Mini-course - Probabilistic Graphical Models: A Geometric, Algebraic and Combinatorial Perspective

Caroline Uhler

Lecture 1: Graphical Models and Markov Properties

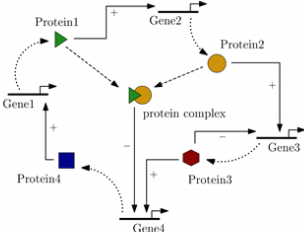
CIMI Workshop on Computational Aspects of Geometry
Toulouse

November 6, 2019

Applications of graphical models

- Probabilistic models that capture the statistical dependencies between variables of interest in the form of a network
- Used throughout the natural sciences, social sciences, and economics for modeling interactions
- Undirected graphical models encode partial correlations, while directed graphical models can be used to represent causality

(a) Weather forecasting



(b) Gene regulation

Graphical models

Motivation: Provide an economic representation of a joint distribution using local relationships between variables

Origins of graphical models can be traced back to 3 communities:

- Statistical physics: use undirected graph to represent distribution over a large system of interacting particles [Gibbs, 1902]
- Genetics: use directed graphs to model inheritance in natural species [Wright, 1921]
- Statistics: use graphs to represent interactions in multi-dimensional contingency tables [Bartlett, 1935]

Graphical models combine **graph theory** with **probability theory** into a powerful framework for multivariate statistical modeling [Lauritzen, 1996]

Algebraic, geometric and combinatorial questions arise naturally when studying graphical models

Overview of mini-course

(1) Introduction to graphical models - Markov properties

(2) Gaussian graphical models - Maximum likleihood estimation

(3) Covariance models with linear structure - Parameter estimation and structure learning

(4) Causal inference - Structure discovery

References: Graphical models

- Bartlett, M. S. (1935). Contingency table interactions. J. Royal Stat. Soc. 2:248-252.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. J. Royal Stat. Soc., 36:192-236.
- Gibbs, J. W. (1902). Elementary Principles in Statistical Mechanics, Yale University Press.
- Lauritzen, S. L. (1996). Graphical Models, Clarendon Press.
- Moussouris, J. (1974). Gibbs and Markov random systems with constraints. J. Stat. Phys., 10:11-33.
- Pearl, J. (1988). Fusion, propagation and structuring in belief networks, Artificial Intelligence, 29, 357-370.
- Shachter, R. D. (1998). The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams), UAI.
- Verma T. & Pearl, J. (1990). Equivalence and synthesis of causal models, UAI.
- Verma T. & Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. UAI.
- Wright, S. (1921). Correlation and causation. J. Agricult. Res., 20:557-585.

Mini-course - Probabilistic Graphical Models: A Geometric, Algebraic and Combinatorial Perspective

Caroline Uhler

Lecture 2: Gaussian Graphical Models

CIMI Workshop on Computational Aspects of Geometry
Toulouse

November 6, 2019

Overview

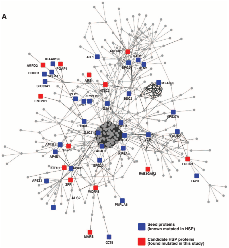
- Lecture is based on a book chapter that I wrote for the Handbook of Graphical Models edited by M. Drton, S. Lauritzen, M. Maathuis and M. Wainwright:

  C. Uhler, "Gaussian graphical models: An algebraic and geometric perspective", available at arXiv:1707.04345

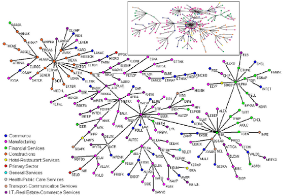
- Goal of this lecture is to give an introduction to Gaussian graphical models and show that algebraic, geometric and combinatorial questions arise naturally when studying graphical models

Gaussian graphical models

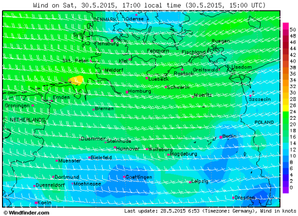
- Goal: Characterize relationship among a large number of variables
- Visualize interactions by graph
- **Gaussian graphical models:** Used throughout the natural sciences, social sciences and economics for modeling interactions among nodes for continuous multivariate data



(a) Gene association network (Novarino et al., Science 343, 2014)



(b) Athens stock exchange (Garos & Argyrakis, Physica A, 2007)



(c) Wind speed forecasting

Gaussian Distribution

A random vector $X \in \mathbb{R}^p$ follows a **multivariate Gaussian distribution** with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{S}^p_{\succ 0}$ if it has density

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

Gaussian Distribution

A random vector $X \in \mathbb{R}^p$ follows a **multivariate Gaussian distribution** with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{S}^p_{\succ 0}$ if it has density

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

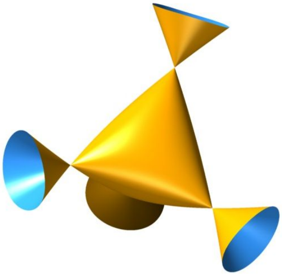
- What can you say about the space of covariance matrices?
- How does the space of 3×3 correlation matrices look like?

Gaussian Distribution

A random vector $X \in \mathbb{R}^p$ follows a **multivariate Gaussian distribution** with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{S}^p_{\succ 0}$ if it has density

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

- What can you say about the space of covariance matrices?
- How does the space of 3×3 correlation matrices look like?



Gaussian Graphical Model

- $G = (V, E)$ undirected graph with vertices $V = \{1, \ldots, p\}$ and edges E
- $\mathcal{K}_G = \{K \in \mathbb{S}^p_{\succ 0} \mid K_{ij} = 0 \text{ for all } i \neq j \text{ with } (i,j) \notin E\}$

A Gaussian vector $X \in \mathbb{R}^p$ is a **Gaussian graphical model** on G if

$$X \sim \mathcal{N}(\mu, \Sigma) \quad \text{and} \quad \Sigma^{-1} \in \mathbb{S}^p_{\succ 0}(G).$$

Gaussian Graphical Model

- $G = (V, E)$ undirected graph with vertices $V = \{1, \ldots, p\}$ and edges E
- $\mathcal{K}_G = \{K \in \mathbb{S}^p_{\succ 0} \mid K_{ij} = 0 \text{ for all } i \neq j \text{ with } (i,j) \notin E\}$

A Gaussian vector $X \in \mathbb{R}^p$ is a **Gaussian graphical model** on G if

$$X \sim \mathcal{N}(\mu, \Sigma) \quad \text{and} \quad \Sigma^{-1} \in \mathbb{S}^p_{\succ 0}(G).$$

Question: Interpretation of missing edges in G?

Marginals and Conditionals of a Gaussian

Theorem

Let $X \sim \mathcal{N}_p(\mu, \Sigma)$ and partition X into two components $X_A \in \mathbb{R}^a$ and $X_B \in \mathbb{R}^b$ such that $a + b = p$. Let μ and Σ be partitioned accordingly, i.e.,

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad and \quad \Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{pmatrix}.$$

Then,

(a) *the marginal distribution of X_A is $\mathcal{N}(\mu_A, \Sigma_{A,A})$;*

(b) *the conditional distribution of $X_A \mid X_B = x_B$ is $\mathcal{N}(\mu_{A|B}, \Sigma_{A|B})$, where*

$$\mu_{A|B} = \mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(x_B - \mu_B) \quad and \quad \Sigma_{A|B} = \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}.$$

Marginals and Conditionals of a Gaussian

Theorem

Let $X \sim \mathcal{N}_p(\mu, \Sigma)$ and partition X into two components $X_A \in \mathbb{R}^a$ and $X_B \in \mathbb{R}^b$ such that $a + b = p$. Let μ and Σ be partitioned accordingly, i.e.,

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad and \quad \Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{pmatrix}.$$

Then,

(a) *the marginal distribution of X_A is $\mathcal{N}(\mu_A, \Sigma_{A,A})$;*

(b) *the conditional distribution of $X_A \mid X_B = x_B$ is $\mathcal{N}(\mu_{A|B}, \Sigma_{A|B})$, where*

$$\mu_{A|B} = \mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(x_B - \mu_B) \quad and \quad \Sigma_{A|B} = \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}.$$

Note: Let $K = \Sigma^{-1}$. Then by Schur complement, $\Sigma_{A|A^c} = (K_{AA})^{-1}$. Hence a missing edge in G means $K_{ij} = 0$, or equivalently, $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$.

Two Main Problems

Given i.i.d. samples $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ from a Gaussian graphical model

- Learn the graph G
  - see tomorrow's lectures (e.g. graphical lasso)
- Estimate the edge weights, i.e. the non-zero entries of Σ^{-1}
  - maximum likelihood estimation

Two Main Problems

Given i.i.d. samples $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^p$ from a Gaussian graphical model

- Learn the graph G
  - see tomorrow's lectures (e.g. graphical lasso)
- Estimate the edge weights, i.e. the non-zero entries of Σ^{-1}
  - maximum likelihood estimation

- These problems don't depend on mean μ; w.l.o.g. assume $\mu = 0$

Two Main Problems

Given i.i.d. samples $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ from a Gaussian graphical model

- Learn the graph G
  - see tomorrow's lectures (e.g. graphical lasso)
- Estimate the edge weights, i.e. the non-zero entries of Σ^{-1}
  - maximum likelihood estimation

- These problems don't depend on mean μ; w.l.o.g. assume $\mu = 0$
- **sample covariance matrix** is given by

$$S = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}(X^{(i)})^T \in \mathbb{S}^p_{\succeq 0}, \quad \mathrm{rk}(S) = n \leq p \text{ with probability } 1$$

Two Main Problems

Given i.i.d. samples $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^p$ from a Gaussian graphical model

- Learn the graph G
  - see tomorrow's lectures (e.g. graphical lasso)
- Estimate the edge weights, i.e. the non-zero entries of Σ^{-1}
  - maximum likelihood estimation

- These problems don't depend on mean μ; w.l.o.g. assume $\mu = 0$
- **sample covariance matrix** is given by

$$S = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}(X^{(i)})^T \in \mathbb{S}^p_{\succeq 0}, \quad \mathrm{rk}(S) = n \leq p \text{ with probability } 1$$

- **log-likelihood** is given by: $\ell(\Sigma; S) \propto -\log\det(\Sigma) - \mathrm{tr}\left(S\Sigma^{-1}\right)$

Two Main Problems

Given i.i.d. samples $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ from a Gaussian graphical model

- Learn the graph G
  - see tomorrow's lectures (e.g. graphical lasso)
- Estimate the edge weights, i.e. the non-zero entries of Σ^{-1}
  - maximum likelihood estimation

- These problems don't depend on mean μ; w.l.o.g. assume $\mu = 0$
- **sample covariance matrix** is given by

$$S = \frac{1}{n} \sum_{i=1}^{n} X^{(i)} (X^{(i)})^T \in \mathbb{S}^p_{\succeq 0}, \quad \mathrm{rk}(S) = n \leq p \text{ with probability } 1$$

- **log-likelihood** is given by: $\ell(\Sigma; S) \propto -\log\det(\Sigma) - \mathrm{tr}\left(S\Sigma^{-1}\right)$
- What can be said about the log-likelihood function?

