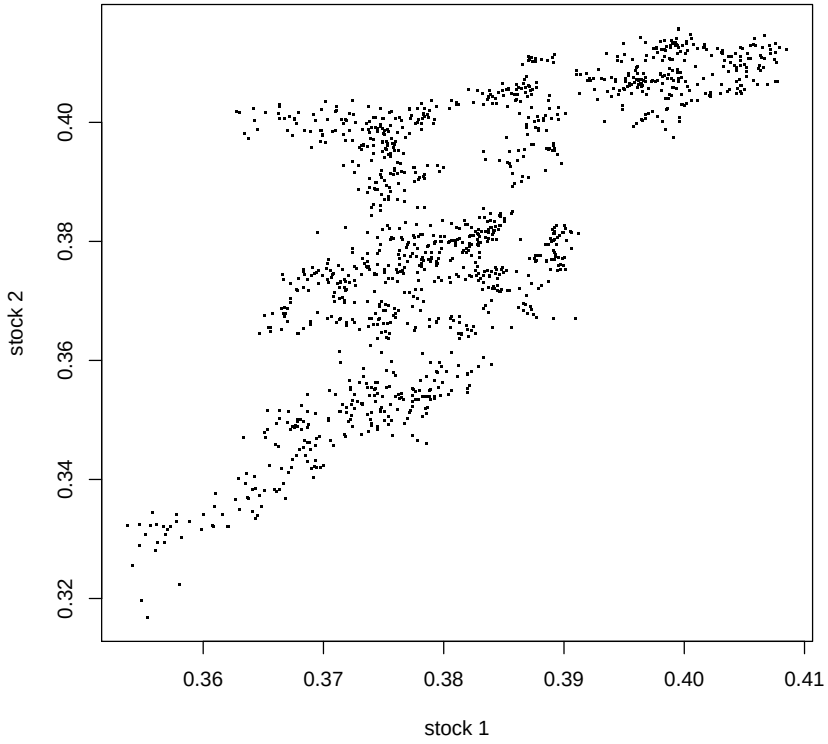
where λ_1 and λ_2 are the top two largest eigenvalues of $\boldsymbol{\Sigma}$.

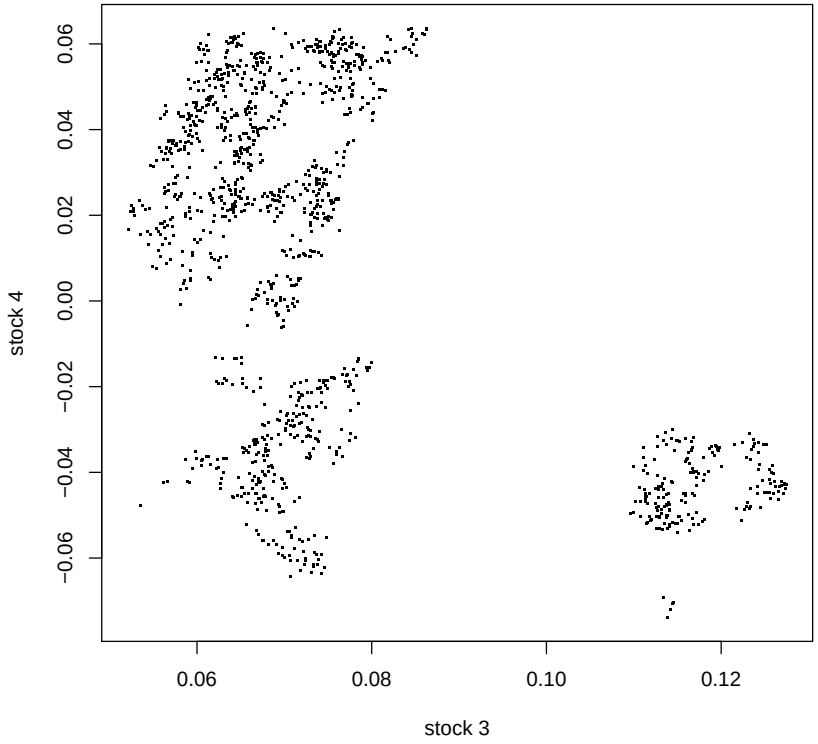
Constraints

The Gaussian or sub-Gaussian assumption is too **constraint**, appearing to be **rare** in applications.

