Fair Structure Learning in Graphical Models

Davoud Ataee Tarzanagh

University of Michigan

September 15, 2021

Overview

Why Graphical Models?

- A graphical model is a probabilistic model for which a graph expresses the conditional dependence structure between random variables.

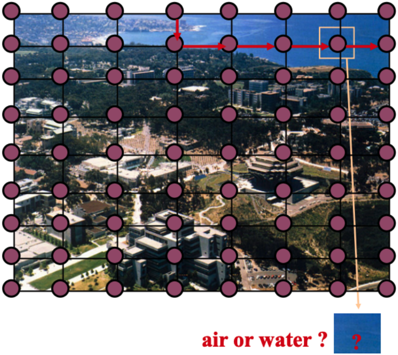
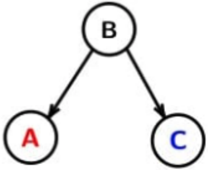


Figure: A Canonical Example: understanding complex scene

- Nodes correspond to random variables
- Edges represent statistical dependencies between the variables

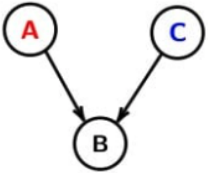
Conditional Independence



B: Train strike
A: Marina is late
C: Caroline is late

A and **C** independent?
No

A and **C** cond. independent given **B**?
Yes



B: Traffic jam
A: Rain
C: Football match

A and **C** independent?
Yes

A and **C** cond. independent given **B**?
No

Gaussian Graphical Models (GGM)

- A random vector $X \in \mathbb{R}^p$ is distributed according to the *multivariate Gaussian distribution* $\mathcal{N}(\mu, \Sigma)$ with parameters $\mu \in \mathbb{R}^p$ (the *mean*) and $\Sigma \in \mathcal{S}^p_{\succ 0}$ (the *covariance matrix*), if it has density function

$$f_{\mu,\Sigma}(x) = (2\pi)^{-p/2}(\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\},$$

where $x \in \mathbb{R}^p$.

Gaussian Graphical Models

- Let $G = (V, E)$ be an undirected graph with vertices $V = [p]$ and edges E, where $[p] = \{1, \ldots, p\}$.
- A random vector $X \in \mathbb{R}^p$ is said to *satisfy the (undirected) Gaussian graphical model with graph G*, if X has a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with

$$\left(\Sigma^{-1}\right)_{i,j} = 0 \quad \text{for all } (i,j) \notin E.$$

Gaussian Graphical Models

Theorem (Conditional Independence)

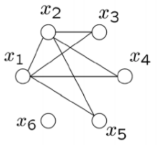
Let $X \in \mathbb{R}^p$ be distributed as $\mathcal{N}(\mu, \Sigma)$ and let $i, j \in [p]$ with $i \neq j$. Then

(a) $X_i \perp\!\!\!\perp X_j$ *if and only if* $\Sigma_{i,j} = 0$;

(b) $X_i \perp\!\!\!\perp X_j \mid X_{[p]\setminus\{i,j\}}$ *if and only if* $Q_{i,j} = \left(\Sigma^{-1}\right)_{i,j} = 0$.

$$Q_{ij} = 0 \Rightarrow X_i \perp\!\!\!\perp X_j \big| X_{[p]\setminus\{i,j\}} \tag{1}$$

$$\mathrm{Q} = \begin{pmatrix} * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & 0 & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}$$



Gaussian Graphical Models

Given n i.i.d. observations $X^{(1)}, \ldots, X^{(n)}$ from $\mathcal{N}(\mu, \Sigma)$, we define the *sample covariance matrix* as

$$S = \frac{1}{n} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T,$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}$ is the *sample mean*. We will see that $\bar{X}$ and S are sufficient statistics for the Gaussian model and hence we can write the log-likelihood function in terms of these quantities. Ignoring the normalizing constant, the Gaussian log-likelihood expressed as a function of (μ, Σ) is

$$\begin{aligned}
\ell(\mu, \Sigma) &\propto -\frac{n}{2} \log\det(\Sigma) - \frac{1}{2} \sum_{i=1}^{n} (X^{(i)} - \mu)^T \Sigma^{-1} (X^{(i)} - \mu) \\
&= -\frac{n}{2} \log\det(\Sigma) - \frac{n}{2} \mathrm{tr}(S\Sigma^{-1}) - \frac{n}{2} (\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu),
\end{aligned}$$

where $X^{(i)} - \mu = (X^{(i)} - \bar{X}) + (\bar{X} - \mu)$ and $\sum_{i=1}^{n} (X^{(i)} - \bar{X}) = 0$.

Gaussian Graphical Models

It can easily be seen that in the *saturated* (unconstrained) *model* where $(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}^p_{\succ 0}$, the MLE is given by

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\Sigma} = S,$$

assuming that $S \in \mathbb{S}^p_{\succ 0}$.

We will restrict ourselves to models where the mean μ is unconstrained, i.e. $(\mu, \Sigma) \in \mathbb{R}^p \times \Theta$, where $\Theta \subseteq \mathbb{S}^p_{\succ 0}$. In this case, $\hat{\mu} = \bar{X}$ and the ML estimation problem for Σ boils down to the optimization problem

$$\begin{aligned} \underset{\Sigma}{\text{maximize}} \quad & -\log\det(\Sigma) - \operatorname{tr}(S\Sigma^{-1}) \\ \text{subject to} \quad & \Sigma \in \Theta. \end{aligned} \tag{2}$$

Gaussian Graphical Models

- Gaussian graphical models are given by linear constraints on Q. So it is convenient to write the optimization problem (2) in terms of the concentration matrix Q:

$$\begin{aligned} \underset{Q}{\text{maximize}} \quad & \log\det(Q) - \operatorname{tr}(SQ) \\ \text{subject to} \quad & Q \in \mathcal{Q}, \end{aligned} \tag{3}$$

  where $\mathcal{Q} = \Theta^{-1}$.