Parameter estimation for Gaussian graphical models

Given a graph G, the maximum likelihood estimator (**MLE**) $\hat{K} := \hat{\Sigma}^{-1}$ solves the following convex optimization problem:

$$\begin{aligned} \text{maximize} \quad & \log\det K - \operatorname{tr}(SK) \\ \text{subject to} \quad & K \in \mathcal{K}_G \end{aligned}$$

Question: What is the MLE when G is the complete graph?

Parameter estimation for Gaussian graphical models

By **strong duality**: Given a graph G, the MLE $\hat{K} := \hat{\Sigma}^{-1}$ solves the following equivalent convex optimization problems:

$$\begin{array}{ll} \text{maximize} & \log\det K - \operatorname{tr}(KS) \\ \text{subject to} & K_{ij} = 0, \ \forall (i,j) \notin E \end{array} \qquad \begin{array}{ll} \text{minimize} & -\log\det \Sigma - p \\ \text{subject to} & \Sigma_{ij} = S_{ij}, \ (i,j) \in E \text{ or } i = j \end{array}$$

Parameter estimation for Gaussian graphical models

By **strong duality**: Given a graph G, the MLE $\hat{K} := \hat{\Sigma}^{-1}$ solves the following equivalent convex optimization problems:

$$\begin{array}{ll} \text{maximize} & \log\det K - \operatorname{tr}(KS) \\ \text{subject to} & K_{ij} = 0, \forall (i,j) \notin E \end{array} \qquad \begin{array}{ll} \text{minimize} & -\log\det \Sigma - p \\ \text{subject to} & \Sigma_{ij} = S_{ij}, (i,j) \in E \text{ or } i = j \end{array}$$

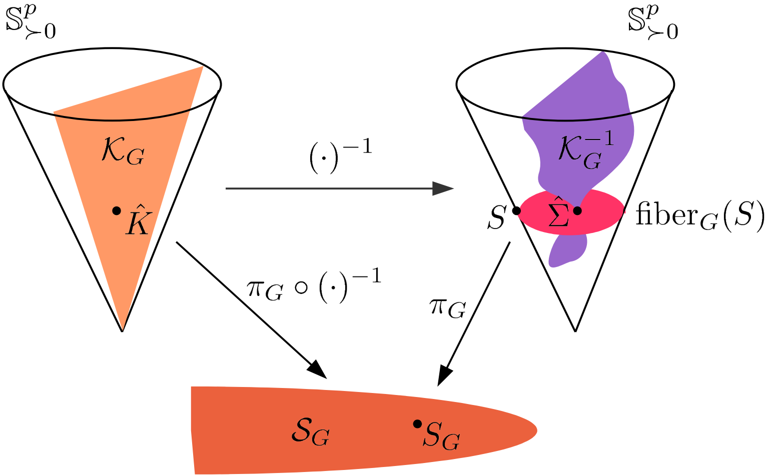
Theorem (Dempster 1972)

*In a Gaussian graphical model on G the MLE $\hat{\Sigma}$ exists if and only if the partial sample covariance matrix $S_G = (S_{ij} \mid (i,j) \in E \text{ or } i = j)$ (**sufficient statistics**) can be extended to a positive definite matrix. Then the MLE $\hat{\Sigma}$ is the unique completion whose inverse satisfies*

$$(\hat{\Sigma}^{-1})_{ij} = 0, \ \forall\, i \neq j, \ (i,j) \notin E.$$

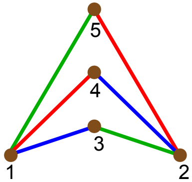
Existence of MLE is equivalent to positive definite matrix completion problem!

Geometric Picture



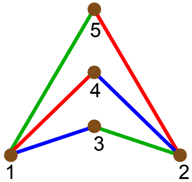
- $S_G := \pi_G(S)$, $\mathcal{S}_G := \pi_G(\mathbb{S}^p_{\succeq 0})$; note that $\mathcal{S}_G = \mathcal{K}_G^\vee$
- $\text{fiber}_G(S) := \{\Sigma \in \mathbb{S}^p_{\succeq 0} \mid \Sigma_G = S_G\}$

Example $K_{2,3}$



$$K = \begin{pmatrix} \lambda_1 & 0 & \lambda_2 & \lambda_3 & \lambda_4 \\ 0 & \lambda_1 & \lambda_4 & \lambda_2 & \lambda_3 \\ \lambda_2 & \lambda_4 & \lambda_1 & 0 & 0 \\ \lambda_3 & \lambda_2 & 0 & \lambda_1 & 0 \\ \lambda_4 & \lambda_3 & 0 & 0 & \lambda_1 \end{pmatrix}$$

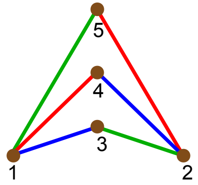
Example $K_{2,3}$



$$K = \begin{pmatrix} \lambda_1 & 0 & \lambda_2 & \lambda_3 & \lambda_4 \\ 0 & \lambda_1 & \lambda_4 & \lambda_2 & \lambda_3 \\ \lambda_2 & \lambda_4 & \lambda_1 & 0 & 0 \\ \lambda_3 & \lambda_2 & 0 & \lambda_1 & 0 \\ \lambda_4 & \lambda_3 & 0 & 0 & \lambda_1 \end{pmatrix}$$

$$\begin{aligned} \det(K) &= \lambda_1 \cdot (\lambda_1^2 - \lambda_2^2 + \lambda_2\lambda_3 - \lambda_3^2 + \lambda_2\lambda_4 + \lambda_3\lambda_4 - \lambda_4^2) \cdot \\ &\quad (\lambda_1^2 - \lambda_2^2 - \lambda_2\lambda_3 - \lambda_3^2 - \lambda_2\lambda_4 - \lambda_3\lambda_4 - \lambda_4^2) \end{aligned}$$

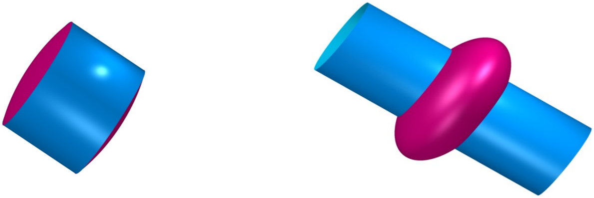
Example $K_{2,3}$



$$K = \begin{pmatrix} \lambda_1 & 0 & \lambda_2 & \lambda_3 & \lambda_4 \\ 0 & \lambda_1 & \lambda_4 & \lambda_2 & \lambda_3 \\ \lambda_2 & \lambda_4 & \lambda_1 & 0 & 0 \\ \lambda_3 & \lambda_2 & 0 & \lambda_1 & 0 \\ \lambda_4 & \lambda_3 & 0 & 0 & \lambda_1 \end{pmatrix}$$

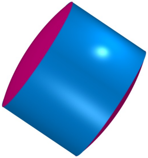
$$\begin{aligned} \det(K) \quad = \quad & \lambda_1 \cdot (\lambda_1^2 - \lambda_2^2 + \lambda_2\lambda_3 - \lambda_3^2 + \lambda_2\lambda_4 + \lambda_3\lambda_4 - \lambda_4^2) \cdot \\ & (\lambda_1^2 - \lambda_2^2 - \lambda_2\lambda_3 - \lambda_3^2 - \lambda_2\lambda_4 - \lambda_3\lambda_4 - \lambda_4^2) \end{aligned}$$

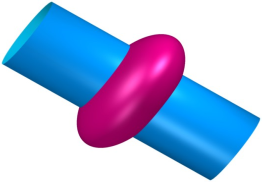
$\mathcal{K}_G$:



Example $K_{2,3}$

$\mathcal{K}_G$:

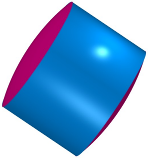


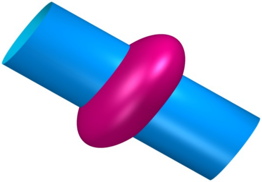


$\mathcal{C}_G$:

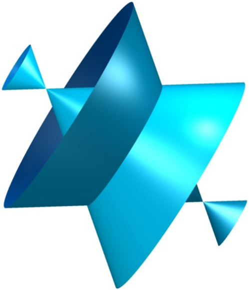
Example $K_{2,3}$

$\mathcal{K}_G$:



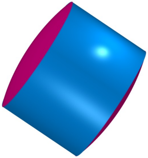


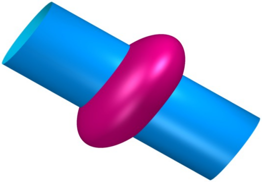
$\mathcal{C}_G$:



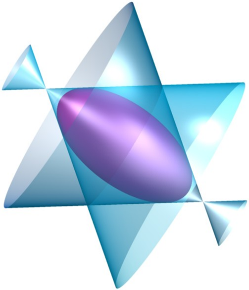
Example $K_{2,3}$

$\mathcal{K}_G$:



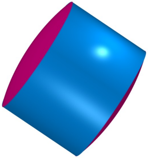


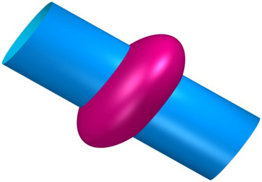
$\mathcal{C}_G$:



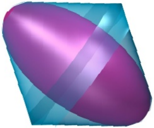
Example $K_{2,3}$

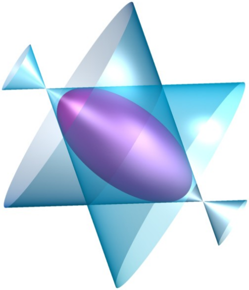
$\mathcal{K}_G$:





$\mathcal{C}_G$:





Positive Definite (pd) Matrix Completion Problem

- Necessary condition for existence of pd completion:

Positive Definite (pd) Matrix Completion Problem

- Necessary condition for existence of pd completion: all specified minors are pd
- However, this is in general not sufficient:

$$S_G = \begin{pmatrix} 1 & 0.9 & ? & -0.9 \\ 0.9 & 1 & 0.9 & ? \\ ? & 0.9 & 1 & 0.9 \\ -0.9 & ? & 0.9 & 1 \end{pmatrix} \text{ does not have a pd completion.}$$

Theorem (Grone, Johnson, Sá & Wolkovicz, 1984)

For a graph G the following statements are equivalent:

(a) *A G-partial matrix $M_G \in \mathbb{R}^{|E^*|}$ has a pd completion if and only if all completely specified submatrices in M_G are positive definite.*

(b) *G is chordal (also known as triangulated), i.e. every cycle of length 4 or larger has a chord.*

Statistical Problem

Current statistical applications:

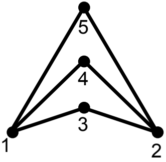
- Number of variables $>>$ Number of observations
- Example: Genetic networks

  Gene expression data of a few individuals to model interaction between large number of genes

$\rightarrow$ **Gaussian graphical models** widely used in this context

Problem: What is the minimum number of observations for existence of the MLE in a given Gaussian graphical model?

Example $K_{2,3}$

$G =$ 

What is the minimal rank n^* such that

$$S_G = \begin{pmatrix} s_{11} & ? & s_{13} & s_{14} & s_{15} \\ ? & s_{22} & s_{23} & s_{24} & s_{25} \\ s_{13} & s_{23} & s_{33} & ? & ? \\ s_{14} & s_{24} & ? & s_{44} & ? \\ s_{15} & s_{25} & ? & ? & s_{55} \end{pmatrix}$$

can be completed to a positive definite matrix for any $S \in \mathbb{S}^p_{\succeq 0}$ of rank $n \geq n^*$?

Bounds

Let n_G^* denote the minimal rank such that every $S \in \mathbb{S}^p_{\succeq 0}$ has a positive definite completion on G

Bounds

Let n_G^* denote the minimal rank such that every $S \in \mathbb{S}_{\succeq 0}^p$ has a positive definite completion on G

- $n_G^* \geq$ maximal clique size of G
- $n_G^* \leq p$

Bounds

Let n_G^* denote the minimal rank such that every $S \in \mathbb{S}_{\succeq 0}^p$ has a positive definite completion on G

- $n_G^* \geq$ maximal clique size of G
- $n_G^* \leq p$

Theorem (Grone, Johnson, Sá & Wolkovicz, 1984)

For chordal (i.e. triangulated) graphs $n^* =$ maximal clique size of G.

Bounds

Let n_G^* denote the minimal rank such that every $S \in \mathbb{S}^p_{\succeq 0}$ has a positive definite completion on G

- $n_G^* \geq$ maximal clique size of G
- $n_G^* \leq p$

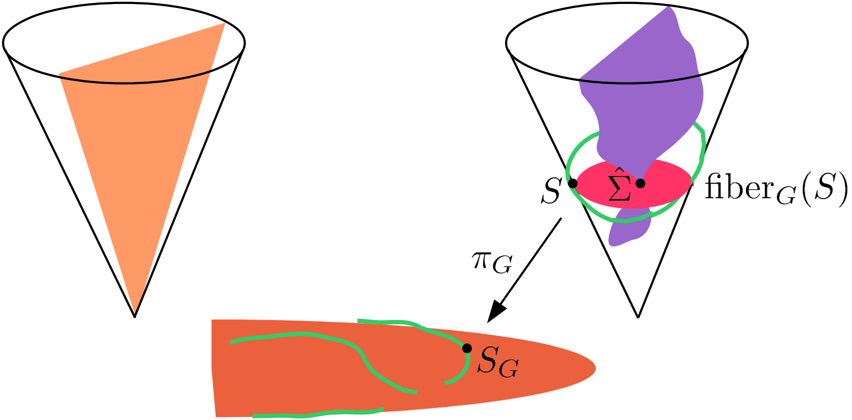
Theorem (Grone, Johnson, Sá & Wolkovicz, 1984)

For chordal (i.e. triangulated) graphs $n^* =$ maximal clique size of G.

Let G be non-chordal. Then

- $n_G^* \geq$ maximal clique size of G
- $n_G^* \leq$ maximal clique size in minimal chordal cover of G

Geometric View



Elimination Criterion

Theorem (Uhler, 2012)

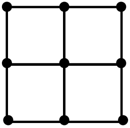
Let I_n be the ideal of $(n+1) \times (n+1)$ minors of a symmetric $p \times p$ matrix of unknowns S. Let $I_{G,n}$ be the elimination ideal obtained from I_n by eliminating all unknowns corresponding to non-edges in the graph. If

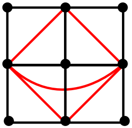
$$I_{G,n} = 0$$

then $n_G^* \leq n$.

- I_n corresponds to all symmetric matrices of rank $\leq n$
- Elimination corresponds to projection onto $\mathcal{S}_G$
- $I_{G,n} = 0$ means that the projection is full-dimensional

3×3 grid



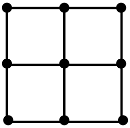


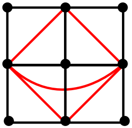
Theorem (Uhler, 2012)

When G is the 3×3 grid, then $n_G^* = 3$.

- First example of a graph for which $n_G^* <$ maximal clique size in minimal chordal cover
- Solves an open problem by Steffen Lauritzen

3×3 grid





Theorem (Uhler, 2012)

When G is the 3×3 grid, then $n_G^* = 3$.

- First example of a graph for which $n_G^* <$ maximal clique size in minimal chordal cover
- Solves an open problem by Steffen Lauritzen

Theorem (Gross and Sullivant, 2018)

For any grid, $n_G^* = 3$. Furthermore, for any planar graph, $n^* \leq 4$.

Computing the MLE

- Convex optimization problem; can be solved e.g. using interior point methods or coordinate descent algorithms (often faster)
- There is a closed-form formula for the MLE $\iff$ G is chordal (*Lauritzen, 1996*)
- ML-degree: maximal number of solutions to the likelihood equations
- There is a rational formula for the MLE (in the entries of S) $\iff$ ML-degree is 1 $\iff$ G is chordal (*Sturmfels & Uhler, 2010*)
- **Conjecture** The p-cycle maximizes the ML-degree over all graphs on p nodes and has ML-degree $(p-3)2^{p-2}+1$

Alternative Approach: Sparsity Order of a Graph

- S_G PD completable if and only if $\langle S_G, X \rangle > 0$ for all $X \in \mathcal{K}_G$ extremal
- Knowledge of extremal rays of $\mathcal{K}_G$ is useful for deciding PD completability

Alternative Approach: Sparsity Order of a Graph

- S_G PD completable if and only if $\langle S_G, X\rangle > 0$ for all $X \in \mathcal{K}_G$ extremal
- Knowledge of extremal rays of $\mathcal{K}_G$ is useful for deciding PD completability

The **sparsity order** of a graph G is defined as

$$\operatorname{ord}(G) = \max\{\operatorname{rk}(X) \mid X \in \mathcal{K}_G \text{ extremal}\}$$

- There should be strong connections between existence of the MLE, ML-degree and sparsity order of a graph, but these are still quite unclear (*Solus, Uhler & Yoshida, 2016*)

Sparsity Order of a Graph

- $\mathrm{ord}(G) = 1$ if and only if G chordal (Agler et al., 1988)
- If H is an induced subgraph of G, then $\mathrm{ord}(H) \leq \mathrm{ord}(G)$ (Agler et al., 1988)
- If G is the clique sum of two graphs G_1 and G_2, then $\mathrm{ord}(G) = \max\{\mathrm{ord}(G_1), \mathrm{ord}(G_2)\}$ (Helton et al. 1989)
- $\mathrm{ord}(G) \leq p - 2$ with equality if and only if G is a p-cycle; the extremal ranks are 1 and $p-2$ (Helton et al. 1989)
- $\mathrm{ord}(K_{m,n}) = \begin{cases} \frac{m^2-m}{2}+1 & \text{if } n \geq \frac{m^2-m}{2}+1 \\ n & \text{otherwise} \end{cases}$; (Grone & Pierce, 1990)
  all ranks $1, \ldots, \mathrm{ord}(K_{m,n})$ are extremal
- All graphs of order 2 have been characterized (Laurent, 2001)
- Many many open problems...