Equity Data

452 stocks that were consistently in the S&P 500 index between January 1, 2003 though January 1, 2008.



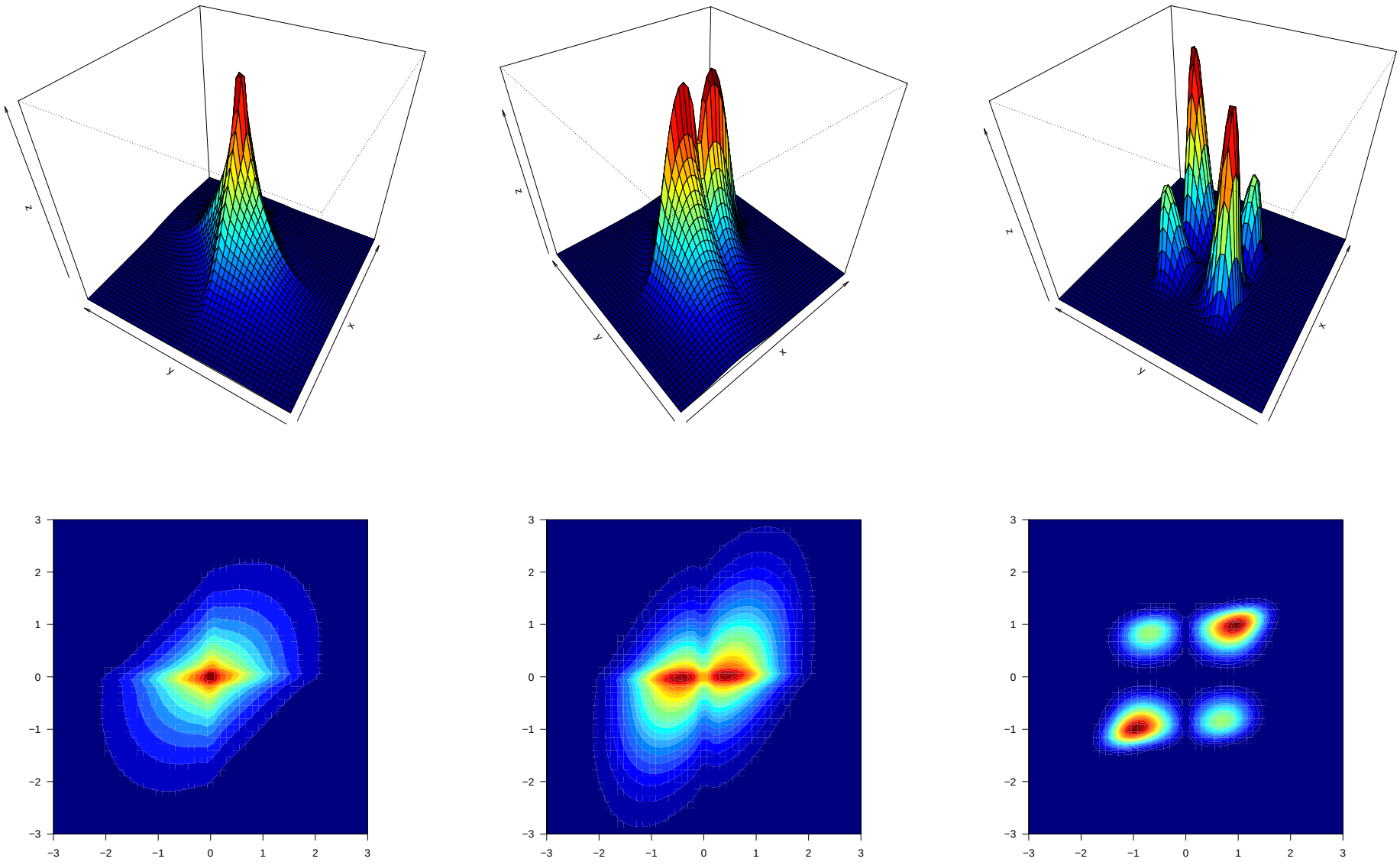


Extensions to non-Gaussian (Abnormal)

