Statistical Foundations of Data Science

Jianqing Fan

Princeton University

https://fan.princeton.edu

ZOOM ID Lectures: **970 4936 8998** Office Hours: **996 4030 7631**

Annotated Lecture Notes: web view



10. Applications of Factor Models and PCA

10.1. Factor-Adjusted Regularized Model Selection(§11.1) *FarmSelect*

10.2. Factor-Adjusted Robust Multiple Testing (§11.2) *FarmTest*

10.3. Factor Augmented Regression Methods for Prediction(§11.3) *FarmPredict*

10.4. Community Detection (§11.4.1)

10.5. Topic Modeeling (§11.4.2)

10.6. Matrix completion(§11.4.3)

10.7. Item Ranking (§11.4.4)

10.1 Factor-adjusted Regularized Model Selection

FarmSelect

Factor Adjusted Model Selection

High-dim model: $Y_t = \alpha + \beta^T \mathbf{X_t} + \varepsilon_t$, **$\beta$ sparse**

★LASSO breaks down due to high-correlation of **X**

Factor-adjusted model selection: $\mathbf{X}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$

$$Y_t = \alpha + \underbrace{\beta^T \mathbf{B}\mathbf{f_t} + \beta^T \mathbf{u_t}}_{\gamma^T \mathbf{f_t} + \beta^T \mathbf{u_t}} + \varepsilon_t$$

★same β ★**u** weak-depend. ★reg. M-est work (Fan, Wang, Ke, 17)

FarmSelect: (factor-adjusted regularized model selection)

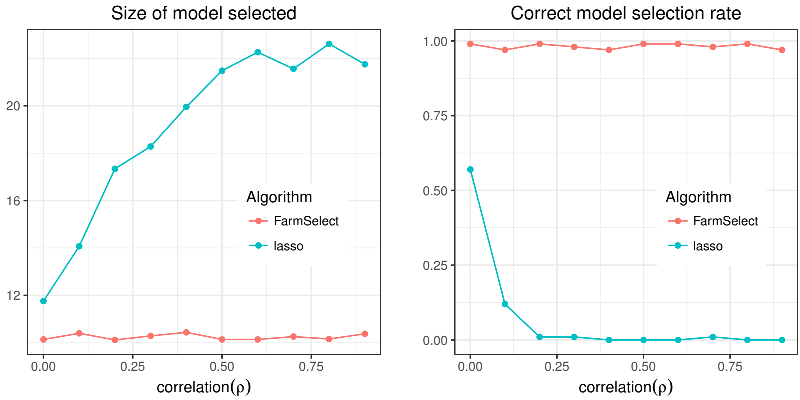
Robustly estimate $\mathbf{f}_t$ and **B**. **Use $\mathbf{u}_t$ and $\mathbf{f}_t$ as predictors**

Factor-adjusted versus direct approach

Model: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\beta = (\overbrace{3, \cdots, 3}^{10}, \overbrace{0, \cdots, 0}^{p-10})^T$, $\{\varepsilon_t\}_{i=1}^n \sim N(0, 0.3^2)$, and $\{\mathbf{X}_t\}_{i=1}^n \sim N_p(\mathbf{0}, \Sigma)$, w/ equi-corr Σ

one-factor: $X_j = \sqrt{\rho}\mathbf{f} + \sqrt{1-\rho}\varepsilon_j$

Parameters: $n = 100, p = 250, N_{sim} = 100.$



Factor-Adjusted Regularized Model Selection

★ Decompose $\mathbf{X}_i = \mathbf{B}\mathbf{f}_i + \mathbf{u}_i$ **u** weak-depend.

★ Fit $Y_i = f(\alpha + \mathbf{X}_i^T \beta, \varepsilon_i)$ via lifting: **New Predictors**: $\{(\mathbf{f}_i, \mathbf{u}_i)\}$

$$Y_i = f(\alpha + \underbrace{\beta^T \mathbf{B}\mathbf{f_i} + \beta^T \mathbf{u_i}}_{\gamma^T \mathbf{f_i} + \beta^T \mathbf{u_i}}, \varepsilon_i).$$

★ Apply to GLIM and M-estimation. R *FarmSelect*

By including $\mathbf{f}_i$, we decorrelate.