References

- Agler, Helton, McCullough & Rodman: Positive semidefinite matrices with a given sparsity pattern (Linear algebra and its applications 107, 1988)
- Boyd & Vandenberghe: Convex Optimization (Cambridge University Press, 2004)
- Dempster: Covariance selection (Biometrics 28, 1972)
- Grone, Johnson, Sà & Wolkowicz: Positive definite completions of partial Hermitian matrices (Linear Algebra and its Applications 58, 1984)
- Grone & Pierce: Extremal bipartite matrices (Linear Algebra & its Applications 131, 1990)
- Gross & Sullivant: The maximum likelihood threshold of a graph (Bernoulli 24, 2018)
- Helton, Pierce & Rodman: The ranks of extremal positive semidefinite matrices with given sparsity pattern (SIAM Journal on Matrix Analysis and Applications 10, 1989)
- Laurent: On the sparsity order of a graph and its deficiency in chordality (Combinatorica 21, 2001)
- Lauritzen: Graphical Models (Oxford University Press, 1996)
- Solus, Uhler & Yoshida: Extremal positive semidefinite matrices whose sparsity pattern is given by graphs without K_5 minors (Linear Algebra and its Applications 509, 2016)
- Sturmfels & Uhler: Multivariate Gaussians, semidefinite matrix completion, and convex algebraic geometry (Annals of the Institute of Statistical Mathematics 62, 2010)
- Uhler: Geometry of maximum likelihood estimation in Gaussian graphical models (Annals of Statistics 40, 2012)

Mini-course - Probabilistic Graphical Models: A Geometric, Algebraic and Combinatorial Perspective

Caroline Uhler

Lecture 3: Structure Learning in Undirected Graphical Models

CIMI Workshop on Computational Aspects of Geometry
Toulouse

November 7, 2019

(Undirected) Gaussian graphical models

- $X \sim \mathcal{N}(0, \Sigma)$, $K := \Sigma^{-1}$, $p =$nr. of variables, $n =$nr. of samples

- Gaussian graphical model: $(i,j) \notin E$ if and only if $K_{ij} = 0$

  if and only if $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$

(Undirected) Gaussian graphical models

- $X \sim \mathcal{N}(0, \Sigma)$, $K := \Sigma^{-1}$, $p =$nr. of variables, $n =$nr. of samples

- Gaussian graphical model: $(i,j) \notin E$ if and only if $K_{ij} = 0$
  if and only if $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$

- Sample covariance matrix S is of rank $\min(n, p)$

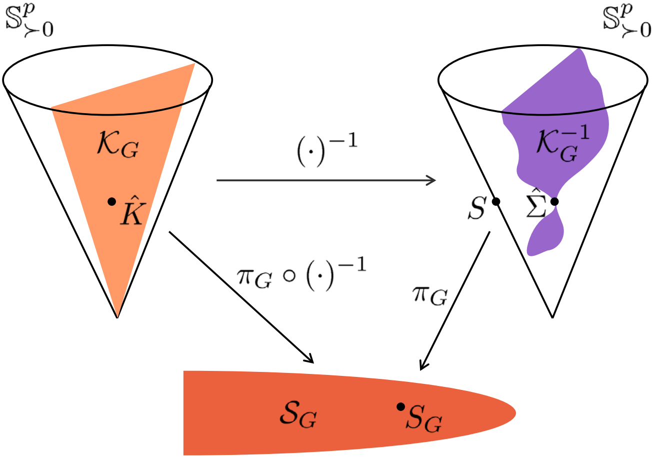
- MLE:

$$\hat{K} = \operatorname{argmax}\{\log\det(K) - \operatorname{trace}(SK) \mid K \succeq 0,\ K_{ij} = 0\ \forall (i,j) \notin E\}$$

  - In general unbounded if $n < p$
  - Given a graph G what is the minimal n such that this problem is bounded (i.e., the MLE exists)?

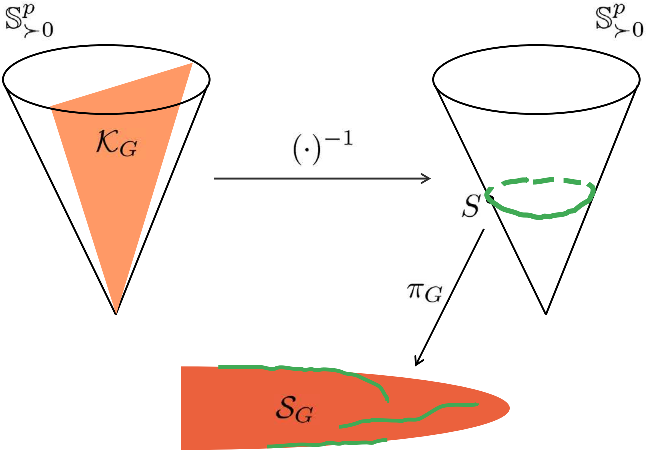
    $\rightarrow$ Geometric problem

Geometric Picture



- π_G : projection onto edge set, $S_G := \pi_G(S)$, $\mathcal{S}_G := \pi_G(\mathbb{S}^p_{\succeq 0})$
- Note that $\mathcal{S}_G = \mathcal{K}_G^\vee$
- MLE for S exists if and only if $S_G \in \mathrm{int}(\mathcal{S}_G)$

Geometric Picture



MLE exists for n samples, if projection of manifold of rank n psd matrices lies in the interior of the cone $\mathcal{S}_G$ [Uhler, arXiv:1707.04345]

Structure learning in (undirected) graphical models

- MLE: $\hat{K} = \mathrm{argmax}\{\log\det(K) - \mathrm{trace}(SK)\}$
  - $\hat{K}$ is dense even if $n \gg p$

Structure learning in (undirected) graphical models

- MLE: $\hat{K} = \text{argmax}\{\log\det(K) - \text{trace}(SK)\}$
  - $\hat{K}$ is dense even if $n \gg p$

- Graphical lasso: $\hat{K}_\lambda = \text{argmax}\{\log\det(K) - \text{trace}(SK) - \lambda|K|_1\}$
  - sparsistent for particular choice of λ (under certain assumptions)
    [Ravikumar, Wainwright, Raskutti & Yu, 2011]
  - $\hat{K}_\lambda$ is not monotone in λ: edges can disappear/appear for increasing λ
    [Fattahi & Sojoudi, 2019]
  - $\hat{K}_\lambda$ is not invariant to rescaling

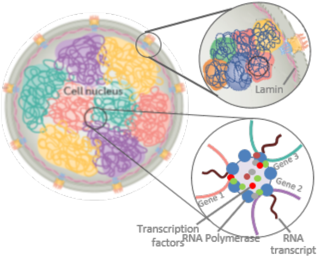
Structure learning in (undirected) graphical models

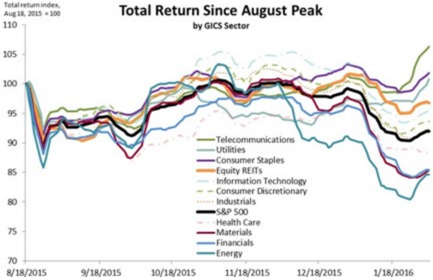
- MLE: $\hat{K} = \text{argmax}\{\log\det(K) - \text{trace}(SK)\}$
  - $\hat{K}$ is dense even if $n \gg p$

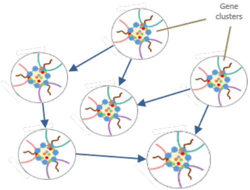
- Graphical lasso: $\hat{K}_\lambda = \text{argmax}\{\log\det(K) - \text{trace}(SK) - \lambda|K|_1\}$
  - sparsistent for particular choice of λ (under certain assumptions) [Ravikumar, Wainwright, Raskutti & Yu, 2011]
  - $\hat{K}_\lambda$ is not monotone in λ: edges can disappear/appear for increasing λ [Fattahi & Sojoudi, 2019]
  - $\hat{K}_\lambda$ is not invariant to rescaling

- Additional approaches include:
  - node-wise regression with the lasso (*Meinshausen & Bühlmann, 2006*)
  - CLIME: constrained ℓ_1-based optimization (*Cai, Liu & Luo, 2011*)
  - Algorithm with false discovery rate control (*Liu, 2013*)
  - ROCKET: for heavy-tailed distributions (*Foygel-Barber & Kolar, 2018*)
  - Conditional independence testing

Motivation: Graphical models under positive dependence







How to model strong forms of positive dependence in data?

Positive dependence and MTP_2 distributions

A distribution (i.e. density function) p on $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$, with $\mathcal{X}_v \subseteq \mathbb{R}$ discrete or open, is **multivariate totally positive of order 2 (MTP_2)** if

$$p(x)p(y) \quad \leq \quad p(x \wedge y)p(x \vee y) \qquad \text{for all } x, y \in \mathcal{X},$$

where $\wedge$ and $\vee$ are applied coordinate-wise.

Theorem (Fortuin Kasteleyn Ginibre inequality, 1971, Karlin & Rinott, 1980)

MTP_2 *implies positive association, i.e.*

$$\mathrm{cov}\{\phi(X), \psi(X)\} \geq 0$$

for any non-decreasing functions $\phi, \psi : \mathbb{R}^m \to \mathbb{R}$.

Positive dependence and MTP_2 distributions

A distribution (i.e. density function) p on $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$, with $\mathcal{X}_v \subseteq \mathbb{R}$ discrete or open, is **multivariate totally positive of order 2 (MTP_2)** if

$$p(x)p(y) \quad \leq \quad p(x \wedge y)p(x \vee y) \qquad \text{for all } x, y \in \mathcal{X},$$

where $\wedge$ and $\vee$ are applied coordinate-wise.

Theorem (FortuinKasteleynGinibre inequality, 1971, Karlin & Rinott, 1980)

MTP_2 *implies positive association, i.e.*

$$\mathrm{cov}\{\phi(X), \psi(X)\} \geq 0$$

for any non-decreasing functions $\phi, \psi : \mathbb{R}^m \to \mathbb{R}$.