Liu, Lafferty and Wasserman (2009, JMLR)

Definition. A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is said to follow a **nonparanormal** distribution if there exists a set of **unspecified** univariate increasing functions $\{f_j\}_{j=1}^d$ such that

$$(f_1(X_1), f_2(X_2), \ldots, f_d(X_d))^T \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}), \quad \text{where } \mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{1}.$$


z
y
x
z
y
x
z
y
x
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Liu et.al. Annals of Statistics (2012)

For each pair (X_j, X_k), **Spearman's rho coefficient** of (X_j, X_k), denoted by $\widehat{\rho}_{jk}$, is the correlation of the ranks. Let

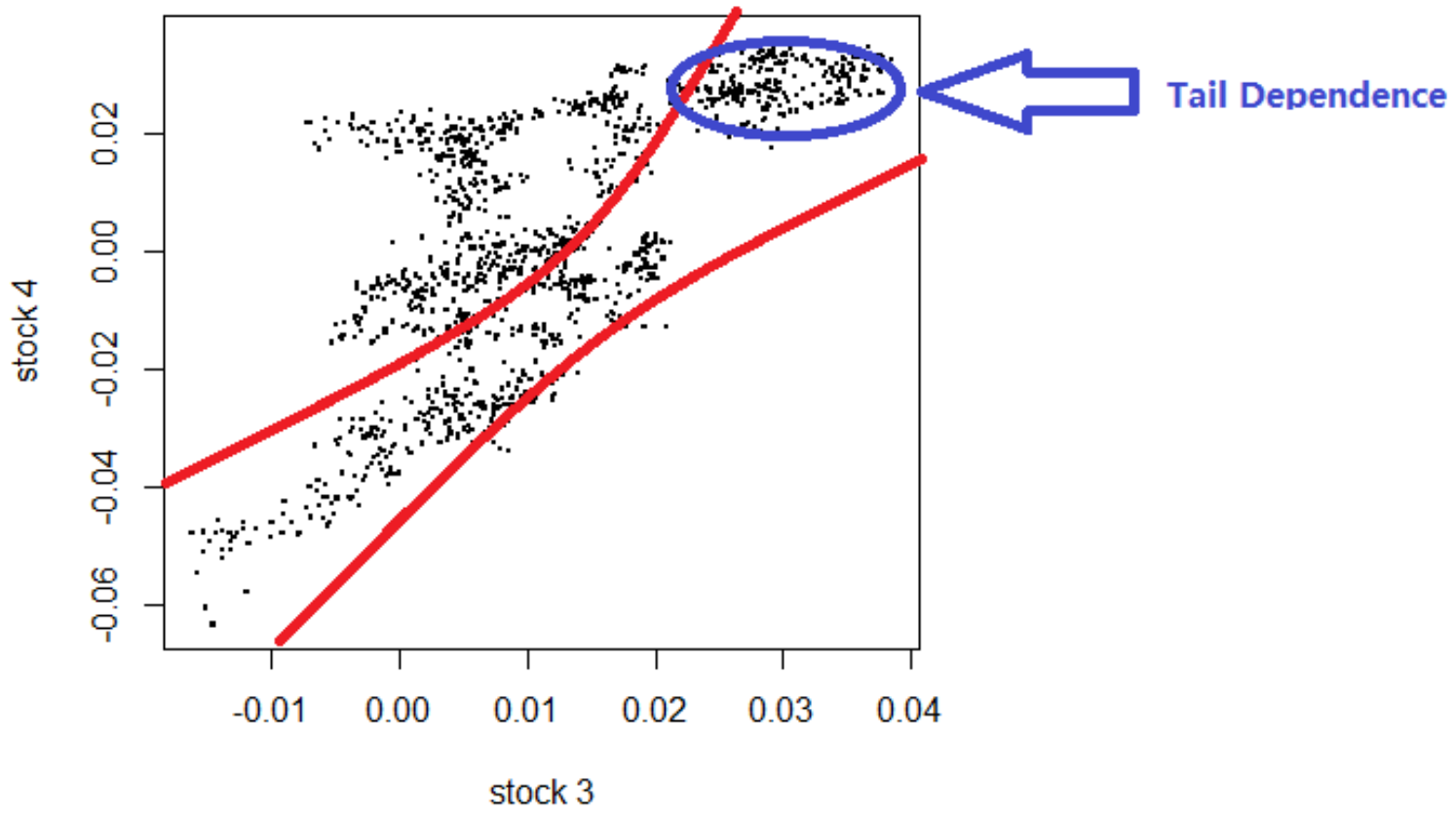
$$\widehat{\mathbf{R}}^{\rho} := \left[2 \sin\left(\frac{\pi}{6}\widehat{\rho}_{jk}\right)\right].$$

