Latent and Network Models with Applications to Finance

Jingchen Liu

Department of Statistics
Columbia University

Joint work with Yunxiao Chen, Xiaoou Li, and Zhiliang Ying

At ISFA-Columbia Workshop, June 28, 2016

November 16, 2015

Modeling multivariate distribution

- Multivariate random vector: $(R_1, ..., R_J)$
- Continuous vectors: multivariate Gaussian, multivariate t-distribution...
- Categorical vectors: loglinear model...
- Copula
- Regression

Modeling multivariate distribution

- Multivariate random vector: $(R_1, ..., R_J)$
- Continuous vectors: multivariate Gaussian, multivariate t-distribution...
- Categorical vectors: loglinear model...
- Copula
- Regression

Modeling multivariate distribution

- Multivariate random vector: $(R_1, ..., R_J)$
- Continuous vectors: multivariate Gaussian, multivariate t-distribution...
- Categorical vectors: loglinear model...
- Copula
- Regression

Modeling multivariate distribution

- Multivariate random vector: $(R_1, ..., R_J)$
- Continuous vectors: multivariate Gaussian, multivariate t-distribution...
- Categorical vectors: loglinear model...
- Copula
- Regression

Modeling multivariate distribution

- Multivariate random vector: $(R_1, ..., R_J)$
- Continuous vectors: multivariate Gaussian, multivariate t-distribution...
- Categorical vectors: loglinear model...
- Copula
- Regression

Latent variable modeling

- There exists α such that $f(R_1, ..., R_J|\alpha)$ is simple.
- What is considered as simple?
- Independence, small variance,...

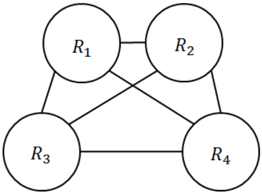
Latent variable modeling

- There exists α such that $f(R_1, ..., R_J|\alpha)$ is simple.
- What is considered as simple?
- Independence, small variance,...

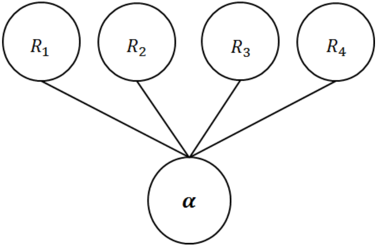
Latent variable modeling

- There exists α such that $f(R_1, ..., R_J|\alpha)$ is simple.
- What is considered as simple?
- Independence, small variance,...

Graphical representation



Local independence



$$f(R_1, ..., R_J|\boldsymbol{\alpha}) = \prod_j f(R_j|\boldsymbol{\alpha})$$

Applications

- Finance, political sciences
- Education
- Psychiatry/psychology
- Marketing and e-commerce

Applications

- Finance, political sciences
- Education
- Psychiatry/psychology
- Marketing and e-commerce

Applications

- Finance, political sciences
- Education
- Psychiatry/psychology
- Marketing and e-commerce

Applications

- Finance, political sciences
- Education
- Psychiatry/psychology
- Marketing and e-commerce

Linear factor models

- $(R_1, ..., R_J)$ is continous.
- Linear factor models: $\alpha = (\alpha_1, ..., \alpha_K)$

$$R_j = a_j^\top \alpha + \varepsilon_j$$

- Principle component analysis

Linear factor models

- $(R_1, ..., R_J)$ is continous.
- Linear factor models: $\alpha = (\alpha_1, ..., \alpha_K)$

$$R_j = a_j^\top \alpha + \varepsilon_j$$

- Principle component analysis

Linear factor models

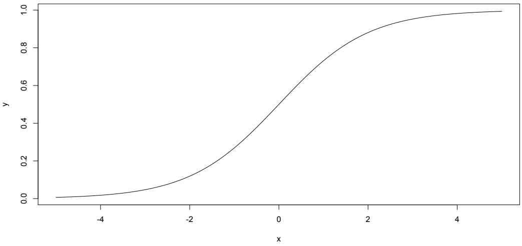
- $(R_1, ..., R_J)$ is continous.
- Linear factor models: $\alpha = (\alpha_1, ..., \alpha_K)$

$$R_j = a_j^\top \alpha + \varepsilon_j$$

- Principle component analysis

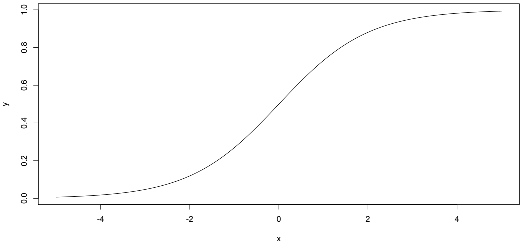
Categorical variable and item response theory model

