Probabilistic graphical models and their application to extreme value statistics

Frank Röttger (University of Twente)

44th Finnish Summer School on Probability and Statistics
May 25-29, 2026

UNIVERSITY
OF TWENTE.

Structure of the Mini-Course

- **Part 1:** Undirected graphical models
- **Part 2:** Directed graphical models
- **Part 3:** Graphical models in extremes

Intended learning goals

- Get an introduction to the vast area of graphical models.
- In particular, learn about graphical models in extremes.

Part 1: Undirected graphical models

Part 1: Table of contents

Literature

- For graphical models, the book of Lauritzen:

  Lauritzen, S. L. (1996).
  ***Graphical models*, Volume 17 of *Oxford Statistical Science Series*.**
  The Clarendon Press, Oxford University Press, New York.
  Oxford Science Publications.

- Recent book on causal inference, including directed graphical models:

  Peters, J., D. Janzing, and B. Schölkopf (2017).
  ***Elements of causal inference: foundations and learning algorithms*.**
  The MIT Press.

- Review article, in particular wrt structure learning:

  Drton, M. and M. H. Maathuis (2017).
  Structure learning in graphical modeling.
  Annual Review of Statistics and Its Application *4*(1), 365–393.

Motivation and preliminaries

Motivation: Why dependence modeling?

Why do we care about dependence modeling?

- Ignoring (conditional) dependence among variables can lead to wrong conclusions.
- Tendency to mix up correlation, dependence and causality. $\Rightarrow$ *Dependence modeling provides a theoretical framework for these notions.*
- Explaining dependence can have a strong impact, think of smoking policies or the claim that one glass of wine per day is healthy.

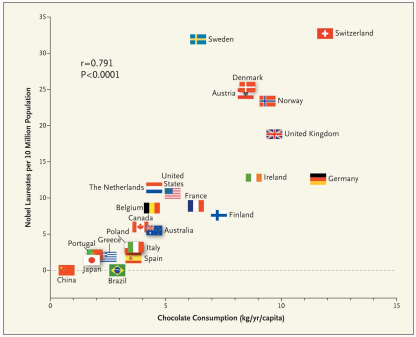


Figure 1: Annual chocolate consumption vs. Nobel laureates quota (Messerli (2012))

Motivation: Binary vectors

- Let's assume a statistical model with three Bernoulli random variables (potentially dependent):
  - X_1 = Rain
  - X_2 = Wet Ground
  - X_3 = Umbrella
- Each can be 0 or 1 $\Rightarrow$ $2^3 = 8$ possible outcomes.

Rain (X_1)	Wet Ground (X_2)	Umbrella (X_3)	Probability
0	0	0	p_1
0	0	1	p_2
0	1	0	p_3
0	1	1	p_4
1	0	0	p_5
1	0	1	p_6
1	1	0	p_7
1	1	1	p_8

- We have eight parameters $p_1, \ldots, p_8 \ldots$ but they must sum to 1.
  $\Rightarrow$ There are seven free parameters in this model!

The Explosion of Parameters

- For three binary variables (as in the previous slide): $2^3 - 1 = 7$ parameters.
- For ten binary variables: $2^{10} - 1 = 1023$ parameters!
- For twenty binary variables: $2^{20} - 1 > 10^6$ parameters!!

Problems

- Exponential parameter growth (2^d) makes storage and computation infeasible.
- Data scarcity: Many configurations are never observed $\Rightarrow$ unreliable estimates and overfitting.

Key Question

How can we reduce the number of parameters by exploiting structure (e.g., independencies), while modeling important relationships between variables?

Conditional independence

Independence: recap

"X_1 and X_2 are independent if observing X_1 provides no information about X_2."

Discrete case

Two discrete random variables X_1 and X_2 are independent if

$$\Pr(X_1 = x_1,\, X_2 = x_2) = \Pr(X_1 = x_1)\,\Pr(X_2 = x_2).$$

for every x_1, x_2.

Continuous case

Two continuous random variables X_1 and X_2 with joint density $f(x_1, x_2)$ and marginal densities $f_1(x_1)$, $f_2(x_2)$ are independent if

$$f(x_1, x_2) = f_1(x_1)\, f_2(x_2), \quad \forall (x_1, x_2) \in \mathbb{R}^2.$$

We typically write $X_1 \perp\!\!\!\perp X_2$ to denote that X_1 and X_2 are independent.