Theorem. Confined in the **nonparanormal**, $||\widehat{\mathbf{R}}^{\rho} - \boldsymbol{\Sigma}||_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right)$.

Parametric (optimal) rates in graph recovery and **near-parametric rate** in principal component analysis (Han and Liu, NIPS 2012).

Good Enough?

Theorem. The tail dependence is **zero** in Nonparanormal (Gaussian Copula).




LEHMAN BROTHERS

Elliptical Distribution

When the density exists, the elliptical distribution has the **density:**

$$f(\boldsymbol{x}) = k \cdot g((\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})),$$

where g is an **unspecified** univariate positive function. In this case, we represent it by $EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$.

Note: Elliptical distributions can be **very heavy tailed** and even have **infinite first or second moments**.

Transelliptical Distribution

Definition. A random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is said to follow a **transelliptical** distribution if there exists a set of **unspecified** univariate increasing functions $\{f_j\}_{j=1}^d$ such that

$$(f_1(X_1), f_2(X_2), \ldots, f_d(X_d))^T \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, g), \quad \text{where } \operatorname{diag}(\boldsymbol{\Sigma}) = \mathbf{1}.$$

In this case, we represent $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, g; f_1, \ldots, f_d)$.

Kendall's tau and Its Invariance Property

Population Kendall's tau:

$$\tau(X_j, X_k) = \mathbb{P}((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) > 0) - \mathbb{P}((X_j - \widetilde{X}_j)(X_k - \widetilde{X}_k) < 0),$$

where $(\widetilde{X}_j, \widetilde{X}_k)$ is a independent copy of (X_j, X_k).

Theorem. Given $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, g; f_1, \ldots, f_d)$,

$$\boldsymbol{\Sigma}_{jk} = \sin\left(\frac{\pi}{2}\tau(X_j, X_k)\right)$$

.

Invariant to both g and f_j.

Transelliptical Component Analysis

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n independent realizations of $\boldsymbol{X}$. Using the Kendall's tau correlation coefficient estimate:

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \text{sign}\left(x_{ij} - x_{i'j}\right)\left(x_{ik} - x_{i'k}\right).$$

Let

$$\widehat{\mathbf{R}}^{\tau} := \left[\sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right)\right].$$

Plugging $\widehat{\mathbf{R}}^{\tau}$ into **any sparse principal component algorithm.**

Theoretical Results

In estimating $\boldsymbol{\theta}_1$ and its support, we have

- the near-parametric rate of convergence under the ℓ_2 norm:

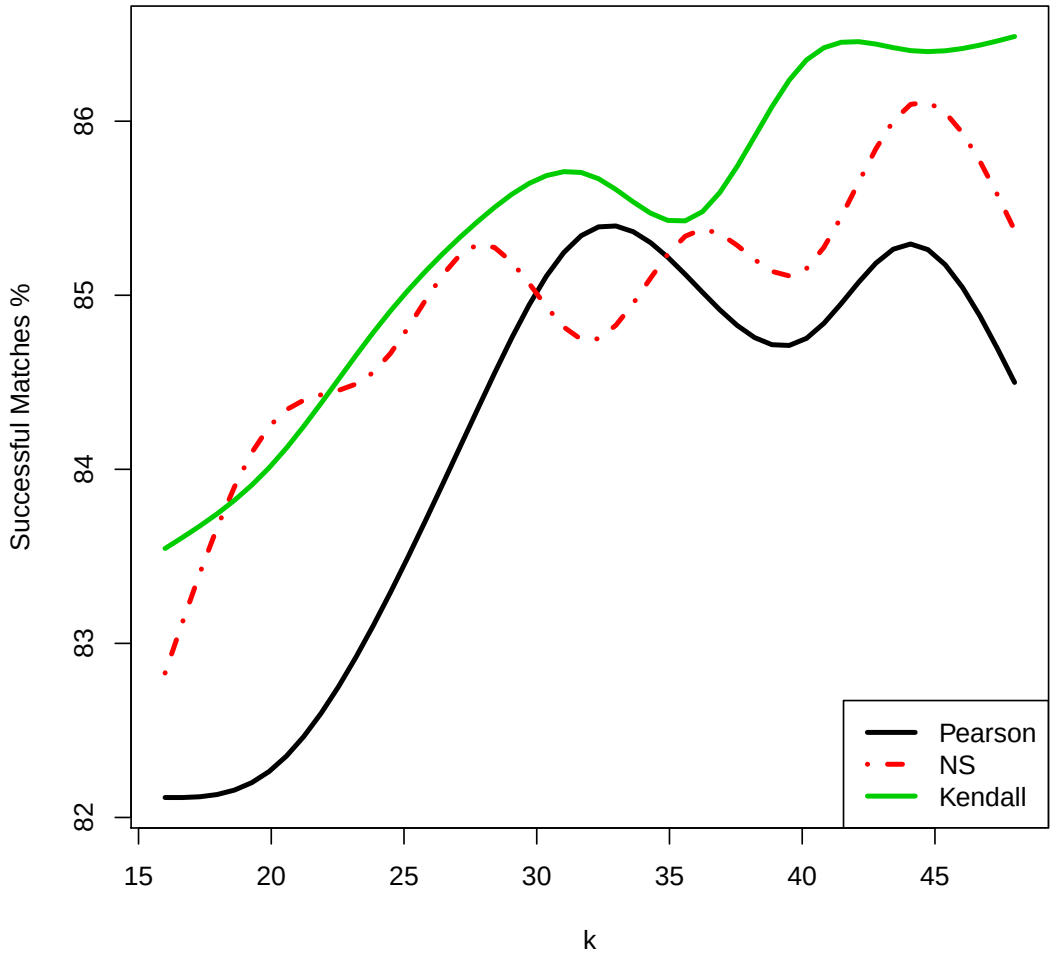
$$O_P\left(\frac{s}{\lambda_1 - \lambda_2}\sqrt{\frac{\log d}{n}}\right)$$

- a support recovery threshold at an order of

$$O_P\left(\frac{s}{\lambda_1 - \lambda_2}\sqrt{\frac{\log d}{n}}\right)$$

Equity Data Again

Prediction of the Market Trend:



Beyond PCA

- **Directed Graphs Estimation**
- **Un-directed Graphs Estimation**
- **Discriminant Analysis**
- **Independent Component Analysis**
- **Canonical Component Analysis**
- ...

Remarks

- **Model Flexibility**
- **Parametric or near-parametric rate**
- **Procedure simple and computation efficiency**

Thanks!

for more information, please come to the poster session (Tu66).