Theorem (FLSUWZ, 2017)

If $p(x) > 0$ *and* MTP_2*, then* $p(x)$ *is faithful to an undirected graph.*

Gaussian MTP_2 distributions

Theorem (Bølviken 1982, Karlin & Rinott, 1983)

A multivariate Gaussian distribution $p(x; K)$ is MTP_2 if and only if the inverse covariance matrix K is an M-matrix, that is

$$K_{uv} \leq 0 \qquad \text{for all } u \neq v.$$

Gaussian MTP_2 distributions

Theorem (Bølviken 1982, Karlin & Rinott, 1983)

A multivariate Gaussian distribution $p(x;K)$ is MTP_2 if and only if the inverse covariance matrix K is an M-matrix, that is

$$K_{uv} \leq 0 \qquad \textit{for all } \; u \neq v.$$

Ex: 2016 Monthly correlations of global stock markets *(InvestmentFrontier.com)*

$$S = \begin{array}{c} \\ \end{array}\begin{pmatrix} 1.000 & 0.606 & 0.731 & 0.618 & 0.613 \\ 0.606 & 1.000 & 0.550 & 0.661 & 0.598 \\ 0.731 & 0.550 & 1.000 & 0.644 & 0.569 \\ 0.618 & 0.661 & 0.644 & 1.000 & 0.615 \\ 0.613 & 0.598 & 0.569 & 0.615 & 1.000 \end{pmatrix}$$

	Nasdaq	Canada	Europe	UK	Australia
Nasdaq	1.000	0.606	0.731	0.618	0.613
Canada	0.606	1.000	0.550	0.661	0.598
Europe	0.731	0.550	1.000	0.644	0.569
UK	0.618	0.661	0.644	1.000	0.615
Australia	0.613	0.598	0.569	0.615	1.000

Gaussian MTP_2 distributions

Theorem (Bølviken 1982, Karlin & Rinott, 1983)

A multivariate Gaussian distribution $p(x; K)$ is MTP_2 if and only if the inverse covariance matrix K is an M-matrix, that is

$$K_{uv} \leq 0 \qquad \text{for all } u \neq v.$$

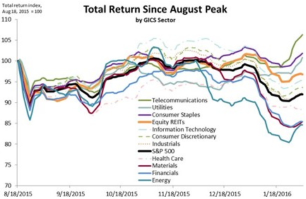
Ex: 2016 monthly correlations of global stock markets *(InvestmentFrontier.com)*

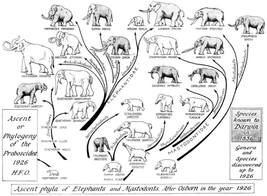
$$S^{-1} = \begin{pmatrix} 2.629 & -0.480 & -1.249 & -0.202 & -0.490 \\ -0.480 & 2.109 & -0.039 & -0.790 & -0.459 \\ -1.249 & -0.039 & 2.491 & -0.675 & -0.213 \\ -0.202 & -0.790 & -0.675 & 2.378 & -0.482 \\ -0.490 & -0.459 & -0.213 & -0.482 & 1.992 \end{pmatrix}$$

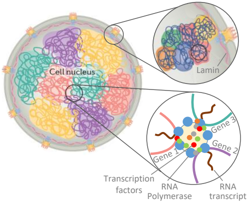
	Nasdaq	Canada	Europe	UK	Australia
Nasdaq	2.629	−0.480	−1.249	−0.202	−0.490
Canada	−0.480	2.109	−0.039	−0.790	−0.459
Europe	−1.249	−0.039	2.491	−0.675	−0.213
UK	−0.202	−0.790	−0.675	2.378	−0.482
Australia	−0.490	−0.459	−0.213	−0.482	1.992

Sample distribution is MTP_2! If you sample a correlation matrix uniformly at random the probability of it being MTP_2 is $< 10^{-6}$!

MTP_2 constraints are often implicit







X is MTP_2 in:

- ferromagnetic Ising models
- Markov chains with MTP_2 transitions
- order statistics of i.i.d. variables
- Brownian motion tree models

$|X|$ is MTP_2 in:

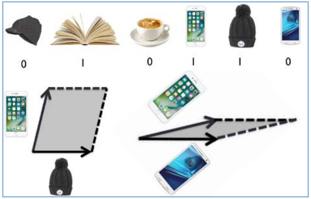
- Gaussian / binary tree models
- Gaussian / binary latent tree models
  - Binary latent class models
  - Single factor analysis models

Negative dependence: NOT analogous!!

- Analog of FKG inequality does not hold: negative association, i.e. $\text{cov}\{\phi(X), \psi(X)\} \leq 0$ for any non-decreasing functions ϕ, ψ is not implied by $p(x)p(y) \geq p(x \wedge y)p(x \vee y)$ for all x, y.
- See Pemantle (1999): Towards a Theory of Negative Association
- Strongly Rayleigh measures: sufficient for conditionally negative association [Borcea, Bränden & Liggett, 2009]

Negative dependence: NOT analogous!!

- Analog of FKG inequality does not hold: negative association, i.e. $\mathrm{cov}\{\phi(X), \psi(X)\} \leq 0$ for any non-decreasing functions ϕ, ψ is not implied by $p(x)p(y) \geq p(x \wedge y)p(x \vee y)$ for all x, y.
- See Pemantle (1999): Towards a Theory of Negative Association
- Strongly Rayleigh measures: sufficient for conditionally negative association [Borcea, Bränden & Liggett, 2009]
- Recently used in various machine learning applications to enforce diversity, e.g. recommender systems, neural network sparsification, matrix sketching, diversity priors



- See NeurIPS 2018 Tutorial by Jegelka & Sra

ML Estimation for Gaussian MTP_2 distributions

Let S be the sample covariance matrix. Then maximum likelihood estimation is a convex optimization problem:

Primal: Max-Likelihood

$$\begin{array}{ll} \underset{K \succeq 0}{\text{maximize}} & \log\det(K) - \mathrm{trace}(KS) \\ \text{subject to} & K_{uv} \leq 0, \ \ \forall\, u \neq v. \end{array}$$

Dual: Entropy

$$\begin{array}{ll} \underset{\Sigma \succeq 0}{\text{minimize}} & -\log\det(\Sigma) - p \\ \text{subject to} & \Sigma_{vv} = S_{vv},\ \Sigma_{uv} \geq S_{uv}. \end{array}$$

ML Estimation for Gaussian MTP_2 distributions

Let S be the sample covariance matrix. Then maximum likelihood estimation is a convex optimization problem:

Primal: Max-Likelihood

$$\begin{array}{ll} \underset{K \succeq 0}{\text{maximize}} & \log\det(K) - \mathrm{trace}(KS) \\ \text{subject to} & K_{uv} \leq 0, \;\; \forall\, u \neq v. \end{array}$$

Dual: Entropy

$$\begin{array}{ll} \underset{\Sigma \succeq 0}{\text{minimize}} & -\log\det(\Sigma) - p \\ \text{subject to} & \Sigma_{vv} = S_{vv},\; \Sigma_{uv} \geq S_{uv}. \end{array}$$

Theorem (Slawski & Hein, 2015)

The MLE in a Gaussian MTP_2 model exists with probability 1 when $n \geq 2$.

New proof: 3 lines using ultrametrics [Lauritzen, U. & Zwiernik, 2019]

ML Estimation for Gaussian MTP_2 distributions

Let S be the sample covariance matrix. Then maximum likelihood estimation is a convex optimization problem:

Primal: Max-Likelihood

$$\begin{aligned} \underset{K \succeq 0}{\text{maximize}} \quad & \log\det(K) - \mathrm{trace}(KS) \\ \text{subject to} \quad & K_{uv} \leq 0, \ \forall\, u \neq v. \end{aligned}$$

Dual: Entropy

$$\begin{aligned} \underset{\Sigma \succeq 0}{\text{minimize}} \quad & -\log\det(\Sigma) - p \\ \text{subject to} \quad & \Sigma_{vv} = S_{vv}, \ \Sigma_{uv} \geq S_{uv}. \end{aligned}$$

Theorem (Slawski & Hein, 2015)

The MLE in a Gaussian MTP_2 model exists with probability 1 when $n \geq 2$.

New proof: 3 lines using ultrametrics [Lauritzen, U. & Zwiernik, 2019]

Theorem (Wang, Roy & U., 2019)

Graphical model inference by testing the signs of the empirical partial correlation coefficients is consistent in the high-dimensional setting without the need of any tuning parameter. With ℓ_1-penalty, the resulting estimator is monotone.

Application: Portfolio selection

[Agrawal, Roy & U., 2019]

Daily stock return data from the Center for Research in Security Prices (CRSP) between 1975-2015 (NYSE, AMEX & NASDAQ stock exchanges).