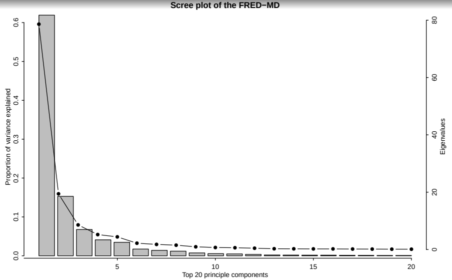
Simulation Results: Logistic regression with $s = 3$ and $n = 200$

	Factor model with $K = 3$					
	FarmSelect			LASSO		
	Selection rate	Screening	Ave size	Selection	Screening	Ave size
$p = 300$	0.94	1.00	3.07	0.07	0.98	12.61
$p = 400$	0.90	0.99	3.12	0.05	0.95	12.94
$p = 500$	0.83	0.98	3.18	0.03	0.93	15.07
	Equal correlated case ($\rho = 0.8$)					
	FarmSelect			LASSO		
	Selection	Screening	Ave size	Selection	Screening	Ave size
$p = 300$	0.93	1.00	3.09	0.07	0.85	9.90
$p = 400$	0.89	1.00	3.14	0.05	0.80	10.82
$p = 500$	0.85	0.99	3.19	0.02	0.69	11.79
	Uncorrelated case					
	FarmSelect			LASSO		
	Selection rate	Screening	Ave size	Selection	Screening	Ave size
$p = 300$	0.97	1.00	3.03	0.95	1.00	3.14
$p = 400$	0.93	1.00	3.07	0.91	1.00	3.34
$p = 500$	0.91	1.00	3.10	0.89	1.00	3.42

Prediction bond risk premia

Covariates: 131 disaggregated macroeconomic times series

■rolling window of 60 months to forecast Y_{t+1}.



Out of sample R^2 and average selected model size

Maturity of Bond	Out of sample R^2			Average model size	
	FarmSelect	LASSO	PCR	FarmSelect	Lasso
2 Years	0.2586	0.2295	0.2012	8.80	22.72
3 Years	0.2295	0.2166	0.1854	8.92	21.40
4 Years	0.2137	0.1801	0.1639	9.03	20.74
5 Years	0.2004	0.1723	0.1463	9.21	20.21

Applications to Neuroblastoma Data

■ Predict '3-y Event Free Survival", using gene expressions

■ Run the penalized logistic regression (56 "$+$" and 190 "$-$")

FarmSelect: 17 genes, **Lasso**: 40, **SCAD**: 34, **Elastic Net**: 86

Comparing bootstrap PE: Learning 200, testing 46

Bootstrap sample average	Model selection methods			
	FARMselect	Lasso	SCAD	elastic net
Model size	**17.6**	46.2	34.0	90.0
Correct prediction rate	**0.813**	0.807	0.809	0.790
Prediction performance with first 17 variables enter the solution path				
	FARMselect	Lasso	SCAD	elastic net
Correct prediction rate	**0.813**	0.733	0.764	0.705

10.2. Factor-adjusted robust multiple testing

FarmTest

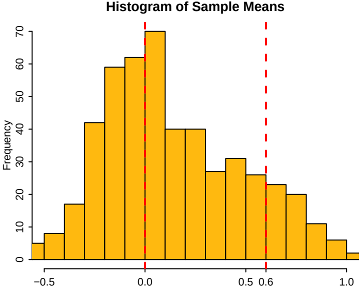
FarmTest

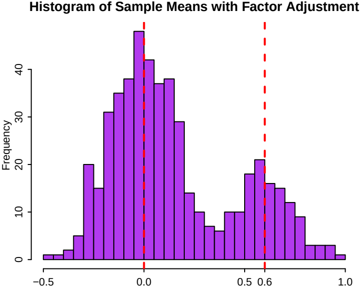
■ How many mutual funds have positive alpha? (Barras, et al, 10, JF)

A synthetic three-factor model: $\mathbf{X}_i = \mu + \underbrace{\mathbf{Bf}_i + \mathbf{u}_i}_{\varepsilon_i}$, $i = 1, \ldots, n$,

$$\mathbf{f}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3), \quad \mathbf{B} = (b_{j\ell}) \sim_{\text{IID}} \mathcal{U}(-1, 1) \;\&\; \mathbf{u_i} \sim \mathbf{t_3}(\mathbf{0}, \mathbf{I_p}).$$

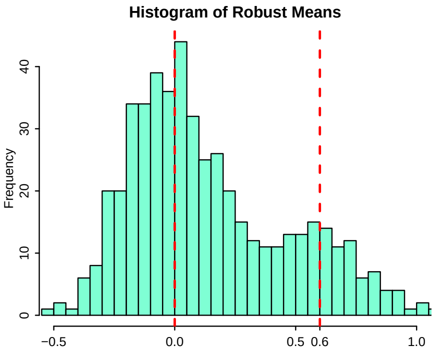
Model setup: $(n, p) = (100, 500)$, $\mu_j = 0.6$ **for** $j \leq p/4$; **0, otherwise**.