- Binary $R_i \in \{0,1\}$.
- $P(R_j = 1|\alpha) = \frac{e^{a_j^\top \alpha - b_j}}{1+e^{a_j^\top \alpha - b_j}}, \quad \alpha \in \mathbb{R}^K$



Categorical variable and item response theory model

- Binary $R_i \in \{0,1\}$.
- $P(R_j = 1|\boldsymbol{\alpha}) = \frac{e^{a_j^\top \boldsymbol{\alpha} - b_j}}{1+e^{a_j^\top \boldsymbol{\alpha} - b_j}}, \quad \boldsymbol{\alpha} \in \mathbb{R}^K$



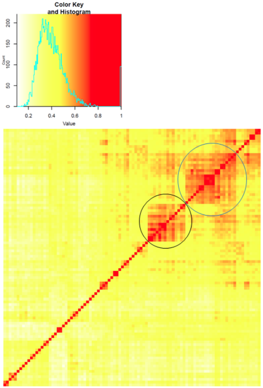
Stock Price Structure

- Data1: 97 stocks selected from S&P100 in 1013 trading days from 2009 to 2014.
- Data2: 117 stocks selected from SSE180 (Shanghai Stock Exchange) in 1159 trading days from 2009 to 2014.

Stock Price Structure

- Data1: 97 stocks selected from S&P100 in 1013 trading days from 2009 to 2014.
- Data2: 117 stocks selected from SSE180 (Shanghai Stock Exchange) in 1159 trading days from 2009 to 2014.

Exploratory Analysis



The heatmap of stock-stock correlation (Data 1; based on daily log return); stocks have been reordered

e.g.

- The block circled by blue contains mostly the energy companies:

  APA (Apache Corp), APC (Anadarko Petroleum), BHI (Baker Hughes), COP (Conoco Phillips), CVX (Chevron), DVN (Devon), ...

- The block circled by black contains the financial companies:

  C (citi), BAC (BOA), MS (Morgan Stanley), BK(Bank of New York Mellon), JPM (JP Morgan), ...

Linear factor model

- Linear factor models

$$R_j = a_j^\top \alpha + \varepsilon_j$$

- Fama-French model:

$$R = R_f + \beta(K - R_f) + b_s SMB + b_v HML + \alpha$$

Linear factor model

- Linear factor models

$$R_j = a_j^\top \alpha + \varepsilon_j$$

- Fama-French model:

$$R = R_f + \beta(K - R_f) + b_s SMB + b_v HML + \alpha$$

Linear factor model

- $(R_1, ..., R_J)$ is not multivariate Gaussian
  in many ways if J is large!
- Marginal tail, joint tail, asymmetric correlation...
- Too many factors!

Linear factor model

- $(R_1, ..., R_J)$ is not multivariate Gaussian

  in many ways if J is large!

- Marginal tail, joint tail, asymmetric correlation...

- Too many factors!

Linear factor model

- $(R_1, ..., R_J)$ is not multivariate Gaussian
  in many ways if J is large!
- Marginal tail, joint tail, asymmetric correlation...
- Too many factors!

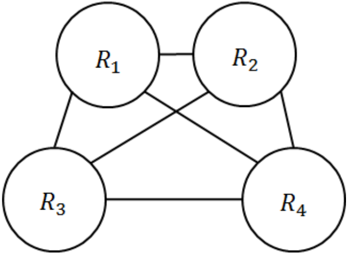
Nonlinear factor model

- Dichotomize $R_{ji} = 1$ if $S_i^{close} > S_i^{open}$ for stock j on day i
- $P(R_j = 1|\alpha) = \frac{e^{a_j^\top \alpha - b_j}}{1+e^{a_j^\top \alpha - b_j}}, \quad \alpha \in \mathbb{R}^K$