M (nr. of assets)	T (lookback period)	EW-TQ	Linear Shrinkage	Approximate Factor Model	MTP_2
100	25	0.694	0.710	0.730	0.803
	50	0.694	0.625	0.637	0.849
	100	0.694	0.600	0.617	0.896
	200	0.694	0.670	0.688	0.899
	400	0.694	0.736	0.782	0.892
	1260	0.694	0.831	0.834	0.890
200	50	0.757	0.742	0.726	0.853
	100	0.757	0.719	0.716	0.829
	200	0.757	0.812	0.800	0.885
	400	0.757	0.864	0.870	0.886
	800	0.757	0.967	0.961	0.970
	1260	0.757	0.906	0.916	0.955
500	125	0.764	0.876	0.872	1.019
	250	0.764	0.985	0.977	1.112
	500	0.764	0.940	0.980	1.045
	1000	0.764	0.918	0.978	1.061

Information ratio (ratio of average return to standard deviation of returns) when weights are estimated based on "full" Markowitz portfolio problem

Conclusions

Graphical models combine graph theory with probability theory into a powerful framework for multivariate statistical modeling

- Total positivity constraints are often implicit and reflect real processes
  - ferromagnetism
  - latent tree models
- MTP_2 implies faithfulness
- MTP_2 is well-suited for high-dimensional applications (also in non-parametric setting, see our recent work)
- Explicit MTP_2 constraints enhance interpretability of graphical models (induce sparsity without the need of a tuning parameter)
- MTP2 distributions not only have broad applications (finance, psychology, genomics), but also lead to beautiful theory (exponential families, convexity, combinatorics, semialgebraic geometry)

References: Graphical models

- Bartlett (1935). Contingency table interactions.
- Cai, Liu & Luo (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation
- Fattahi & Sojoudi (2019). Graphical lasso and thresholding: Equivalence and closed-form solutions.
- Foygel-Barber & Kolar (2018) ROCKET: robust confidence intervals via Kendall's Tau for transelliptical graphical models
- Friedman, Hastie, and Tibshirani (2007). Sparse inverse covariance estimation with the graphical lasso.
- Gibbs (1902). Elementary Principles in Statistical Mechanics, Yale Univ. Press.
- Lauritzen, S. L. (1996). Graphical Models, Clarendon Press.
- Liu (2013) Gaussian graphical model estimation with false discovery rate control
- Meinshausen & Bühlmann (2006). High-dimensional graphs and variable selection with the lasso
- Ravikumar, Wainwright, Raskutti & Yu (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence.
- Uhler (2017). Gaussian graphical models: An algebraic and geometric perspective.
- Wright, S. (1921). Correlation and causation. J. Agricult. Res., 20:557-585.

References: Total positivity

- Agrawal, Roy & Uhler (2019). Covariance matrix estimation under total positivity for portfolio selection
- Bolviken (1982). Probability inequalities for the multivariate normal with non-negative partial correlations.
- Fallat, Lauritzen, Sadeghi, Uhler, Wermuth & Zwiernik (2017). Total positivity in Markov structures.
- Fortuin, Kasteleyn & Ginibre, J. (1971). Correlation inequalities on some partially ordered sets.
- Karlin & Rinott (1983). M-matrices as covariance matrices of multinormal distributions.
- Lauritzen, Uhler & Zwiernik (2019). Maximum likelihood estimation in Gaussian models under total positivity.
- Lauritzen, Uhler & Zwiernik (2019). Total positivity in structured binary distributions (arXiv:1905.00516)
- Lebowitz (1972). Bounds on the correlations and analyticity properties of ferromagnetic Ising spin systems.
- Slawski & Hein (2014). Estimation of positive definite M-matrices and structure learningfor attractive Gaussian Markov random fields
- Wang, Roy, & Uhler (2019). Learning high-dimensional Gaussian graphical models under total positivity without adjustment of tuning parameters (arXiv:1906.05159)

Mini-course - Probabilistic Graphical Models: A Geometric, Algebraic and Combinatorial Perspective

Caroline Uhler

Lecture 4: Causal Structure Discovery

CIMI Workshop on Computational Aspects of Geometry
Toulouse

November 7, 2019

Causal inference

- Framework for causal inference from observational data (structural equation models) developed in 1920's by J. Neyman and S. Wright
- Skepticism amongst statisticians halted the developments for 50 years
- Reemergence in the 1970's after major contributions by J. Pearl (CS), J. Robins (epidemiology), D. Rubin (stats) & P. Spirtes (philosophy)

Causal inference

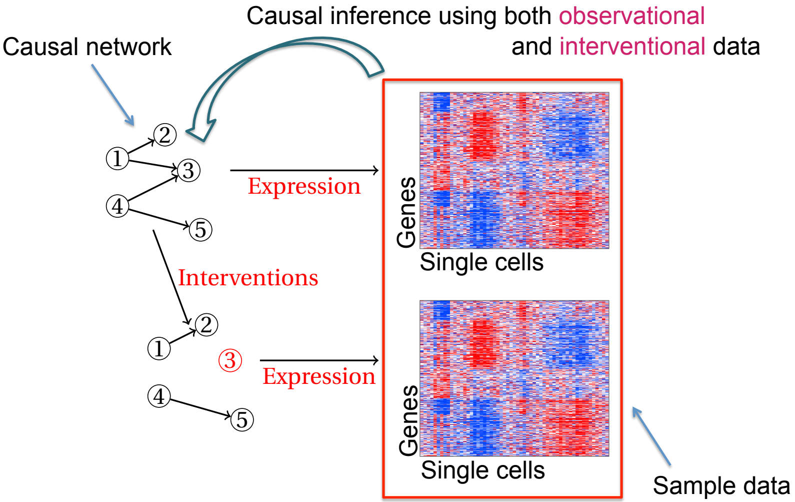
- Framework for causal inference from observational data (structural equation models) developed in 1920's by J. Neyman and S. Wright
- Skepticism amongst statisticians halted the developments for 50 years
- Reemergence in the 1970's after major contributions by J. Pearl (CS), J. Robins (epidemiology), D. Rubin (stats) & P. Spirtes (philosophy)

✶ Interaction between genetics and causal inference could be particularly beneficial:

- Geneticists can perform interventional experiments relatively easily
- Drop-seq and Perturb-seq: High-throughput (100,000-1 mio single-cell measurements on all 20,000 genes per experiment) observational and interventional single-cell RNA-seq data is now available

✶ Unique data and challenges!

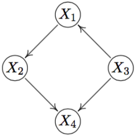
Gene expression data - single-cell RNA-seq



Perturb-seq: High-throughput observational and interventional single-cell RNA-seq data is now available [*Dixit et al., 2016*]

Structural equation models

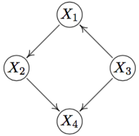
- Introduced by Sewell Wright in the 1920s
- Represent causal relationships by a directed acyclic graph (DAG)
- Each node is associated with a random variable; stochasticity is introduced by independent noise variables ϵ_i



$$
\begin{aligned}
X_1 &= f_1(X_3, \epsilon_1)\\
X_2 &= f_2(X_1, \epsilon_2)\\
X_3 &= f_3(\epsilon_3)\\
X_4 &= f_4(X_2, X_3, \epsilon_4)
\end{aligned}
$$

Structural equation models

- Introduced by Sewell Wright in the 1920s
- Represent causal relationships by a directed acyclic graph (DAG)
- Each node is associated with a random variable; stochasticity is introduced by independent noise variables ϵ_i

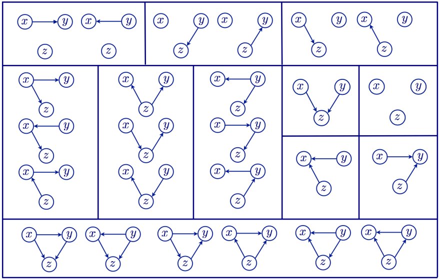


$$
\begin{aligned}
X_1 &= f_1(X_3, \epsilon_1) \\
X_2 &= f_2(X_1, \epsilon_2) \\
X_3 &= f_3(\epsilon_3) \\
X_4 &= f_4(X_2, X_3, \epsilon_4)
\end{aligned}
$$

- Structural equation model also defines interventional distribution:
  - Perfect (hard) intervention on X_2: $X_2 = c$
  - General intervention on X_2: $X_2 = \tilde{f}_2(X_1, \tilde{\epsilon}_2)$

Markov equivalence classes on 3 nodes & talk overview

- Markov equivalence: different DAGs can encode same conditional independence relations (through factorization of the joint distribution)



✶ Interventional Markov equivalence classes?

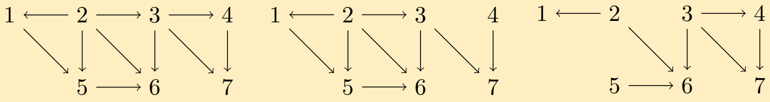
✶ How do they depend on the type of intervention? Do perfect interventions provide smaller equivalence classes than imperfect interventions?

- Algorithms for learning the interventional Markov equivalence class?

Interventional Markov equivalence class

- Let $\mathcal{I}$ be a set of intervention targets

Ex: Perfect interventions $\mathcal{I} = \{\emptyset, \{4\}, \{3, 5\}\}$

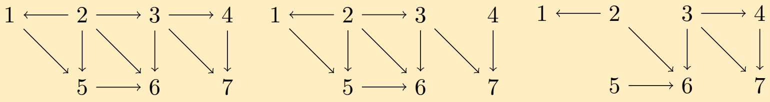


(a) $G^{\emptyset}$ (b) $G^{(\{4\})}$ (c) $G^{(\{3,5\})}$

Interventional Markov equivalence class

- Let $\mathcal{I}$ be a set of intervention targets

Ex: Perfect interventions $\mathcal{I} = \{\emptyset, \{4\}, \{3,5\}\}$



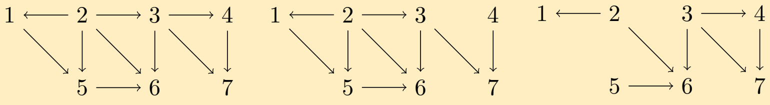
(a) $G^{\emptyset}$ (b) $G^{(\{4\})}$ (c) $G^{(\{3,5\})}$

- *Hauser and Bühlmann (2012)*: characterized $\mathcal{I}$-Markov equivalence classes under perfect interventions: an edge is orientable if it is
  - orientable from observational data
  - adjacent to an intervened node

Interventional Markov equivalence class

- Let $\mathcal{I}$ be a set of intervention targets

Ex: Perfect interventions $\mathcal{I} = \{\emptyset, \{4\}, \{3,5\}\}$



(a) $G^{\emptyset}$ (b) $G^{(\{4\})}$ (c) $G^{(\{3,5\})}$

- *Hauser and Bühlmann (2012)*: characterized $\mathcal{I}$-Markov equivalence classes under perfect interventions: an edge is orientable if it is
  - orientable from observational data
  - adjacent to an intervened node

Theorem (Yang, Katcoff & Uhler, ICML 2018)

The $\mathcal{I}$-Markov equivalence classes under perfect and imperfect interventions are the same.