- For a Gaussian graphical model with graph $G = (V, E)$ the constraints are given by $Q \in \mathcal{Q}_G$, where

$$\mathcal{Q}_G := \{Q \in \mathbb{S}^p_{\succ 0} \mid Q_{i,j} = 0 \text{ for all } i \neq j \text{ with } (i,j) \notin E\}.$$

- Since $\mathcal{Q}_G$ is a convex cone, this implies that ML estimation for Gaussian graphical models is a convex optimization problem.

Gaussian Graphical Models (Friedman, 2007)

- Sparse GGM learns a graph via the following optimization problem

$$\underset{\Theta \in \mathscr{M}}{\text{minimize}} \quad \text{trace}(S\Theta) - \log\det\Theta + \lambda \|\Theta\|_1,$$

where

- S is the empirical covariance matrix of X
- $\mathscr{M}$ is the set of $p \times p$ symmetric positive definite matrices
- λ is a non-negative tuning parameter.

- Why do we need sparse solution?
    - feature/variable selection
    - better interprete the data
    - shrinkage the size of model
    - computatioal savings
    - discourage overfitting

Ising Graphical Model (Ising, 1925; Lee, 2007)

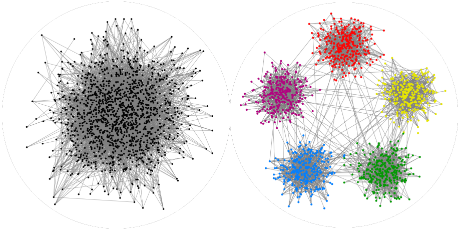
- Ising Graphical Model (IGM) is suitable for binary or categorical data
- Let $\mathbf{y} = (y_1, \dots, y_p) \in \{0,1\}^p$ denote a binary random vector. The Ising model specifies the probability mass function

$$p(\mathbf{y}) = \frac{1}{\mathscr{W}(\Theta)} \exp\Big(\sum_{j=1}^{p} \theta_{jj} y_j + \sum_{1 \le j < j' \le p} \theta_{jj'} y_j y_{j'} \Big). \tag{4}$$

  Here, $\mathscr{W}(\Theta)$ is the partition function.
- Sparse IGM learns a graph via the following optimization problem

$$\min_{\Theta \in \mathscr{M}} \sum_{j=1}^{p} \sum_{j'=1}^{p} \theta_{jj'} s_{jj'} - \sum_{i=1}^{n} \sum_{j=1}^{p} \log\big(1 + \exp(\theta_{jj} + \sum_{j' \neq j} \theta_{jj'} y_{ij'})\big) + \lambda \|\Theta\|_1.$$

Community Detection



- The above two graphs are the same graph re-organized and drawn from the SBM model with 1000 vertices, 5 balanced communities, within-cluster probability of $1/50$ and across-cluster probability of $1/1000$.
- The goal of community detection in this case is to obtain the right graph (with the true communities) from the left graph (scrambled) up to some level of accuracy.

Cluster-Based Graphical Models

- Suppose there exist K disjoint communities of nodes denoted by $\mathscr{V} = \mathscr{C}_1 \cup \cdots \cup \mathscr{C}_K$ where $\mathscr{C}_k$ is the subset of nodes from $\mathscr{G}$ that belong to the k–th community.
- For each candidate partition of n nodes into K communities, we associate it with a *partition matrix* $\mathbf{Q} \in \{0,1\}^{p \times p}$, such that $q_{ij} = 1/|C_k|$ if and only if nodes i and j are assigned to the kth community.
- Let $\mathscr{Q}_{pK}$ be the set of all such partition matrices, and $\bar{\mathbf{Q}}$ the true partition matrix associated with the ground-truth clusters $\{\bar{\mathscr{C}_k}\}_{k=1}^K$.

Demographic Balance Clusters (Chierichetti et. al 2018)

- $\mathscr{V}$ contains H demographic groups such that $\mathscr{V} = \mathscr{D}_1 \cup \cdots \cup \mathscr{D}_H$.
- Demographic Balance Clusters: representation in each cluster to preserve the global fraction of each demographic group $\mathscr{D}_h$, i.e.,

$$\frac{|\mathscr{D}_h \cap \mathscr{C}_k|}{|\mathscr{C}_k|} = \frac{|\mathscr{D}_h|}{p} \quad \text{for all} \quad k \in [K].$$

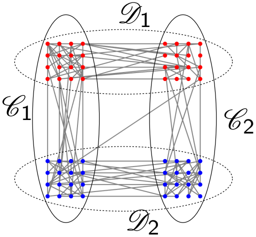


Figure: Fair Clustering. There are two meaningful ground-truth clusterings into two clusters: $\mathscr{V} = \mathscr{C}_1 \cup \mathscr{C}_2$ and $\mathscr{V} = \mathscr{D}_1 \cup \mathscr{D}_2$. Only the first one is fair.