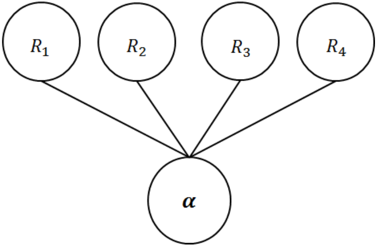
Nonlinear factor model

- Dichotomize $R_{ji} = 1$ if $S_i^{close} > S_i^{open}$ for stock j on day i
- $P(R_j = 1|\boldsymbol{\alpha}) = \frac{e^{a_j^\top \boldsymbol{\alpha} - b_j}}{1+e^{a_j^\top \boldsymbol{\alpha} - b_j}}, \quad \boldsymbol{\alpha} \in \mathbb{R}^K$

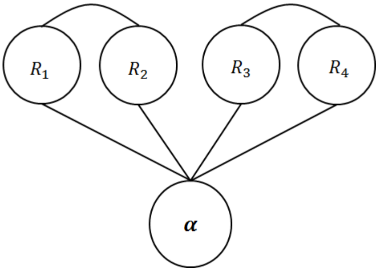
Latent graphical model



Latent graphical model



Latent graphical model



Issues to concern

- Parametric/nonparametric models: latent variable and graph
- Inference: identifiability

Issues to concern

- Parametric/nonparametric models: latent variable and graph
- Inference: identifiability

The latent variable component – IRT model

- Alternative formulation:

$$P(R_j = 1|\boldsymbol{\alpha}) = \frac{e^{a_j^\top \boldsymbol{\alpha} - b_j}}{1 + e^{a_j^\top \boldsymbol{\alpha} - b_j}} \qquad \Leftrightarrow \qquad P(R_j|\boldsymbol{\alpha}) \propto e^{R_j(a_j^\top \boldsymbol{\alpha} - b_j)}$$

- Local independence

$$P(R_1, ..., R_J|\boldsymbol{\alpha}) = \prod_{j=1}^{J} P(R_j|\boldsymbol{\alpha}) \propto e^{\sum_j R_j(a_j^\top \boldsymbol{\alpha} - b_j)}$$

The latent variable component – IRT model

- Alternative formulation:

$$P(R_j = 1|\boldsymbol{\alpha}) = \frac{e^{a_j^\top \boldsymbol{\alpha} - b_j}}{1 + e^{a_j^\top \boldsymbol{\alpha} - b_j}} \qquad \Leftrightarrow \qquad P(R_j|\boldsymbol{\alpha}) \propto e^{R_j(a_j^\top \boldsymbol{\alpha} - b_j)}$$

- Local independence

$$P(R_1, ..., R_J|\boldsymbol{\alpha}) = \prod_{j=1}^{J} P(R_j|\boldsymbol{\alpha}) \propto e^{\sum_j R_j(a_j^\top \boldsymbol{\alpha} - b_j)}$$

Graphical component component – Ising model

$$P(R_1, ..., R_J|S) \propto e^{\frac{1}{2}\sum_{i,j} s_{ij} R_i R_j}$$

- Physics
- Graphical representation

Graphical component component – Ising model

$$P(R_1, ..., R_J|S) \propto e^{\frac{1}{2}\sum_{i,j} s_{ij} R_i R_j}$$

- Physics
- Graphical representation

Latent graphical model: IRT model + Ising model

- Nonlocal independence

$$P(R_1, ..., R_J | \boldsymbol{\alpha}) \propto e^{\sum_j R_j (a_j^\top \boldsymbol{\alpha} - b_j) + \frac{1}{2} \sum_{i,j} s_{ij} R_i R_j}$$

- Simplification: $R_j^2 = R_j$

$$P(R_1, ..., R_J | \alpha) \propto e^{\sum_j R_j a_j^\top \alpha + \frac{1}{2} \sum_{i,j} s_{ij} R_i R_j}$$

Latent graphical model: IRT model + Ising model

- Nonlocal independence

$$P(R_1, ..., R_J|\boldsymbol{\alpha}) \propto e^{\sum_j R_j(a_j^\top \boldsymbol{\alpha} - b_j) + \frac{1}{2}\sum_{i,j} s_{ij} R_i R_j}$$

- Simplification: $R_j^2 = R_j$

$$P(R_1, ..., R_J|\boldsymbol{\alpha}) \propto e^{\sum_j R_j a_j^\top \boldsymbol{\alpha} + \frac{1}{2}\sum_{i,j} s_{ij} R_i R_j}$$