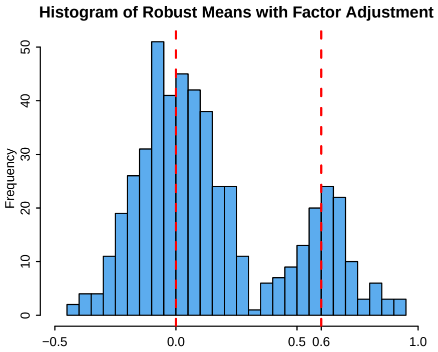




Naive : $\overline{\mathbf{X}_i}$ **FarmTest:** $\overline{\mathbf{X}_i - \mathbf{Bf}_i}$

Importance of Robust Adjustments





Decreased noise!

Factors Adjusted Robust Multiple testing

Factor adjusted data: $\widehat{Y}_{tj} \approx \mu_j + u_{tj}$, ★ $\widehat{Y}_{tj} = X_{tj} - \widehat{\mathbf{b}}_j^T \widehat{\mathbf{f}}_t$.

- Control false discovery rate as if independent normal data $\{Y_{tj}\}$.
- Proposed and studied by Fan, Ke, Sun, and Zhou (2020) R *FarmTest*

- ★ For each $H_{j0} : \mu_j = 0$, compute the robust two-sided test $\widehat{T}_j$
- ★ Compute the P-value $P_j = 2\Phi(-|\widehat{T}_j|)$ for testing H_{0j}.
- ★ Provide alternative ranking of P-values from those without adjust.
- ★ Reduce noise and increase power.

FDP approximation

Total discoveries: $\widehat{R}(z) = \sum_{j=1}^{p} 1(|\widehat{T}_j| \geq z)$ **z**: critical value

$$\mathrm{FDP}(z) = \underbrace{\frac{\sum_{j\in \text{Null}} 1(|\widehat{T}_j| \geq z)}{\widehat{R}(z)}}_{\textit{definition}, \textbf{ unknown}}, \qquad \widehat{\mathrm{FDP}}_N(z) = \underbrace{\frac{2p_0\Phi(-z)}{\widehat{R}(z)}}_{\textit{estimate}, \textbf{ known}}.$$

- $p_0 = \#\text{true null}$ ⟵ bounded by p **(conservative)**.
- **Usage**: Report sig. tests and FDP for given z ★Choose z to control FDP

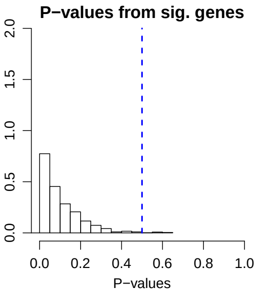
Valid approximation of FDP over a wide range

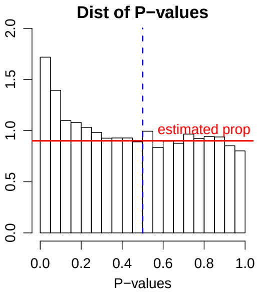
$$\max_{0\leq z\leq \Phi^{-1}(1-m_p/(2p))} \left| \frac{\widehat{\mathrm{FDP}}(z)}{\widehat{\mathrm{FDP}}_\mathrm{N}(z)} - 1 \right| \to 0 \quad \text{in probability.}$$

Remarks

1. True FDP can be well approximated by **normal** dist., after factor adj, as if data are **weakly** dep. with **reduced noise**.
2. Proportion of true nulls p_0/p can be estimated (Storey, 02):

$$\widehat{\pi}_0(\lambda) = \frac{1}{(1-\lambda)p}\sum_{j=1}^{p} 1(\widehat{P}_j > \lambda).$$





FDP and power comparisons: Models and Methods

Factor model: $\mathbf{X}_i = \mu + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i$, $n \in \{100, 150, 200\}, p = 500$.

$$\mathbf{B} = (b_{j\ell}) \sim_{\text{IID}} \mathcal{U}(-1,1), \ \ \mathbf{f}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3), \ \ \mathbf{u_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{4I_3}) \text{ or } \mathbf{t_3}(\mathbf{0}, \mathbf{I_3}),$$

$$\mu_j = 0.5, \ \ 1 \le j \le 25; \quad \mu_j = 0, \quad \text{otherwise}.$$

Competing methods:

1. FARM-H: FARM-Test with adaptive Huber covariance estimator;
2. FARM-U: FARM-Test with U-type covariance estimator;
3. FAM: A non-robust counterpart of FARM (sample mean + cov.);
4. PFA: Principal factor approximation *((Fan and Han, 17+))*;
5. Naive: Multiple t-tests ignoring factors.