Example: High School Friendship Networks

- Vertices correspond to students and are split into two groups of males and females.
- Eedge between two students indicates that one of them reported friendship with the other one.

- Gender should be balanced

Fair Graphical Models

- Let $\mathbf{R} \in \{0,1\}^{p \times p}$ be such that $r_{ij} = 1$ if and only if nodes i and j are assigned to the same group, with the convention that $r_{ii} = 1, \forall i$.
-
$$\underbrace{\mathbf{R}(\mathbf{I} - \mathbf{1}\mathbf{1}^\top/p)}_{=:\mathbf{A}_1}\mathbf{Q} = 0 \Leftrightarrow \frac{|\mathscr{D}_h \cap \mathscr{C}_k|}{|\mathscr{C}_k|} = \frac{|\mathscr{D}_h|}{p}.$$
- Let $\mathbf{B}_1 := \mathrm{diag}(\varepsilon)\mathbf{J}_p$ for some $\varepsilon > 0$ that controls how close we are to exact demographic parity.
- Fairness Constraint:
$$\mathbf{A}_1\mathbf{Q} \leq \mathbf{B}_1.$$

Fair Graphical Models

Fair Structured Graph Learning:

$$\begin{array}{ll} \underset{\Theta,\ Q}{\text{minimize}} & L(\Theta;\mathbf{Y}) + \rho_1 \|\Theta\|_{1,\text{off}} + \rho_2 \operatorname{trace}\big((\mathbf{S}+\mathbf{Q})G(\Theta)\big) \\ \text{subj. to} & \Theta \in \mathscr{M}, \quad \mathbf{A}_1\mathbf{Q} \leq \mathbf{B}_1, \quad \text{and} \quad \mathbf{Q} \in \cup_K \mathscr{Q}_{pK}. \end{array} \tag{5}$$

Here, $G(\Theta) : \mathscr{M} \to \mathscr{M}$ is a function of Θ.

- Fair GLasso: $L(\Theta;\mathbf{Y}) = -\log\det(\Theta) + \operatorname{trace}(\mathbf{S}\Theta)$ and $G(\Theta) = \Theta$.

Fair Graphical Models

$\mathbf{Q}$ satisfies several convex constraints:

- all entries of $\mathbf{Q}$ are nonnegative,
- all diagonal entries of $\mathbf{Q}$ are 1,
- $\mathbf{Q}$ is positive semi-definite.

(Bi-) Convex Relaxation:

$$\begin{aligned} \underset{\Theta,\ Q}{\text{minimize}} \quad & L(\Theta;\mathbf{Y}) + \rho_1 \|\Theta\|_{1,\text{off}} + \rho_2 \,\text{trace}\left((\mathbf{S}+\mathbf{Q})G(\Theta)\right) \\ \text{subj. to} \quad & \Theta \in \mathscr{M}, \quad \text{and} \quad \mathbf{Q} \in \mathscr{N}. \end{aligned} \tag{6a}$$

Here,

$$\begin{aligned} \mathscr{M} &= \left\{\Theta \in \mathbb{R}^{p\times p}: \ \theta_{ij} = \theta_{ji}, \ \text{and} \ \theta_{ii} > 0, \ \text{for every} \ 1 \le i,j \le p\right\}, \\ \mathscr{N} &= \left\{\mathbf{Q} \in \mathbb{R}^{p\times p}: \ \mathbf{Q} \succeq \mathbf{0}, \ \mathbf{0} \le \mathbf{A}\mathbf{Q} \le \mathbf{B}\right\}, \\ \mathbf{A} &= [\mathbf{A}_1; \mathbf{I}_p], \quad \text{and} \quad \mathbf{B} = [\mathbf{B}_1; \mathbf{J}_p]. \end{aligned} \tag{6b}$$

Alternating Direction Method of Multipliers (ADMM)

- Let $\Omega = (\boldsymbol{\Theta}, \mathbf{Q})$, $\dot{\Omega} = (\dot{\boldsymbol{\Theta}}, \dot{\mathbf{Q}})$.
- The scaled augmented Lagrangian takes the form

$$\begin{aligned} \Upsilon_\gamma(\Omega, \dot{\Omega}, \mathbf{W}) := L(\Theta; \mathbf{Y}) + \rho_1 \|\dot{\Theta}\|_1 + \rho_2 \operatorname{trace}\left((\mathbf{S} + \mathbf{Q}) G(\Theta)\right) \\ + \imath(\mathbf{Q} \succeq \mathbf{0}) + \imath(\mathbf{0} \leq \mathbf{A}\dot{\mathbf{Q}} \leq \mathbf{B}) + \frac{\gamma}{2} \|\Omega - \dot{\Omega} + \mathbf{W}\|_F^2. \end{aligned} \tag{7}$$

Here,

- $\Theta \in \mathscr{M}$
- Ω and $\dot{\Omega}$ are the primal variables
- $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2)$ is the dual variable
- $\gamma > 0$ is a dual parameter
- $\imath(\cdot)$ denote the indicator function.