Latent variable and network modeling

- The item response function

$$f_{A,S}(\mathbf{R}|\boldsymbol{\alpha}) \propto \exp\{\boldsymbol{\alpha}^\top A\mathbf{R} + \frac{1}{2}\mathbf{R}^\top S\mathbf{R}\}$$

where $A_{K\times J} = (a_1, ..., a_J)$ and $S_{J\times J} = (s_{ij})$

- Population (prior) distribution such that

$$f_{A,S}(\mathbf{R}, \alpha) \propto \exp\{-|\alpha|^2/2 + \alpha^\top A\mathbf{R} + \frac{1}{2}\mathbf{R}^\top S\mathbf{R}\}$$

Latent variable and network modeling

- The item response function

$$f_{A,S}(\mathbf{R}|\boldsymbol{\alpha}) \propto \exp\{\boldsymbol{\alpha}^\top A\mathbf{R} + \frac{1}{2}\mathbf{R}^\top S\mathbf{R}\}$$

where $A_{K\times J} = (a_1, ..., a_J)$ and $S_{J\times J} = (s_{ij})$

- Population (prior) distribution such that

$$f_{A,S}(\mathbf{R}, \boldsymbol{\alpha}) \propto \exp\{-|\boldsymbol{\alpha}|^2/2 + \boldsymbol{\alpha}^\top A\mathbf{R} + \frac{1}{2}\mathbf{R}^\top S\mathbf{R}\}$$

Latent variable and network modeling

- Marginalized likelihood

$$\mathcal{L}(A, S) = \int f(\mathbf{R}, \boldsymbol{\alpha}) d\boldsymbol{\alpha} \propto \exp\{\frac{1}{2}\mathbf{R}^\top (A^\top A + S)\mathbf{R}\}$$

- Let $L_{J\times J} = A^\top A$

$$\mathcal{L}(L, S) = f(\mathbf{R}|L, S) \propto \exp\{\frac{1}{2}\mathbf{R}^\top (L + S)\mathbf{R}\}$$

Latent variable and network modeling

- Marginalized likelihood

$$\mathcal{L}(A,S) = \int f(\mathbf{R},\boldsymbol{\alpha})d\boldsymbol{\alpha} \propto \exp\{\frac{1}{2}\mathbf{R}^\top(A^\top A + S)\mathbf{R}\}$$

- Let $L_{J\times J} = A^\top A$

$$\mathcal{L}(L,S) = f(\mathbf{R}|L,S) \propto \exp\{\frac{1}{2}\mathbf{R}^\top(L + S)\mathbf{R}\}$$

Identifiability

- Identifiability of L and S
- Low dimension latent factor :

  $L_{J\times J} = A^\top A$ is positive semi-definite of rank $K \ll J$
- Small remaining dependence

  S is sparse

Identifiability

- Identifiability of L and S
- Low dimension latent factor :

  $L_{J\times J} = A^\top A$ is positive semi-definite of rank $K \ll J$
- Small remaining dependence

  S is sparse

Identifiability

- Identifiability of L and S
- Low dimension latent factor :

  $L_{J \times J} = A^\top A$ is positive semi-definite of rank $K \ll J$
- Small remaining dependence

  S is sparse

Regularized estimation

- Sparse network:

$$\|O(S)\|_1 = \sum_{i \neq j} |s_{ij}|$$

- Rank of $L = T^\top \Lambda T$ = number of nonzero eigenvalues

$$\|L\|_* \triangleq \sum_{i=1}^{J} |\lambda_i| = \sum_{i=1}^{J} \lambda_i = \mathrm{Trace}(L)$$

Regularized estimation

- Sparse network:

$$\|O(S)\|_1 = \sum_{i \neq j} |s_{ij}|$$

- Rank of $L = T^\top \Lambda T$ = number of nonzero eigenvalues

$$\|L\|_* \triangleq \sum_{i=1}^{J} |\lambda_i| = \sum_{i=1}^{J} \lambda_i = \mathrm{Trace}(L)$$

Regularized pseudo-likelihood estimation

$$\max_{L,S}\{\log \mathcal{L}(L,S) - \gamma\|O(S)\|_1 - \delta\|L\|_*\}$$

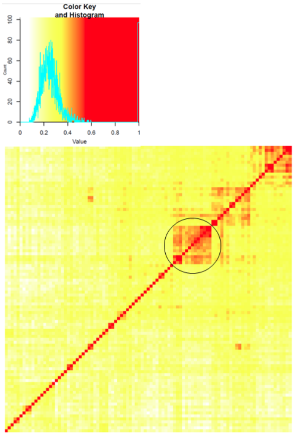
S&P 100

- $\hat{K} = 1$: a unidimensional market factor is preferred (general market movement).
- $Cor(\hat{\theta}, r_{s\&p}) = 0.77$.