Proof: By introducing & providing a graphical criterion for the $\mathcal{I}$-Markov property for $\mathcal{I}$-DAGs.

Algorithms for learning causal graphs

There are two main types of algorithms for learning causal graphs from observational data:

- **Constraint-based:** treat causal search as constraint satisfaction problem; constraints given by conditional independence; main example: PC algorithm [Spirtes, Glymour & Scheines, 2001]

  Properties: very fast, with consistency guarantees (with prob. 1 as $n \to \infty$), require large sample size, tend to miss edges

- **Score-based:** maximize score (e.g. BIC) of a Markov equivalence class with respect to a data set by greedy search; main example: Greedy Equivalence Search (GES) [Chickering, 2002]

  Properties: higher accuracy for same sample size, huge search space, theoretical consistency guarantees

Limitation of score-based approaches

Table 1: Equivalence Class Counts

n	Equivalence classes	Cl/ADG	Cl_1/Cl
1	1	1.00000	1.00000
2	2	0.66667	0.50000
3	11	0.44000	0.36364
4	185	0.34070	0.31892
5	8782	0.29992	0.29788
6	1067825	0.28238	0.28667
7	312510571	0.27443	0.28068
8	212133402500	0.27068	0.27754
9	326266056291213	0.26888	0.27590
10	1118902054495975141	0.26799	0.27507

(*Gillispie & Perlman, 2001*)

Problem of enumerating Markov equivalence classes and their sizes leads to hard and beautiful combinatorics problems: e.g., formula for number of equivalence classes on p nodes? Average size of equivalence classes?

[Radhakrishnan, Solus, Uhler, UAI 2017]

[Katz-Rogozhnikov, Shanmugam, Squires, Uhler, AISTATS 2019]

Limitation of constraint-based approaches

Constraint-based methods require the faithfulness assumption:

$$(i,j) \in E \quad \Longleftrightarrow \quad X_i \not\perp\!\!\!\perp X_j \mid X_S \qquad \forall S \subset V \backslash \{i,j\}$$

[*Zhang &Spirtes, 2003*]

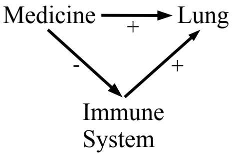
Limitation of constraint-based approaches

Constraint-based methods require the faithfulness assumption:

$$(i,j) \in E \quad \Longleftrightarrow \quad X_i \not\perp\!\!\!\perp X_j \mid X_S \qquad \forall S \subset V \backslash \{i,j\}$$

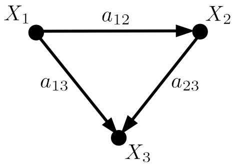
[*Zhang &Spirtes, 2003*]

Ex:



Faithfulness means that causal effects cannot cancel out!

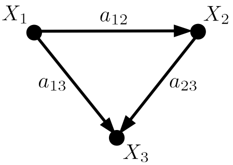
Unfaithful distributions: 3-node example



$X_1 = \epsilon_1$
$X_2 = a_{12}X_1 + \epsilon_2$
$X_3 = a_{13}X_1 + a_{23}X_2 + \epsilon_3$
$\epsilon \sim \mathcal{N}(0, I)$

$\Longrightarrow X \sim \mathcal{N}(0, \Sigma), \;\; \Sigma^{-1} = (I - A)(I - A)^T$

Unfaithful distributions: 3-node example



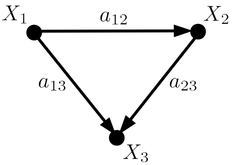
$X_1 = \epsilon_1$
$X_2 = a_{12}X_1 + \epsilon_2$
$X_3 = a_{13}X_1 + a_{23}X_2 + \epsilon_3$
$\epsilon \sim \mathcal{N}(0, I)$

$\implies X \sim \mathcal{N}(0, \Sigma), \ \ \Sigma^{-1} = (I - A)(I - A)^T$

Faithfulness is **NOT** satisfied if any of the following relations hold:

- $X_1 \perp\!\!\!\perp X_2 \iff \det((\Sigma^{-1})_{13,23}) = a_{12} = 0$
- $X_1 \perp\!\!\!\perp X_3 \iff \det((\Sigma^{-1})_{12,23}) = a_{13} + a_{12}a_{23} = 0$
- $X_2 \perp\!\!\!\perp X_3 \iff \det((\Sigma^{-1})_{12,13}) = a_{12}^2 a_{23} + a_{12}a_{13} + a_{23} = 0$
- $X_1 \perp\!\!\!\perp X_2 \mid X_3 \iff \det((\Sigma^{-1})_{1,2}) = a_{13}a_{23} - a_{12} = 0$
- $X_1 \perp\!\!\!\perp X_3 \mid X_2 \iff \det((\Sigma^{-1})_{1,3}) = -a_{13} = 0$
- $X_2 \perp\!\!\!\perp X_3 \mid X_1 \iff \det((\Sigma^{-1})_{2,3}) = -a_{23} = 0$

Unfaithful distributions: 3-node example



$X_1 = \epsilon_1$
$X_2 = a_{12}X_1 + \epsilon_2$
$X_3 = a_{13}X_1 + a_{23}X_2 + \epsilon_3$
$\epsilon \sim \mathcal{N}(0, I)$

$\implies X \sim \mathcal{N}(0, \Sigma), \;\; \Sigma^{-1} = (I - A)(I - A)^T$

Faithfulness is **NOT** satisfied if any of the following relations hold:

- $X_1 \perp\!\!\!\perp X_2 \iff \det((\Sigma^{-1})_{13,23}) = a_{12} = 0$
- $X_1 \perp\!\!\!\perp X_3 \iff \det((\Sigma^{-1})_{12,23}) = a_{13} + a_{12}a_{23} = 0$
- $X_2 \perp\!\!\!\perp X_3 \iff \det((\Sigma^{-1})_{12,13}) = a_{12}^2 a_{23} + a_{12}a_{13} + a_{23} = 0$
- $X_1 \perp\!\!\!\perp X_2 \mid X_3 \iff \det((\Sigma^{-1})_{1,2}) = a_{13}a_{23} - a_{12} = 0$
- $X_1 \perp\!\!\!\perp X_3 \mid X_2 \iff \det((\Sigma^{-1})_{1,3}) = -a_{13} = 0$
- $X_2 \perp\!\!\!\perp X_3 \mid X_1 \iff \det((\Sigma^{-1})_{2,3}) = -a_{23} = 0$

$\implies$ Faithfulness not satisfied on collection of **hypersurfaces** in $\mathbb{R}^{|E|}$

3-node example continued [Uhler, Raskutti, Bühlmann & Yu, Ann. Stat. 2013]

3-node example continued [Uhler, Raskutti, Bühlmann & Yu, Ann. Stat. 2013]

- For consistency of constraint-based algorithms data has to be bounded away from these hypersurfaces by $\sqrt{\log(p)/n}$
- For high-dimensional consistency: $p_n = o(\log(n))$

Alternative approach: Permutation-based searches

Idea: DAG defined by ordering of vertices (**permutation**) and **skeleton**

- For $p = 10$ search space is of size $10! = 3,628,800$ versus 10^{18}

Alternative approach: Permutation-based searches

Idea: DAG defined by ordering of vertices (**permutation**) and **skeleton**

- For $p = 10$ search space is of size $10! = 3,628,800$ versus 10^{18}
- For each permutation π construct a DAG $G_\pi = (V, E_\pi)$ by
  $(\pi(i), \pi(j)) \in E_\pi \iff X_{\pi(i)} \not\perp\!\!\!\perp X_{\pi(j)} \mid X_{\{\pi(1),\dots,\pi(i-1),\pi(i+1),\dots\pi(j-1)\}}$

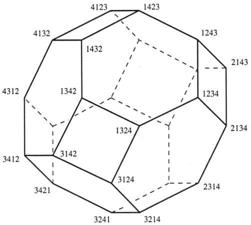
Alternative approach: Permutation-based searches

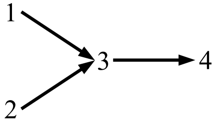
Idea: DAG defined by ordering of vertices (**permutation**) and **skeleton**

- For $p = 10$ search space is of size $10! = 3,628,800$ versus 10^{18}
- For each permutation π construct a DAG $G_\pi = (V, E_\pi)$ by
  $$(\pi(i), \pi(j)) \in E_\pi \iff X_{\pi(i)} \not\perp\!\!\!\perp X_{\pi(j)} \mid X_{\{\pi(1),\ldots,\pi(i-1),\pi(i+1),\ldots\pi(j-1)\}}$$
- Greedy search for sparsest permutation G_{π^*} (GSP) is consistent under strictly weaker conditions than faithfulness

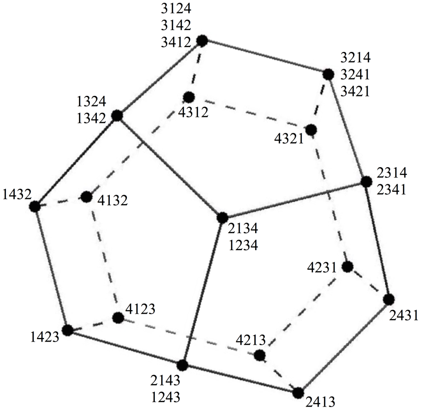
  [Mohammadi, Uhler, Wang & Yu, SIAM J. Discr. Math., 2018]
  [Solus, Wang, Matejovicova & Uhler, arXiv:1702.03530]