Alternating Direction Method of Multipliers (ADMM)

The proposed ADMM algorithm requires the following updates:

$$\Omega^{(t+1)} \leftarrow \underset{\Omega}{\operatorname{argmin}}\ \Upsilon_\gamma(\Omega, \dot{\Omega}^{(t)}, \mathbf{W}^{(t)}), \tag{8a}$$

$$\dot{\Omega}^{(t+1)} \leftarrow \underset{\dot{\Omega}}{\operatorname{argmin}}\ \Upsilon_\gamma(\Omega^{(t+1)}, \dot{\Omega}, \mathbf{W}^{(t)}), \tag{8b}$$

$$\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} + \Omega^{(t+1)} - \dot{\Omega}^{(t+1)}. \tag{8c}$$

where $\Omega = (\boldsymbol{\Theta}, \mathbf{Q})$, $\dot{\Omega} = (\dot{\boldsymbol{\Theta}}, \dot{\mathbf{Q}})$.

Algorithm 1 Fair Graph Learning via ADMM

Initialize the parameters: (a) primal variables $\Theta, \mathbf{Q}, \dot{\Theta}$ and $\dot{\mathbf{Q}}$ to the $p \times p$ identity matrix; (b) dual variables $\mathbf{W}_1$ and $\mathbf{W}_2$ to the $p \times p$ zero matrix; (c) constants $\gamma > 0$ and $\nu > 0$.

Iterate until the stopping criterion $\max\left\{ \frac{\|\boldsymbol{\Theta}^{(t)} - \boldsymbol{\Theta}^{(t-1)}\|_F^2}{\|\boldsymbol{\Theta}^{(t-1)}\|_F^2}, \frac{\|\mathbf{Q}^{(t)} - \mathbf{Q}^{(t-1)}\|_F^2}{\|\mathbf{Q}^{(t-1)}\|_F^2} \right\} \leq \nu$ is met, where $\boldsymbol{\Theta}^{(t)}$ and $\mathbf{Q}^{(t)}$ are the value of $\boldsymbol{\Theta}$ and $\mathbf{Q}$, respectively, obtained at the t-th iteration:

1. Update graph adjacency matrices Θ and $\dot{\Theta}$:
   1. $\boldsymbol{\Theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\Theta} \in \mathscr{M}}{\arg\min} \left\{ L(\Theta; \mathbf{Y}) + \rho_2 \operatorname{trace}\left((\mathbf{S} + \mathbf{Q}^{(t)}) G(\Theta) \right) + \frac{\gamma}{2} \|\boldsymbol{\Theta} - \dot{\boldsymbol{\Theta}}^{(t)} + \mathbf{W}_1^{(t)}\|_F^2 \right\}.$
   2. $\dot{\Theta}^{(t+1)} \leftarrow \textsc{Shrink}\left(\boldsymbol{\Theta}^{(t+1)} + \mathbf{W}_1^{(t)}, \rho_1/\gamma \right)$, where $\textsc{Shrink}(a_{ij}, b) = \operatorname{sign}(a_{ij}) \max(|a_{ij}| - b, 0).$
2. Update partition matrices $\mathbf{Q}$ and $\dot{\mathbf{Q}}$:
   1. $\mathbf{Q}^{(t+1)} \leftarrow \left(\dot{\mathbf{Q}}^{(t)} - \mathbf{W}_2^{(t)} - \frac{\rho_2}{\gamma} G(\Theta^{t+1}) \right)_+.$
   2. $\dot{\mathbf{Q}}^{t+1} \leftarrow \textsc{Proj}_{\dot{\mathscr{N}}}\left(\mathbf{Q}^{(t+1)} + \mathbf{W}_2^{(t+1)} \right)$, where $\dot{\mathscr{N}} = \left\{ \dot{\mathbf{Q}} \in \mathbb{R}^{p \times p} : \mathbf{0} \leq \mathbf{A}\dot{\mathbf{Q}} \leq \mathbf{B} \right\}.$
3. Update dual variables $\mathbf{W}_1$ and $\mathbf{W}_2$:
   1. $\mathbf{W}_1^{(t+1)} = \mathbf{W}_1^{(t)} + \boldsymbol{\Theta}^{(t+1)} - \dot{\Theta}^{(t+1)}$
   2. $\mathbf{W}_2^{(t+1)} = \mathbf{W}_2^{(t)} + \mathbf{Q}^{(t+1)} - \dot{\mathbf{Q}}^{(t+1)}.$

Convergence and Computational Complexity of ADMM

Theorem

The iterates generated by Algorithm 1 converge to a stationary point of the augmented Lagrangian (7).

- Computational Complexity:
  - Unknown K: $O(p^3)$.
  - Known K and $\varepsilon = 0$: $\max\left(\min\left(O(np^2), O(p^3)\right), (p - H + 1)^2 K\right)$.

Fair CONCORD (FCONCORD)

Letting $G(\Theta) = \Theta^2$ in (6), our problem takes the form

$$\begin{aligned}
\underset{\Theta,\ Q}{\text{minimize}} \quad & F(\Theta, \mathbf{Q}; \mathbf{Y}) := \frac{n}{2}\Big[-\log |\mathrm{diag}(\Theta)^2| \\
& \qquad + \mathrm{trace}\left(((1+\rho_2)\mathbf{S} + \rho_2 \mathbf{Q})\Theta^2\right)\Big] + \rho_1 \|\Theta\|_{1,\mathrm{off}} \\
\text{subj. to} \quad & \Theta \in \mathscr{M} \quad \text{and} \quad \mathbf{Q} \in \mathscr{N}.
\end{aligned} \tag{9}$$

Here, $\mathscr{M}$ and $\mathscr{N}$ are the graph adjacency and fairness constraints, respectively.