S&P 100

- $\hat{K} = 1$: a unidimensional market factor is preferred (general market movement).
- $Cor(\hat{\theta}, r_{s\&p}) = 0.77$.

S&P 100


Color Key
and Histogram
Count
0
20
40
60
80
100
0
0.2
0.4
0.6
0.8
1
Value

S&P 100

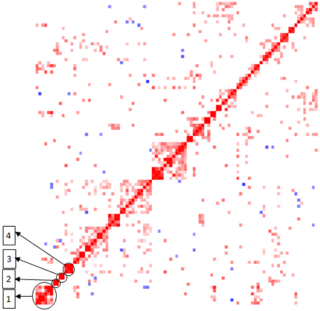


Figure: The estimated graph has 462 edges.

Structure Learning of S&P Stock Movement

Most positive stock-pairs:

Stock 1	Stock 2	$\hat{s}_{ij}$	cor
MA (Mastercard)	V (Visa)	1.90	0.53
NSC (Norfolk Southern)	UNP (Union Pacific)	1.90	0.60
HD (Home Depot)	LOW (Lowe's Cos)	1.81	0.51
AEP (American Electric Power)	SO (Southern)	1.54	0.48
FDX (FedEx)	UPS (United Parcel Service)	1.47	0.52
CVX (Chevron)	XOM (Exxon Mobil)	1.44	0.59
BAC (Bank of America)	C (Citigroup)	1.40	0.56
T (AT&T)	VZ (Verizon)	1.40	0.45
INTC (Intel)	TXN (Texas Instruments)	1.36	0.48
LMT (Lockheed Martin)	RTN (Raytheon)	1.32	0.49
BHI (Baker Hughes)	HAL (Halliburton)	1.31	0.58
HAL (Halliburton)	SLB (Schlumberger)	1.12	0.57
JPM (JP Morgan)	WFC (Wells Fargo)	1.09	0.57
KO (Coca-Cola)	PEP (Pepsi)	1.03	0.39
USB (US Bancorp)	WFC (Wells Fargo)	1.02	0.54
AXP (American Express)	COF (Capital One)	0.96	0.44

Structure Learning of S&P Stock Movement

Most negative stock-pairs:

Stock 1	Stock 2	$\hat{s}_{ij}$	cor
DOW (Dow Chemical)	LLY (Eli Lilly)	-0.73	0.13
AMZN (Amazon)	BAC (Bank of America)	-0.69	0.17
GILD (Gilead Sciences)	PEP (Pepsi)	-0.61	0.10
DVN (Devon Energy)	SBUX (Starbucks)	-0.60	0.20
GILD (Gilead Sciences)	SPG (Simon Property Group)	-0.55	0.10
CAT (Caterpillar)	PEP (Pepsi)	-0.53	0.13
F (Ford Motor)	PFE (Pfizer)	-0.53	0.17
LLY (Eli Lilly)	WFC (Wells Fargo)	-0.53	0.13

Structure Learning of S&P Stock Movement

Possible explanations: different reactions to the business cycle.

- Eli Lilly VS Dow Chemical: Health care stocks are viewed as more defensive investments compared with more cyclically sensitive stocks (such as Dow Chemical), since consumers are less likely to cut back on healthcare expenses during times of economic stress.
- BOA VS Amazon: Financial stocks tend to do well at the beginning of the business cycle, and may weaken as the economy heads into recession. However, Amazon benefited from the general shift to online commerce and the careful shopping behavior that consumers were exhibiting during the downturn (Bloomberg, 2009).