FDP Control

Empirical mean abs. error between estimated & oracle FDP $(t = 0.01, z = 2.576)$

		$p = 500$				
$\mathbf{u}_i$	n	**FARM-H**	**FARM-U**	**FAM**	**PFA**	**Naive**
Normal	100	0.0601	0.0683	**0.0594**	0.0611	0.1902
	150	0.0559	0.0645	**0.0544**	0.0563	0.1582
	200	0.0525	0.0538	**0.0510**	0.0531	0.1348
t_3	100	**0.0799**	0.0848	0.1540	0.1796	0.3305
	150	**0.0723**	0.0712	0.1329	0.1510	0.2944
	200	**0.0643**	0.0619	0.1228	0.1366	0.2663

Non-robust methods break down!

Power Comparisons

Empirical power

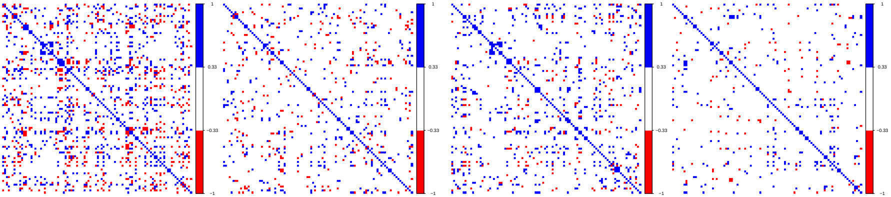
$\mathbf{u}_i$	n	FARM-H	FARM-U	FAM	PFA	Naive
		$p = 500$				
Normal	100	0.844	0.835	**0.881**	0.867	0.591
	150	0.868	0.861	**0.902**	0.893	0.629
	200	0.896	0.891	**0.927**	0.914	0.679
t_3	100	**0.882**	0.874	0.633	0.582	0.389
	150	**0.904**	0.889	0.675	0.607	0.417
	200	**0.914**	0.905	0.709	0.628	0.457

Little price to pay for robustness!

Applications to Neuroblastoma Data

■ **671 and 420** genes respect. have kurtosis heavier than t_5 ($p = 10,707$).

Before After Before After



Positive group before adjustment | Positive group after adjustment | Negative group before adjustment | Negative group after adjustment

■ corr $> 1/3$ ■ *corr* $< -1/3$.

■ At $t = 0.01$, FARM-U, FAM and naive methods identify

3855, **3509**, **3236** differently expressed genes.

■ Penalized logistic regression gives 56 "$+$" and 190 "$-$"

FarmSelect: 17 genes, **Lasso**: 40, **SCAD**: 34, **Elastic Net**: 86

10.3. Factor augmented regression methods for prediction

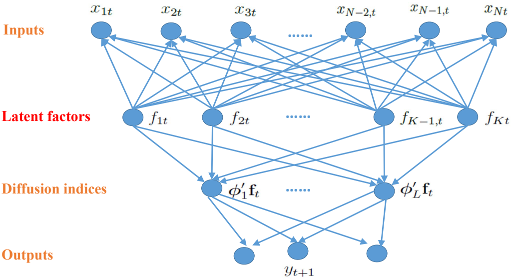
FarmPredict

Principal Component Regression

Data: $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, $\mathbf{X}_i$ is high-dim

★Regression Y on the principal components of $\mathbf{X}$

why?