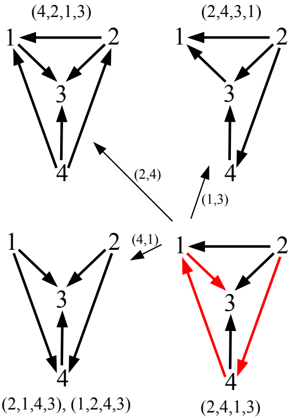
edges in polytope of permutations (i.e., permutohedron) connect neighboring transpositions, e.g. $(3,1,4,2) - (3,4,1,2)$





CI relations: $1 \perp\!\!\!\perp 2,\ 1 \perp\!\!\!\perp 4 \mid 3,\ 1 \perp\!\!\!\perp 4 \mid \{2,3\}$
$2 \perp\!\!\!\perp 4 \mid 3,\ 2 \perp\!\!\!\perp 4 \mid \{1,3\}$



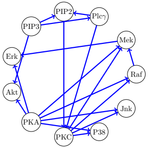


Learning the interventional Markov equivalence class

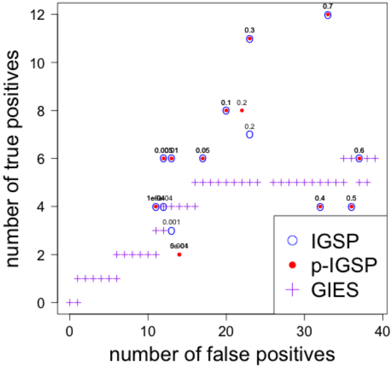
- **GIES:** perfect intervention adaptation of GES [Hauser & Bühlmann, 2012]
  - In general not consistent [Wang, Solus, Yang & Uhler, *NIPS* 2017]

- **IGSP:** interventional adaptation of GSP: provably consistent algorithm that can deal with interventional data
  - for perfect interventions [Wang, Solus, Yang & Uhler, *NIPS* 2017]
  - for general interventions [Yang, Katcoff & Uhler, *ICML* 2018]

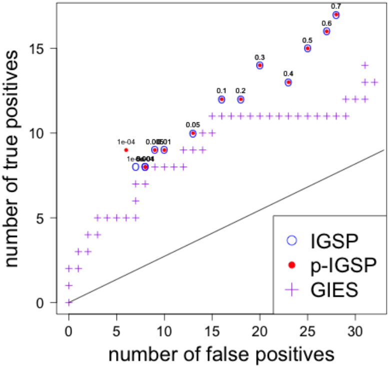
Note: While for perfect interventions it is sufficient to perform conditional independence tests, for general interventions we need to test whether a conditional distribution is invariant to the interventions



Protein signaling network described by Sachs et al. (2005); 7466 measurements of the abundance of phosphoproteins and phospholipids recorded under different interventional experiments;



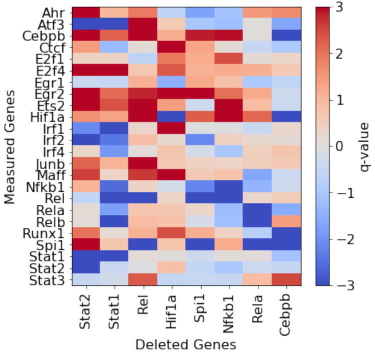
(a) Directed edge recovery

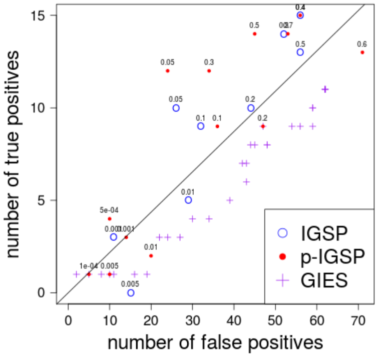


(b) Skeleton recovery

Perturb-seq data

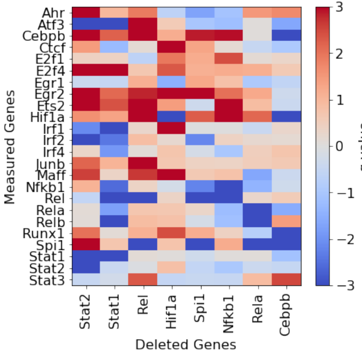
[Yang, Katcoff & Uhler, 2018]

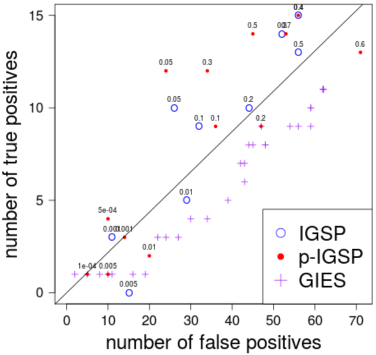




- After preprocessing: 992 observational samples and 13,435 interventional samples from 8 gene deletions; analyzed 24 genes of interest
- Predicted effect of each intervention when leaving out that data

Perturb-seq data



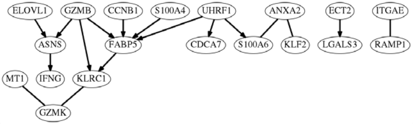


- After preprocessing: 992 observational samples and 13,435 interventional samples from 8 gene deletions; analyzed 24 genes of interest
- Predicted effect of each intervention when leaving out that data
- Much work remains to be done to deal with zero-inflated data, off-target intervention effects, and latent variables; see our recent work [arXiv:1906.00928, 1910.09014, 1910.09007]

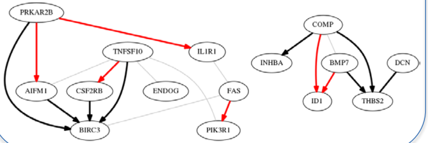
Causal inference and genomics

- Often interested in difference of regulatory network, e.g. between normal / diseased states; learn difference directly without estimating each network separately! [Wang, Squires, Belyaeva & Uhler, NeurIPS 2018]

Difference network of naïve versus activated T-cells (estimated from single-cell RNA-seq)



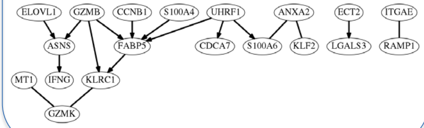
Difference network of ovarian cancer cells from 2 patient cohorts with different survival rates



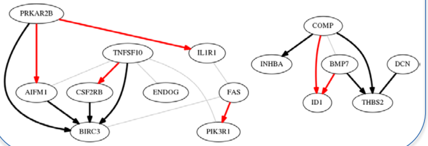
Causal inference and genomics

- Often interested in difference of regulatory network, e.g. between normal / diseased states; learn difference directly without estimating each network separately! [Wang, Squires, Belyaeva & Uhler, NeurIPS 2018]

Difference network of naïve versus activated T-cells (estimated from single-cell RNA-seq)



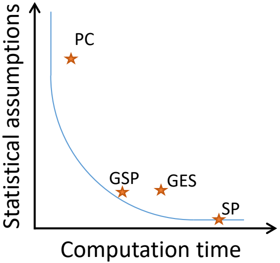
Difference network of ovarian cancer cells from 2 patient cohorts with different survival rates



- Tractable strategy to select interventions in batches under budget constraints for causal inference with provable guarantees on both approximation and optimization quality based on submodularity [Agrawal, Squires, Yang, Shanmugam & Uhler, AISTATS 2019]

Statistical-computational trade-off

Open problem: Characterize the statistical-computational trade-off that is inherent to causal inference



- What is the optimal algorithm for unlimited computation time? (Conjecture: SP algorithm)
- How much weaker than faithfulness are SMR (necessary and sufficient assumption for SP) or triangle-faithfulness assumption (only violations that are undetectable)?
- What is the optimal tradeoff curve?

References

- Peters: Causality script (http://web.math.ku.dk/~peters/jonas_files/scriptChapter1-4.pdf)
- Pearl: The Book of Why (Basic Books, 2018)

- Wright: Correlation and causation (J. Agricult. Res. 10, 1921)
- Neyman: Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes (Roczniki Nauk Rolniczych 10, 1923)
- Spirtes, Glymour & Scheines: Causation, Prediction and Search (MIT Press, 2001)
- Pearl: Causality: Models, Reasoning and Inference (Cambridge University Press, 2000)
- Rubin: Causal inference using potential outcomes (J. Am. Stat. Ass. 100, 2005)
- Robins: Association, causation, and marginal structural models (Synthese 121, 1999)

- Verma & Pearl: Equivalence and synthesis of causal models (UAI, 1990)
- Gillispie and Perlman: Enumerating Markov equivalence classes of acyclic digraph models (UAI 2001)
- Radhakrishnan, Solus & Uhler: A combinatorial perspective of Markov equivalence classes for DAG models (UAI 2017)
- Katz, Shanmugam, Squires & Uhler: Size of interventional Markov equivalence classes in random DAG models (AISTATS, 2019)

References

- Zhang & Spirtes: The three faces of faithfulness (Synthese 193, 2016)
- Uhler, Raskutti, Bühlmann & Yu: Geometry of faithfulness assumption in causal inference (Ann. Statist. 41, 2013)

- Chickering: Optimal structure identification with greedy search (J. Mach. Learn. Res. 3, '02)
- Friedman & Koller: Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks (Mach. Learn. 50, 2003)
- Mohammadi, Uhler, Wang & Yu: Generalized permutohedra from probabilistic graphical models (SIAM J. Discr. Math. 32, 2018)
- Agrawal, Broderick & Uhler: Minimal I-MAP MCMC for scalable structure discovery in causal DAG models (ICML 2018)

- Hauser & Bühlmann: Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs (J. Royal Stat. Soc. 77, '15)
- Wang, Solus, Yang & Uhler: Permutation-based causal inference algorithms with interventions (NIPS 2017)
- Yang, Katcoff & Uhler: Characterizing and learning equivalence classes of causal DAGs under Interventions (ICML 2018)
- Eberhardt, Glymour & Scheines: On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables (UAI, 2005)
- Agrawal, Squires, Yang, Shanmugam & Uhler: ABCD-Strategy: Budgeted experimental design for targeted causal structure discovery (AISTATS 2019)