Large Sample Properties of FCONCORD

Let

- $\theta^o = (\theta_{ij})_{1 \le i < j \le p}$ and $\mathbf{q}^o = (q_{ij})_{1 \le i < j \le p}$ denote the vector of off-diagonal entries of Θ and $\mathbf{Q}$, respectively.
- θ^d and $\mathbf{q}^d$ denote the vector of diagonal entries of Θ and $\mathbf{Q}$, respectively.
- $\bar{\theta}^o, \bar{\theta}^d, \bar{\mathbf{q}}^o$, and $\bar{\mathbf{q}}^d$ denote the true value of $\theta^o, \theta^d, \mathbf{q}^o$, and $\mathbf{q}^d$, respectively.
- $\mathscr{B}$ denote the set of non-zero entries in the vector $\bar{\theta}^o$
- $F_n(\theta^d, \theta^o, \mathbf{q}^d, \mathbf{q}^o; \mathbf{Y})$ stands for for $\frac{F}{n}$ in (9).

Restricted version of criterion (9):

$$\underset{\theta^o, q^o}{\text{minimize}} \quad F_n(\bar{\theta}^d, \theta^o, \bar{\mathbf{q}}^d, \mathbf{q}^o; \mathbf{Y}) + \rho_{1n} \|\theta^o\|_1, \quad \text{subj. to} \quad \theta^o_{\mathscr{B}^c} = 0. \tag{10}$$

Large Sample Properties of FCONCORD

The following standard assumptions are required.

1. The random vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are *i.i.d.* sub-Gaussian for every $n \geq 1$, i.e., there exists $M > 0$ such that $\|\mathbf{u}^\top \mathbf{y}_i\|_{\psi_2} \leq M\sqrt{\mathbb{E}(\mathbf{u}^\top \mathbf{y}_i)^2}$, $\forall \mathbf{u} \in \mathbb{R}^p$. Here, $\|\mathbf{y}\|_{\psi_2} = \sup_{t\geq 1}(\mathbb{E}|\mathbf{y}|^t)^{\frac{1}{t}}/\sqrt{t}$.
2. There exist constants $\tau_1, \tau_2 \in (0,\infty)$ such that
$$\tau_1 < \Lambda_{\min}(\bar{\Theta}) \leq \Lambda_{\max}(\bar{\Theta}) < \tau_2.$$
3. There exists a constant $\tau_3 \in (0,\infty)$ such that
$$0 \leq \Lambda_{\min}(\bar{\mathbf{Q}}) \leq \Lambda_{\max}(\bar{\mathbf{Q}}) < \tau_3.$$
4. For any $K, H \in [p]$, we have $K \leq p - H + 1$.
5. There exists a constant $\delta < 1$ such that for all $(i,j) \in \mathscr{B}$,
$$\left| \bar{L}''_{ij,\mathscr{B}}(\bar{\theta}^d, \bar{\theta}^o) \left(\bar{L}''_{\mathscr{B},\mathscr{B}}(\bar{\theta}^d, \bar{\theta}^o) \right)^{-1} \operatorname{sign}(\bar{\theta}^o_{\mathscr{B}}) \right| \leq \delta,$$
where for $1 \leq i,j,t,s \leq p$ satisfying $i < j$ and $t < s$,
$$\bar{L}''_{ij,kl}(\bar{\theta}^d, \bar{\theta}^o) := \mathbb{E}_{\bar{\theta}^d, \bar{\theta}^o}\left(\frac{\partial^2 L(\theta^d, \theta^o; \mathbf{Y})}{\partial \theta_{ij} \partial \theta_{kl}} \Big|_{\theta^d = \bar{\theta}^d, \theta^o = \bar{\theta}^o} \right).$$

Large Sample Properties of FCONCORD

Theorem (Restricted version)

Suppose Assumptions 1–4 are satisfied. Assume further that $\rho_{1n} = O(\sqrt{\frac{\log p}{n}})$, $n = O(|\mathscr{B}|\log(p))$, $\frac{p-H+1}{n}(\frac{p-H+1}{K}-1) = o(1)$, *and* $\varepsilon = 0$. *Then, there exists a finite constant* $\Pi_1(\bar{\theta}^o, \bar{\mathbf{q}}^o)$, *such that for any* $\eta > 0$, *the following events hold with probability at least* $1 - O(\exp(-\eta \log p))$:

- *there exists a local minimizer* $(\widehat{\theta}^o_{\mathscr{B}}, \widehat{\mathbf{q}}^o)$ *of* (10);
- *any local minimizer* $(\widehat{\theta}^o_{\mathscr{B}}, \widehat{\mathbf{q}}^o)$ *of the problem* (10) *satisfies*

$$\|\widehat{\theta}^o_{\mathscr{B}} - \bar{\theta}^o_{\mathscr{B}}\| + \|\widehat{\mathbf{q}}^o - \bar{\mathbf{q}}^o\| \leq \Pi_1(\bar{\theta}^o, \bar{\mathbf{q}}^o)\left(\rho_{1n}\sqrt{|\mathscr{B}|} + \rho_{2n}\sqrt{\frac{(p-H+1)^2}{nK} - \frac{p-H+1}{n}}\right);$$

- *If* $\min_{(i,j)\in\mathscr{B}} \bar{\theta}_{ij} \geq \sqrt{q}(\rho_{1n} + \tau_2\rho_{2n})$, *then* $\widehat{\theta}^o_{\mathscr{B}^c_k} = 0$ *for the k-th community.*