Explanatory model: $\begin{cases} \mathbf{X}_i = \mathbf{a} + \mathbf{B}\mathbf{f}_i + \mathbf{u}_i, \\ Y_i = g(\mathbf{f}_i) + \varepsilon_i, \end{cases}$

★Provide very different motivation for classical one

Augmented Factor Models

Data: $\{(\mathbf{W}_i, \mathbf{X}_i, Y_i)\}_{i=1}^n$, $\mathbf{X}_i$ is high-dim

Model: $Y_i = g(\widehat{\mathbf{f}}_i, \mathbf{W}_i) + \varepsilon_i, \quad i \in [n].$ also called augmented PCR

■ $\mathbf{X}_i^* = (\mathbf{f}_i', \mathbf{W}_i')'$ = (**latent factors**, **augmented variables**)

- ★ **linear model**: $Y_i = \alpha + \beta^T \mathbf{X}_i^* + \varepsilon_i, \qquad i \in [n].$
- ★ **multi-index model**: $Y_i = g(\phi_1^T \mathbf{X}_i^*, \cdots, \phi_L^T \mathbf{X}_i^*) + \varepsilon_i$
- ★ **Machine Learning**: kernel, RandomForest, Deep Learning

Extension: $Y_i = g(\widehat{\mathbf{f}}_i, \widehat{\mathbf{u}}_i, \mathbf{W}_i) + \varepsilon_i.$ New data: $\{(\widehat{\mathbf{f}}_i, \widehat{\mathbf{u}}_i, \mathbf{W}_i, y_i)\}_{i=1}^n$

Forecast Bond Risk Premia

Linear prediction: Out-of-sample R^2 (%)

predictors	PPCA				PCA			
	Maturity(Year)				Maturity(Year)			
	2	3	4	5	2	3	4	5
$\mathbf{f}_t$	**38.0**	32.7	25.6	22.9	**23.0**	20.7	16.8	16.5
$(\mathbf{f}_t^T, \mathbf{W}_t^T)^T$	**37.7**	32.4	25.4	22.7	**23.9**	21.4	17.4	17.5

Multi-index prediction: Out-of-sample R^2 (%)

Predictors	PPCA				PCA			
	Maturity(Year)				Maturity(Year)			
	2	3	4	5	2	3	4	5
$\mathbf{f}_t$	**44.6**	43.0	38.8	37.3	**30.1**	25.5	23.2	21.3
$(\mathbf{f}_t^T, \mathbf{W}_t^T)^T$	**41.5**	38.7	35.2	33.8	**30.8**	26.3	24.6	22.0

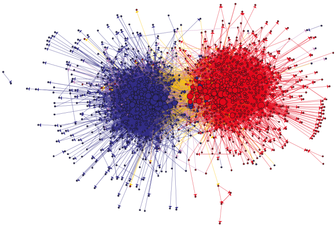
10.4. Community Detection

Community detection

Data: adjacency matrix $A \in \{0,1\}^{n\times n}$, indicating if nodes i and j has a link.

Stochastic Block Model: K disjoint communities $\mathcal{C}_1, \cdots, \mathcal{C}_K$, with $P(A_{kl} = 1) = p_{ij}$, for $k \in \mathcal{C}_i, l \in \mathcal{C}_j$, indep.

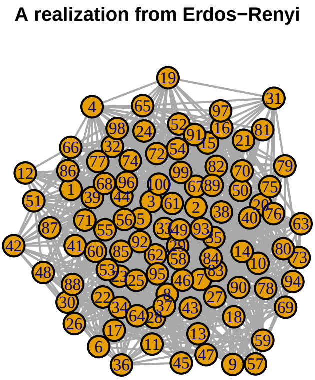
Edge probability: $\mathbf{P} = (p_{i,j})_{K\times K}$.



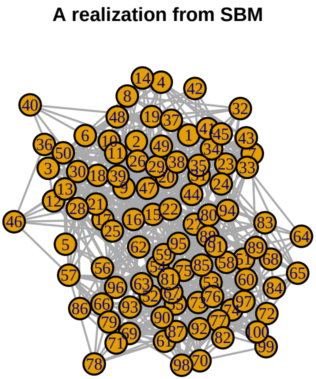
Erdös-Rényi graph: $p_{ij} = p$, **degenerate**

Planted partition model: $p_{ii} = p$ and $p_{ij} = q$ for $i \neq j$, $p \neq q$.

Simulated graphs



(a)



(b)

Simulated network data from (a) Erdös-Rényi graph and (b) SBM with $n = 100$, $p = 5\log(n)/n$, and $q = p/4$.

R-functions: `sample_gnp` and `sample_sbm` in R package `igraph`

Methods of Estimation

Likelihood: $\prod_{i>j} p_{C(i)C(j)}^{a_{ij}}(1-p_{C(i)C(j)})^{1-a_{ij}}$,

parameters: $\{C(i)\}_{i=1}^{K}$ and $\mathbf{P}=(p_{ij})\in R^{K\times K}$. ★hard to opt

Method of Moment: Let $\Gamma =$ membership matrix, with i^{th} row = membership of node i. Then (except diagnal elements, negligible)

$$\mathrm{E}\,\mathbf{A} = \underbrace{\Gamma}_{n\times K}\underbrace{\mathbf{P}}_{K\times K}\Gamma^T.$$

■ $\Gamma =$ eigen-space spanned by top K eigenvectors.