Structure Learning of S&P Stock Movement

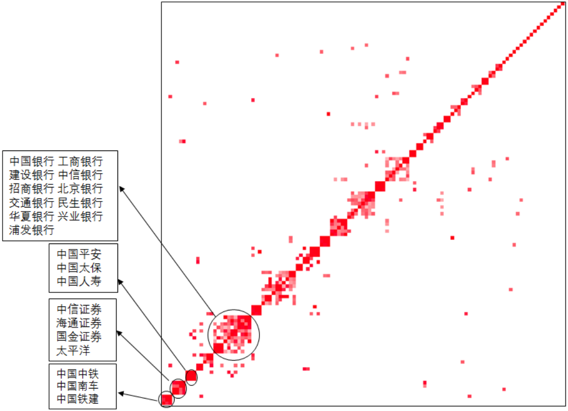
Maximal graph cliques (stock clusters): 10 6-stock cliques, 13 5-stock cliques, 64 4-stock cliques, and 94 3-stock cliques. e.g.

	Stocks	
1	HAL (Halliburton), BHI (Baker Hughes), NOV (National Oilwell Varco), SLB (Schlumberger), DVN (Devon), APC (Anadarko Petroleum)	Energy
2	BAC (Bank of America), JPM (JP Morgan), WFC (Wells Fargo), MS (Morgan Stanley), C (Citigroup)	Finance
3	BA (Boeing), GD (General Dynamics), LMT (Lockheed Martin), HON (Honeywell), RTN (Raytheon)	Industrials
4	EBAY (eBay), EMC (EMC), TXN (Texas Instruments), QCOM (QUALCOMM), AMZN (Amazon)	IT (although AMZN is listed in Consumer Discretionary)
5	CAT (Caterpillar), BHI (Baker Hughes), SLB (Schlumberger), NOV (National Oilwell Varco)	Industrial & energy

Structure Learning of Chinese Stock Movement

Results:

- $\hat{K} = 1$ and the estimated graph has 208 edges.



Structure Learning of Chinese Stock Movement

Most positive pairs:

Stock1	Stock2	s_ij	Cor	Stock1	Stock2	s_ij	Cor	Stock1	Stock2	s_ij	Cor
中国铁建	中国中铁	2.34	0.68	川投能源	国投电力	1.22	0.46	天津港	营口港	0.93	0.47
海通证券	中信证券	2.26	0.71	浦发银行	中信银行	1.19	0.64	阳泉煤业	中国神华	0.92	0.57
阳泉煤业	潞安环能	2.11	0.68	上汽集团	华域汽车	1.18	0.50	中信证券	太平洋	0.90	0.58
中金黄金	紫金矿业	2.11	0.60	民生银行	兴业银行	1.15	0.65	华能国际	国电电力	0.89	0.45
中国太保	中国人寿	1.77	0.65	北京银行	兴业银行	1.11	0.65	宝钢股份	包钢股份	0.89	0.43
建设银行	工商银行	1.69	0.61	华电国际	国电电力	1.09	0.48	浦发银行	华夏银行	0.89	0.66
北方稀土	厦门钨业	1.69	0.59	同仁堂	康美药业	1.07	0.48	北京银行	交通银行	0.88	0.59
中国石油	中国石化	1.66	0.57	外高桥	陆家嘴	1.06	0.54	中国卫星	中国船舶	0.88	0.48
浦发银行	兴业银行	1.57	0.71	北方导航	中国卫星	1.06	0.50	中国银行	建设银行	0.86	0.55
保利地产	金地集团	1.55	0.58	华发股份	上实发展	1.04	0.56	外高桥	浦东金桥	0.86	0.51
中国平安	中国人寿	1.49	0.62	江西铜业	金钼股份	1.03	0.61	华夏银行	民生银行	0.86	0.62
中国平安	中国太保	1.46	0.62	阳泉煤业	吉恩镍业	1.03	0.54	伊利股份	光明乳业	0.85	0.38
山西汾酒	贵州茅台	1.46	0.44	用友网络	东软集团	1.03	0.44	中国中铁	中国南车	0.85	0.50
中国银行	工商银行	1.43	0.58	首开股份	北京城建	0.98	0.54	招商银行	工商银行	0.85	0.53
吉恩镍业	金钼股份	1.40	0.59	陆家嘴	张江高科	0.97	0.54	中国神华	潞安环能	0.84	0.55
陆家嘴	浦东金桥	1.37	0.59	厦门钨业	金钼股份	0.96	0.52	中信证券	辽宁成大	0.83	0.52
华能国际	华电国际	1.35	0.48	保利地产	世茂股份	0.96	0.53	康美药业	天士力	0.81	0.42
航天电子	中国卫星	1.35	0.56	保利地产	冠城大通	0.96	0.53	北方导航	洪都航空	0.81	0.43
中海油服	海油工程	1.25	0.51	吉恩镍业	江西铜业	0.95	0.55	江西铜业	紫金矿业	0.81	0.50
国金证券	太平洋	1.25	0.56	保利地产	首开股份	0.93	0.56	招商银行	中国人寿	0.80	0.52