Conditional independence for discrete random variables

"Conditional independence is independence for conditional probability distributions."

Discrete case

Two discrete random variables X_1 and X_2 are conditionally independent given a third discrete variable X_3 if

$$\Pr(X_1 = x_1, X_2 = x_2 \mid X_3 = x_3) = \Pr(X_1 = x_1 \mid X_3 = x_3) \Pr(X_2 = x_2 \mid X_3 = x_3), \quad \forall (x_1, x_2, x_3).$$

We typically write $X_1 \perp\!\!\!\perp X_2 \mid X_3$ to denote that X_1 and X_2 are conditionally independent given X_3.

Example

- X_1 = Rain
- X_2 = Wet Ground
- X_3 = Umbrella

$\Longrightarrow$ $X_2 \perp\!\!\!\perp X_3 \mid X_1$ says that given Rain, Wet Ground carries no information for Umbrella (and vice versa).

Example

Example

- $X_1 = \text{Rain}$
- $X_2 = \text{Wet Ground}$
- $X_3 = \text{Umbrella}$

$\Longrightarrow$ $X_2 \perp\!\!\!\perp X_3 \mid X_1$ says that given Rain, Wet Ground carries no information for Umbrella (and vice versa).

For example, this imposes

$$
\begin{aligned}
& \Pr(X_1 = 1,\, X_2 = 1 \mid X_3 = 1) = \Pr(X_1 = 1 \mid X_3 = 1)\, \Pr(X_2 = 1 \mid X_3 = 1) \\
\Longleftrightarrow\ & \Pr(X_1 = 1,\, X_2 = 1, X_3 = 1)\, \Pr(X_3 = 1) = \Pr(X_1 = 1, X_3 = 1)\, \Pr(X_2 = 1, X_3 = 1) \\
\Longleftrightarrow\ & p_8(p_2 + p_4 + p_6 + p_8) = (p_6 + p_8)(p_4 + p_8) \\
\Longleftrightarrow\ & p_8 p_2 - p_6 p_4 = 0,
\end{aligned}
$$

effectively reducing the dimension of the parameter space.

$\rightarrow$ Note that the model is "carved out" by polynomial constraints!

Conditional independence for continuous random vectors

Conditional independence for continuous random vectors

Conditional density

Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$ be random vectors with joint density $f_{X,Y}$. Assume the marginal density $f_X(x) = \int_{\mathbb{R}^m} f_{X,Y}(x, y)\, \mathrm{d}y$ satisfies $f_X(x) > 0$. Then the conditional density of Y given $X = x$ is

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \qquad y \in \mathbb{R}^m.$$

Conditional independence

X and Y are conditionally independent given $Z = (Z_1, \dots, Z_k)$ (denoted $X \perp\!\!\!\perp Y \mid Z$) if, for all z with $f_Z(z) > 0$,

$$f_{X,Y|Z=z}(x, y) = f_{X|Z=z}(x)\, f_{Y|Z=z}(y), \quad (x, y) \in \mathbb{R}^n \times \mathbb{R}^m.$$

Example: Conditional Independence in a Trivariate Gaussian

Model: Let $X = (X_1, X_2, X_3)^\top \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}$.

Conditional distribution

$$(X_1, X_2) \mid X_3 \sim \mathcal{N}(\mu_{|3}, \Sigma_{|3})$$

with conditional covariance

$$\Sigma_{|3} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} \sigma_{13} \\ \sigma_{23} \end{pmatrix} \frac{1}{\sigma_{33}} \begin{pmatrix} \sigma_{13} & \sigma_{23} \end{pmatrix}$$

Key observation:

$$X_1 \perp\!\!\!\perp X_2 \mid X_3 \quad \Longleftrightarrow \quad \mathrm{Cov}(X_1, X_2 | X_3) = (\Sigma_{|3})_{12} = \sigma_{12} - \frac{\sigma_{13}\sigma_{23}}{\sigma_{33}} = 0.$$

$\rightarrow$ Conditional independence can be rephrased as a polynomial constraint! We observe:

Conditional independence $\Longleftrightarrow$ vanishing partial correlation

We call $K = \Sigma^{-1}$ the precision or concentration matrix.