- ★ Get top K eigenvector matrix $\widehat{\Gamma}$ from $\mathbf{A}$
- ★ Run k-means algorithm on n (normalized) rows of $\widehat{\Gamma}$ to cluster

Example: Stochastic block model

Example: $K = 2$, $|J| = \frac{n}{2}$, E(A) = $\begin{pmatrix} p\mathbf{1}_{J,J} & q\mathbf{1}_{J,J^c} \\ q\mathbf{1}_{J^c,J} & p\mathbf{1}_{J^c,J^c} \end{pmatrix}$ has spectral

$$u_1^* = \frac{1}{\sqrt{n}}\mathbf{1},\ u_2^* = \frac{1}{\sqrt{n}}(\mathbf{1}_J - \mathbf{1}_{J^c}),\ \lambda_1^* = \frac{n(p+q)}{2},\ \lambda_2^* = \frac{n(p-q)}{2}.$$

- 2^{nd} eigenvector identifies memberships

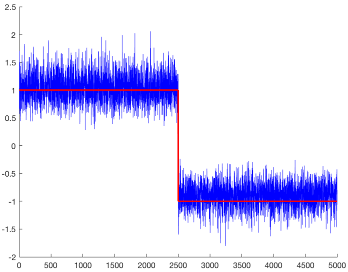
$n = 5000$, $p = \frac{4.5 \log n}{n}$, $q = \frac{\log n}{4n}$.

Red: entries of $\sqrt{n}u_2^*$.

Blue: entries of $\sqrt{n}u_2$.

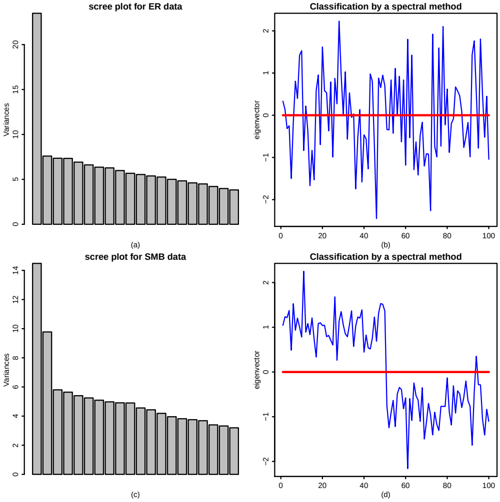
$\mathrm{sgn}(u_2)$ **recovers memberships,**

if uniformly approx.



Spectral analysis of network data

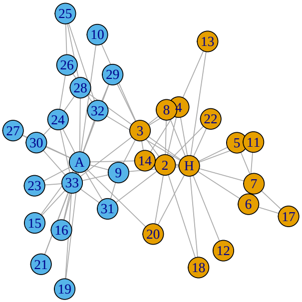
■ Simulated network data given before.



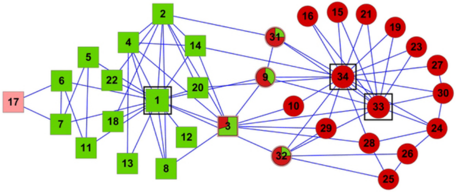
★Top panel: data generated from Erdös-Rényi model ★Bottom: data generated from SMB.

Mixed Memberships

- A university *karate club network* data for 34 members (Girvan & Newman, 2002)
- Edge links two members spent much time together outside club meetings
- At some point members split into *two communities* (led by *H* and by *A*)



(a) Non-overlapping



(b) Overlapping

Mixed Membership Model

Data: Adjacency matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{n \times n}$ follow

$$A_{ij} \sim_{indep} \text{Bernoulli}(h_{ij}), \text{for } i > j$$

Connection Probability: With degree heterogeneity θ_i,

$$P(A_{ij} = 1 | i \in \mathcal{C}_k, j \in \mathcal{C}_l) = \theta_i \theta_j p_{kl},$$

Edge probability: $\boldsymbol{\pi}_\mathbf{i} = (\pi_i(1), \cdots, \pi_i(K))^T \in \mathbb{R}^K$ is membership profile

$$\mathbf{P(A_{ij} = 1) = \theta_i \theta_j \sum_{k=1}^{K} \sum_{l=1}^{K} \pi_i(k) \pi_j(l) p_{kl} = h_{ij}}.$$