Structure Learning of Chinese Stock Movement

- There are stock pairs with moderate correlation but high residual association according to the graphical model (marked by yellow).
- Interesting patterns may be discovered: e.g., there seems to be nothing in common between "Zhongxin Zhengquan" and "Liaoning Chengda" (marked by green). However, "Zhongxin Zhengquan" kept to be one of the top ten shareholders of "Liaoning Chengda" during the period being considered.

Structure Learning of Chinese Stock Movement

Maximal graph cliques (stock clusters): 5 6-stock cliques, 6 5-stock cliques, 18 4-stock cliques, and 23-stock cliques. e.g.

华夏银行 浦发银行 兴业银行 交通银行 招商银行 北京银行

北方导航 航天电子 中国船舶 洪都航空 中国卫星

人福医药 天士力 康美药业 恒瑞医药 同仁堂

海通证券 中信证券 太平洋 国金证券

北京城建 保利地产 金地集团 首开股份

中国平安 中国太保 中国人寿

中国南车 中国铁建 中国中铁

潞安环能 阳泉煤业 中国神华

Senate voting

- The US senate voting record from 108th congress (2003) consists of 100 senators voting to 459 bills.
- Major factor: party membership
- Graph: personal relationship

Main factor

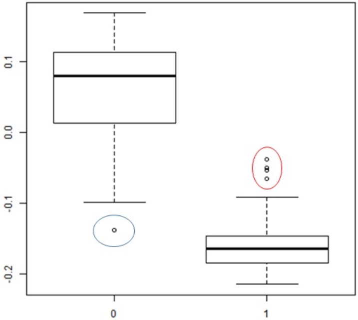
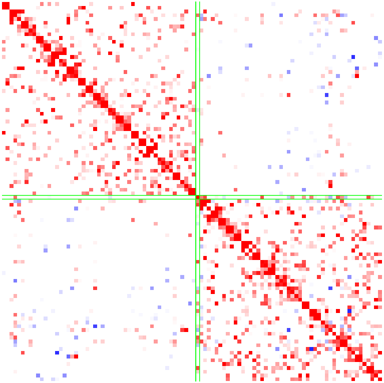


Figure: 0: Democratic, 1: Republican

The graph



Senate voting

- Positive links: same state or close relationship
- Example: Edwards.D.NC. & Kerry.D.MA.

  John Edwards became the 2004 Democratic candidate for vice president, the running mate of presidential nominee Senator John Kerry of Massachusetts.
- Example2: Baucus.D.MT. & Breaux.D.LA.

  Both of them supported the Bush's administration's effort to enact a Medicare drug benefit.

Senate voting

- Positive links: same state or close relationship
- Example: Edwards.D.NC. & Kerry.D.MA.

  John Edwards became the 2004 Democratic candidate for vice president, the running mate of presidential nominee Senator John Kerry of Massachusetts.
- Example2: Baucus.D.MT. & Breaux.D.LA.

  Both of them supported the Bush's administration's effort to enact a Medicare drug benefit.

Senate voting

- Positive links: same state or close relationship
- Example: Edwards.D.NC. & Kerry.D.MA.

  John Edwards became the 2004 Democratic candidate for vice president, the running mate of presidential nominee Senator John Kerry of Massachusetts.
- Example2: Baucus.D.MT. & Breaux.D.LA.

  Both of them supported the Bush's administration's effort to enact a Medicare drug benefit.

Senate voting

- Positive link while negative correlation.
- Example: Lieberman.D.CT. & Stevens.R.AK.

  Here is what Lieberman said about Stevens's death (2010). "... I have lost a dear friend. I am deeply saddened by Ted's death. I knew him for many years as a valued friend, a neighbor and a colleague. We shared many great experiences and I am grateful for all of the wisdom he offered me personally ?"

Summary

- Latent variable models
- Conditional graphical model