- Main result: If $n \to \infty$,

$$\widehat{\theta}^o_{\mathscr{B}} \to \bar{\theta}^o_{\mathscr{B}}, \quad \text{and} \quad \widehat{\mathbf{q}}^o \to \bar{\mathbf{q}}^o.$$

Large Sample Properties of FCONCORD

Theorem

Assume the conditions of Theorem 2 and Assumption 5 hold. Then, there exists a constant $\Pi_1(\bar{\theta}^o, \bar{\mathbf{q}}^o)$ *such that for any* $\eta > 0$*, the following events hold with probability at least* $1 - O(\exp(-\eta \log p))$*:*

- *There exists a minimizer* $(\widehat{\theta}^o, \widehat{\mathbf{q}})$ *of* (10).
- *Any minimizer* $(\widehat{\theta}^o, \widehat{\mathbf{q}})$ *of* (10) *satisfies*

$$\|\widehat{\theta}^o - \bar{\theta}^o\| + \|\widehat{\mathbf{q}}^o - \bar{\mathbf{q}}^o\| \leq \Pi_1(\bar{\theta}^o, \bar{\mathbf{q}}^o)\left(\rho_{1n}\sqrt{|\mathscr{B}|} + \rho_{2n}\sqrt{\frac{(p-H+1)^2}{nK} - \frac{p-H+1}{n}}\right).$$

- *If* $\min_{(i,j)\in\mathscr{B}} \bar{\theta}_{ij} \geq \sqrt{q}(\rho_{1n} + \tau_2\rho_{2n})$*, then* $\widehat{\theta}^o_{\mathscr{B}_k^c} = 0$ *for the k-th community.*

- Main result: If $n \to \infty$,

$$\widehat{\theta}^o \to \bar{\theta}^o, \quad \text{and} \quad \widehat{\mathbf{q}}^o \to \bar{\mathbf{q}}^o.$$

Fair Ising Graphical Model (FBLASSO)

Letting $G(\Theta) = \Theta$ in (6), our problem takes the form

$$\begin{aligned} \underset{\Theta, Q}{\text{minimize}} \quad & L(\Theta, \mathbf{Q}; \mathbf{Y}) + \rho_1 \|\Theta\|_{1,\text{off}} := \sum_{j=1}^{p} \sum_{j'=1}^{p} \theta_{jj'}(s_{ij'} + \rho_2 q_{jj'}) \\ & - \sum_{i=1}^{n} \sum_{j=1}^{p} \log\left(1 + \exp(\theta_{jj} + \sum_{j' \neq j} \theta_{jj'} y_{ij'})\right) + \rho_1 \sum_{1 \le i < j \le p} |\theta_{ij}|, \\ \text{subj. to} \quad & \Theta \in \mathscr{M} \quad \text{and} \quad \mathbf{Q} \in \mathscr{N}. \end{aligned} \tag{11}$$

Here, $\mathscr{M}$ and $\mathscr{N}$ are the graph and fairness constraints, respectively.

Fair Ising Graphical Model (FBLASSO)

- Denote the log-likelihood for the i-th observation by

$$L_i(\Theta) = \sum_{j=1}^{p} y_{ij}\Big(\sum_{j \neq j'} \theta_{jj'} y_{ij'}\Big) - \log\Big(1 + \exp\big(\sum_{j \neq j'} \theta_{jj'} y_{ij'}\big)\Big). \tag{12}$$

- The second derivative of $L_i(\Theta)$ is given by

$$\nabla^2 L_i(\Theta) = \mathbf{y}_i^\top \Pi_i(\Theta) \mathbf{y}_i, \tag{13}$$

where $\Pi_i(\Theta) = \mathrm{diag}(\pi_{i_1}(\Theta), \ldots, \pi_{i_p}(\Theta))$ is a $p \times p$ diagonal matrix, and

$$\pi_{i_j}(\Theta) = \frac{\exp(\sum_{j' \neq j} \theta_{jj'} y_{ij'})}{1 + \exp(\sum_{j' \neq j} \theta_{jj'} y_{ij'})}.$$

- The population Fisher information matrix of L at $\bar{\theta}^o$ can be expressed as $\bar{\mathbf{H}} = \mathbb{E}(\mathbf{y}_i^\top \Pi_i(\bar{\theta}^o) \mathbf{y}_i)$.

Fair Ising Graphical Model (FBLASSO)

Our results rely on Assumptions 3–4 and the following regularity conditions:

1. There exist constants $\tau_4, \tau_5 \in (0, \infty)$ such that
$$\Lambda_{\min}(\bar{\mathbf{H}}_{\mathscr{B}\mathscr{B}}) \geq \tau_4 \quad \text{and} \quad \Lambda_{\max}(\mathbf{T}) \leq \tau_5.$$
2. There exists a constant $\alpha \in (0, 1]$, such that
$$\|\bar{\mathbf{H}}_{\mathscr{B}^c\mathscr{B}} \left(\bar{\mathbf{H}}_{\mathscr{B}\mathscr{B}}\right)^{-1}\|_\infty \leq (1 - \alpha). \tag{14}$$

Fair Ising Graphical Model (FBLASSO)