Spectral Clustering and Inference

Mixed Membership Model: With $\Pi = (\pi_1, \cdots, \pi_n)^T \in \mathbb{R}^{n \times K}$

$$\mathbf{EA} = \mathbf{H} = \Theta \Pi \mathbf{P} \Pi^T \Theta, \qquad \Theta = \mathrm{diag}(\theta_1, \cdots, \theta_n)$$

1. Compute top K eigenvector matrix $\widehat{\Gamma}$ from $\mathbf{A}$
2. Get SCORE $Y_{ik} = \widehat{\gamma}_{ik} / \widehat{\gamma}_{i1}, \qquad k = 2, \cdots, K$ (ratio eliminates Θ, Jin, 2015)
3. Run k-means algorithm on n rows of $\mathbf{Y}_i \in R^{K-1}$ to cluster

★ Applicable to both cases. No ratios is better in homogeneous case.

★ Fan et al (2021) gives uncertainty quantification.

10.5. Topic Modeling

Latent Semantic Indexing

Data: Text corpus $\mathbf{D} = (\mathbf{d}_1, \cdots, \mathbf{d}_n)$, $\quad d_{ij}$ = prop. of word i in doc j.

Multinomial: n_j = document length, $d_{m+1,j}$ = other words

$$n_j(d_{1,j}, \cdots, d_{m,j}) \sim \text{Multinomial}(n_j, d^*_{1,j}, \cdots, d^*_{m,j}).$$

Topic: doc j mixes K topics, with prob. $\pi_{j1}, \cdots, \pi_{jK}$. (Hofmann, 1999)

$$d^*_{ij} = \sum_{k=1}^{K} \pi_{jk} p_{ik}, \quad E\mathbf{D} = \underbrace{\mathbf{P}\Pi^T}_{\text{rank } \kappa}, \qquad \Pi = (\pi_{jk})_{n \times K}, \quad \mathbf{P} = (p_{ik})_{m \times K}$$

where p_{ik} = prob of word i in topic k.

Goal: learn Π and Θ and cluster them.

For $\pm$ news, π_{i1} measures sentiment; p_{jk} semantic.

Document Classification and Vertice Hunting

Anchor words: words with only one non-zero row.

1. Singular-value-decomposition: $\mathbf{D} = \mathbf{L}\Lambda\mathbf{R}$. $\mathbf{L}$ estimates $\mathbf{P}$ up to a right $K \times K$ matrix $\mathbf{U}$, identified by anchor words.
2. Use K-mean algorithm and rows $\mathbf{L}$ for grouping words and rows of $\mathbf{R}$ for clustering documents.
3. Use anchor words to identify estimates Π and $\mathbf{P}$ (*Ke and Wang, 19*).

Vertex Hunting: Anchor words correspond to vertices of eigen-vector $(\mathbf{L}_2/\mathbf{L}_1, \cdots, \mathbf{L}_K/\mathbf{L}_1)$. Use this to identify $\mathbf{U}$ and hence $\mathbf{P}$.

10.6. Matrix completion

The problem

Netflix problem: Customer i rates movie j if watched ; otherwise missing.

Similarly for books, music and CDs in **collaborative filtering**

Models: Let Θ be preference matrix. Observe $\mathbf{X}$ on a subset Ω, corrupted with noise: $X_{ij} = \theta_{ij} + \varepsilon_{ij}, \qquad$ **for** $(i,j) \in \Omega$**.**

Factor models: $\Theta = \underbrace{\mathbf{B}}_{n\times K} \underbrace{\mathbf{F}^T}_{K\times m}$ low rank

Missing at random: entry (i,j) is observed with prob p, indep.

Methods of Estimation

Penalized least-squares: min $\sum_{(i,j)\in\Omega}(X_{ij}-\theta_{ij})^2+\lambda\|\Theta\|_*$

Data matrix: Let $\mathbf{Y}=(X_{ij}I_{(i,j)\in\Omega}/p)$. Then, $E\,\mathbf{Y}=\Theta$.