Theorem

Suppose Assumptions 1–4 are satisfied. Assume further that $\rho_{1n} = D_{\rho_1}\sqrt{\log p/n}$ *for some constant* $D_{\rho_1} > 16(2-\alpha)/\alpha$, $q\sqrt{(\log p)/n} = o(1)$, $\frac{p-H+1}{n}(\frac{p-H+1}{K} - 1) = o(1)$, *and* $\varepsilon = 0$. *Then, there exist finite constants* $\Pi_2(\bar{\theta}^o, \bar{\mathbf{q}}^o)$ *and* η, *such that the following events hold with probability at least* $1 - O(\exp(-\eta \log p))$:

- *There exists a minimizer* $(\widehat{\theta}^o, \widehat{\mathbf{q}})$ *of* (10).
- *Any minimizer* $(\widehat{\theta}^o, \widehat{\mathbf{q}})$ *of* (10) *satisfies*

$$\|\widehat{\theta}^o - \bar{\theta}^o\| + \|\widehat{\mathbf{q}}^o - \bar{\mathbf{q}}^o\| \leq \Pi_2(\bar{\theta}^o, \bar{\mathbf{q}}^o)\left(\rho_{1n}\sqrt{|\mathscr{B}|} + \rho_{2n}\sqrt{\frac{(p-H+1)^2}{nK} - \frac{p-H+1}{n}}\right).$$

- *If* $\min_{(i,j)\in\mathscr{B}} \bar{\theta}_{ij} \geq \sqrt{q}(\rho_{1n} + \tau_5\rho_{2n})$, *then* $\widehat{\theta}^o_{\mathscr{B}_k^c} = 0$ *for the k-th community.*

- Main result: If $n \to \infty$,

$$\widehat{\theta}^o \to \bar{\theta}^o, \quad \text{and} \quad \widehat{\mathbf{q}}^o \to \bar{\mathbf{q}}^o.$$

Stochastic Block Model (SBM)

- i and j belong to the same cluster: they have a higher probability of connection between them for a fixed value of π_d.
- The vertices also have a higher tendency to connect: if they are connected in the graph specified by π_d, even if they do not belong to the same cluster.
- When π_d itself has a community structure, there are two natural ways to cluster the vertices:
    1. based on the ground-truth clusters $\mathscr{C}_1, \ldots \mathscr{C}_K$ specified by π_c;
    2. based on the clusters specified by π_d.

Stochastic Block Model (SBM)

- Given the matrix $\mathbf{A}$, we set $\mathbf{\Sigma}^{-1}$ equal to $\mathbf{A} + (0.01 - \Lambda_{\min}(\mathbf{A}))\mathbf{I}$, where $\Lambda_{\min}(\mathbf{A})$ is the smallest eigenvalue of $\mathbf{A}$.
- We generate the data matrix $\mathbf{Y}$ according to $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{\text{i.i.d.}}{\sim} \mathbb{N}(\mathbf{0}, \mathbf{\Sigma})$.
- Variables are standardized to have standard deviation one.

Stochastic Block Model (SBM)

We compare our algorithm with two clustering and graphical model estimation methods:

- **FK-means**: Fair K-means clustering [Chierichetti et al. 2017].
- **TCONCORD**: A three-stage approach which (i) uses the joint neighborhood selection approach [Kare et al. 2015] to estimate precision matrices, (ii) applies the robust community detection approach [Cai et al. 2015] to compute partition matrix $\hat{\mathbf{Q}}$, and (iii) employs a fair K-means clustering [Chierichetti et al. 2017] to obtain clusters.

Stochastic Block Model (SBM)

- Clustering error: calculates the distance between an estimated community assignment $\hat{z}_i$ and the true assignment z_i of the sample data $\mathbf{y}_i$:

$$\frac{1}{\binom{n}{2}} \left| \{ (i,j) : \mathbf{1}(\hat{z}_i = \hat{z}_j) \neq \mathbf{1}(z_i = z_j), i < j \} \right|.$$

- Precision matrix error:

$$\frac{1}{K} \sum_{k=1}^{K} \left\| \hat{\Theta}_k - \bar{\Theta}_k \right\|_F .$$

- Balance:

$$\min_{k \in [K]} \frac{\min_{h \in [H]} \left| \mathscr{C}_k \cap \mathscr{D}_h \right|}{\left| \mathscr{C}_k \right|}.$$

	Method	Clustering Error	Precision Matrix Error	Balance
$n = 300$	**FK-means**	0.287(0.005)	N/A	0.292(0.009)
	TCONCORD	0.264(0.005)	5.150 (0.090)	0.387(0.007)
	FCONCORD	0.260(0.005)	5.155 (0.050)	0.404(0.007)
$n = 400$	**FK-means**	0.273(0.007)	N/A	0.287 (0.009)
	TCONCORD	0.248(0.011)	4.561 (0.012)	0.350(0.009)
	FCONCORD	0.213(0.011)	4.147 (0.012)	0.420(0.010)
$n = 500$	**FK-means**	0.229(0.002)	N/A	0.292(0.011)
	TCONCORD	0.217(0.002)	4.032 (0.050)	0.311(0.011)
	FCONCORD	0.201(0.001)	3.501(0.050)	0.419(0.019)

Table: Simulation results of SBM network. The results are for $p = 1000$, $H = 5$, and $K = 5$.

Stochastic Ising Block Model (SIBM)

We consider the SIBM give in
Berthet, Quentin, Philippe Rigollet, and Piyush Srivastava. "Exact recovery in the Ising blockmodel." The Annals of Statistics 47.4 (2019): 1805-1834.

1. We use SBM to generate a graph $\mathscr{G}$ based on an (unknown) partition of the vertex set.
2. We use $\mathscr{G}$ as the underlying dependency graph of the Ising model and draw m i.i.d. samples from it.
3. The objective is to exactly recover the partition of the vertex set in SBM from the samples generated by the Ising model, without observing the graph G.

Stochastic Ising Block Model (SIBM)

We compare the performance of FBLasso to the following clustering and graphical model algorithms:

- **FK-means**: Fair K-means clustering.
- **TBLasso**: A three-stage approach which (i) applies a joint binary neighborhood selection approach [Ravikumar et al. 2010] to estimate precision matrices, (ii) uses a robust community detection approach [Cai2015robust] to compute partition matrix $\hat{\mathbf{Q}}$, and (iii) utilizes a fair K-means clustering to obtain clusters.

Stochastic Ising Block Model (SIBM)

	Method	Clustering Error	Precision Matrix Error	Balance
$n = 300$	**FK-means**	0.318(0.009)	N/A	0.284(0.005)
	TBLasso	0.329(0.009)	10.081 (0.080)	0.334(0.005)
	FBLasso	0.297(0.009)	9.143 (0.025)	0.385(0.005)
$n = 400$	**FK-means**	0.376(0.006)	N/A	0.259 (0.013)
	TBLasso	0.325(0.008)	9.401 (0.031)	0.235(0.005)
	FBLasso	0.307(0.008)	9.001 (0.050)	0.380(0.006)
$n = 500$	**FK-means**	0.348(0.004)	N/A	0.371(0.005)
	TBLasso	0.297(0.004)	9.024 (0.029)	0.406(0.008)
	FBLasso	0.256(0.004)	8.031(0.065)	0.453(0.008)

Table: Simulation results of SIBM network. The results are for $p = 2000$, $H = 5$, and $K = 5$.

FGL on Recommender Systems

1. Kamishima, Toshihiro, et al. "Recommendation independence." Conference on Fairness, Accountability and Transparency. PMLR, 2018.
   - MovieLens 10k dataset: use the year of the movie as a sensitive attribute and consider movies before 1990 as old movies.
2. Abdollahpouri, Himan, et al. "The unfairness of popularity bias in recommendation." arXiv preprint arXiv:1907.13286 (2019).
   - Three different groups of users according to their interest in popular items (Niche, Diverse and Blockbuster-focused) and show the impact of popularity bias on the users in each group.

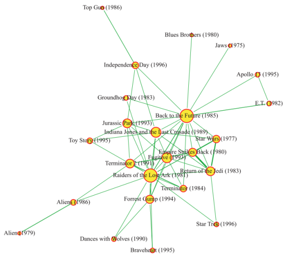
FGL on Real Datasets: Recommender Systems

- We apply FGL to Movielens, a dataset containing rating scores for 1682 movies by 943 users.
- The rating scores have five levels, where 1 corresponds to strong dissatisfaction and 5 to strong satisfaction.

	Method	Clustering Error	Normalized Mutual Information	Balance
$H = 2, K = 3$	**FK-means**	0.380(0.005)	0.110 (0.005)	0.272(0.008)
	TCONCORD	0.244(0.005)	0.129 (0.005)	0.312(0.011)
	FCONCORD	0.219(0.005)	0.151 (0.005)	0.324 (0.011)

Table: The clustering errors, normalized mutual information, and balance of various methods in the Crime Dataset.

FGL on Real Datasets:Recommender Systems



- The estimated network for 32 movies within a community.
- The three large communities mainly consists of mass marketed commercial movies.
- As expected, movies within the same series are most strongly associated.
- Further, Raiders of the Lost Ark (1981) and Back to the Future (1985) form two hub nodes: their common feature is that they were directed/produced by Spielberg.

FGL on Real Datasets: Detection of Toxic Comments

- Detection of Toxic Comments
  - Class distribution of Wikipedia dataset: Clean (201,081), Toxic (21,384), Obscene (12,140), Insult (11,304), Identity Hate (2,117) Severe Toxic (1,962) Threat (689).

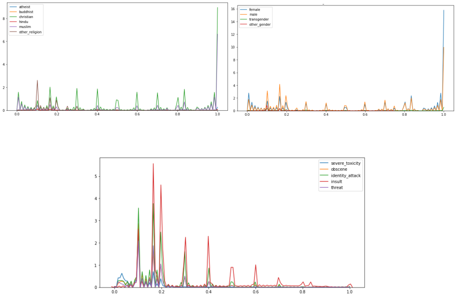


Figure: Distribution of features in the Toxicity dataset: We see that (i) they are lot of values where the target value is 0 and fewer values greater than 1; (ii) values which are less that 0.5 are non–toxic and greater than 0.5 are toxic.

FGL on Real Datasets: Detection of Toxic Comments

- Detection of Toxic Comments
  - The identity label female is regarded as the protected attribute ($H = 2$).
  - There are two neighborhoods defined by whether the comment is regarded toxic or not ($K = 2$).

	Method	Clustering Error	Normalized Mutual Information	Balance
$H = 2, K = 2$	**FK-means**	0.366(0.005)	0.008(0.001)	0.301(0.003)
	TBLasso	0.233(0.009)	0.014(0.001)	0.419(0.003)
	FBLasso	0.214(0.009)	0.017 (0.001)	0.461(0.003)

Table: The clustering errors, normalized mutual information, and balance of various methods in the Toxicity data set.