★ p estimated by observed frequencies

Spectral method: Let $\mathbf{Y}=\sum_{i=1}^{\min(m,n)}\lambda_i\mathbf{u}_i\mathbf{v}_i^T$ be SVD. Set

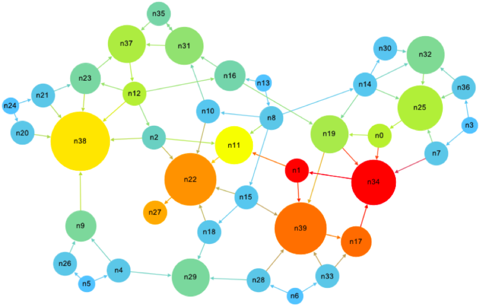
$$\widehat{\Theta}=\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{K}}\lambda_{\mathbf{i}}\mathbf{u}_{\mathbf{i}}\mathbf{v}_{\mathbf{i}}^{\mathbf{T}}.$$

10.7. Item Ranking

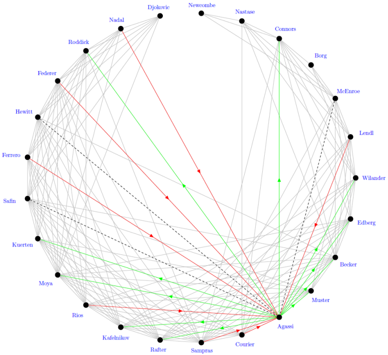
Ranking top K items based on many pairwise comparisons

Example: Top K Ranking

- web search, recomm. systems, admissions, sports, voting, ...



Page ranking (Dzenan Hamzic)

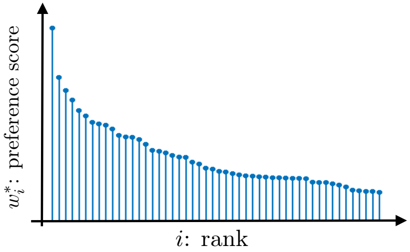


Ranking of tennis players (Bozóki, Csató)

Bradley-Terry-Luce model

Assign **latent score** to each of n items $\mathbf{w}^* = [w_1^*, \cdots, w_n^*]$

$$P\{\textbf{item } j \textbf{ beats item } i\} = \frac{w_j^*}{w_i^* + w_j^*}$$

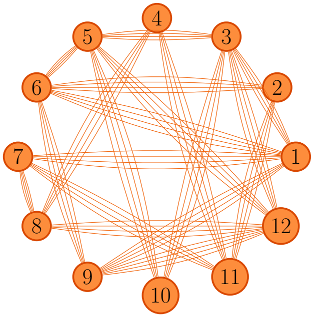


★ **Goal:** identify the set of **top-K items**

Sampling and likelihood based methods

★Comparison graph: Erdős–Rényi graph $\mathcal{G} \sim \mathcal{G}(n,p)$

★For each $(i,j) \in \mathcal{G}$, obtain L_{ij} paired comparisons



$$y_{i,j}^{(k)} \overset{\text{ind.}}{=} \begin{cases} 1, & \text{with prob. } \frac{w_j^*}{w_i^* + w_j^*} \\ 0, & \text{else} \end{cases} \qquad 1 \le k \le L_{ij}$$

Likelihood: $L(\mathbf{w}) = \prod_{(i,j)\in\mathcal{E}} \left(\frac{w_j}{w_i+w_j}\right)^{L_{ij}\widehat{p}_{ij}} \left(\frac{w_i}{w_i+w_j}\right)^{L_{ij}(1-\widehat{p}_{ij})}$ or its reparametrized $\mathbf{w} = \exp(\theta)$ **regularized version** $-\log L(\exp(\theta)) + \lambda\|\theta\|^2$.

A spectral method

Transition matrix: $P^*_{ij} = \begin{cases} \frac{1}{d} \cdot \frac{w_j^*}{w_i^* + w_j^*}, & \text{if } (i,j) \in \mathcal{G} \\ \textbf{remaining} & \text{if } i = j \\ 0, & \text{if } (i,j) \notin \mathcal{G} \end{cases}$, for given d.

Invariant distribution: $\pi^* \propto \mathbf{w}^*$, due to the reversibility:

$$\pi \mathbf{P}^* = \pi, \qquad \sum_{i=1}^{n} \pi_i P^*_{ij} = \sum_{i=1}^{n} \pi_j P^*_{ji} = \pi_j$$

Spectral ranking: based on $\mathbf{P} = \left(\frac{1}{d} \widehat{p}_{ij} I_{(i,j) \in \mathcal{G}}\right)$, ranked by its 1^{st} left-eigenvector.

- $d \geq d_{\max}$, maximum degree; e.g. $d = 2 * d_{\max}$