Conditional Covariance via the Precision Matrix

Let $X \sim \mathcal{N}_d(\mu, \Sigma)$ with precision matrix $K = \Sigma^{-1}$. Partition $X = (X_A, X_B)$. Then:

$$\mathrm{Cov}(X_A \mid X_B) = (K_{AA})^{-1}.$$

In particular, this means that

$$\mathrm{Cov}\left(\begin{pmatrix} X_i \\ X_j \end{pmatrix} \middle| X_{[d]\setminus\{i,j\}}\right) = \begin{pmatrix} K_{ii} & K_{ij} \\ K_{ij} & K_{jj} \end{pmatrix}^{-1}.$$

From Partial Correlation to the Precision Matrix

By Cramer's rule, we obtain

$$\mathrm{Cov}(X_i, X_j \mid X_{[d]\setminus\{i,j\}}) = -\frac{K_{ij}}{\det\begin{pmatrix} K_{ii} & K_{ij} \\ K_{ij} & K_{jj} \end{pmatrix}}.$$

Conditional Independence and Zeros in K

Let $K = \Sigma^{-1}$ be the precision matrix of a Gaussian vector X. Then:

$$K_{ij} = 0 \quad \Longleftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid X_{[d]\setminus\{i,j\}}.$$

Conditional independence (given all other variables) is encoded by zeros in Σ^{-1}

Undirected graphical models

Graph terminology

- Let $G = (V, E)$ be a graph with vertex set $V = [d] := \{1, \ldots, d\}$ and edge set $E \subset V \times V$, where $(v, v) \notin E$ for any $v \in V$ (no loops).
- When $(i, j) \in E$ and $(j, i) \in E$, the nodes i and j are connected by an undirected edge.
- When $(i, j) \in E$ and $(j, i) \notin E$, a directed edge $i \to j$ points from i to j.
- When all edges are undirected, we call G an undirected graph.
- We assume that there are no multiple edges.

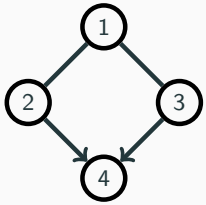


Figure 2: Graph with $V = [4]$ and $E = \{(1, 2), (2, 1), (1, 3), (3, 1), (2, 4), (3, 4)\}$

Graph terminology

- Let $G = (V, E)$ be a graph with vertex set $V = [d] := \{1, \ldots, d\}$ and edge set $E \subset V \times V$, where $(v, v) \notin E$ for any $v \in V$ (no loops).
- When $(i, j) \in E$ and $(j, i) \in E$, the nodes i and j are connected by an undirected edge.
- When $(i, j) \in E$ and $(j, i) \notin E$, a directed edge $i \to j$ points from i to j.
- When all edges are undirected, we call G an undirected graph.
- We assume that there are no multiple edges.

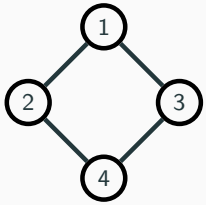


Figure 2: Undirected graph with $V = [4]$ and $E = \{(1, 2), (2, 1), \ldots, (3, 4), (4, 3)\}$

Undirected graph separation

- Let A, B, S be disjoint subsets of V.
- We say that S separates A from B when every path between A and B has a non-empty intersection with S.
- This is denoted as $A \perp_G B|S$.
- Example: The diamond graph on the right satisfies

$$\{1\} \perp_G \{4\}|\{2,3\},$$
$$\{2\} \perp_G \{3\}|\{1,4\}.$$

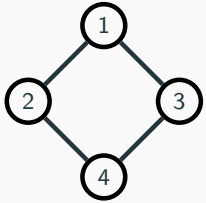


Figure 3: Undirected graph

Undirected graphical models

Definition

Let $G = (V, E)$ be an undirected graph and X a random vector. X satisfies

(P) the pairwise Markov property wrt G when

$$(i, j) \notin E \implies X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}},$$

(G) the global Markov property wrt G when

$$A \perp_G B | S \implies X_A \perp\!\!\!\perp X_B | X_S.$$

Proposition

It holds that (G) $\implies$ (P).

Theorem

When $\mathbb{P}_X$ has positive and continuous density on a product space, then

$$(G) \iff (P).$$

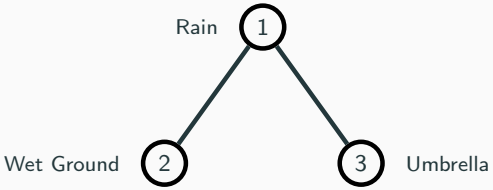
Example

Example

- $X_1 = \text{Rain}$
- $X_2 = \text{Wet Ground}$
- $X_3 = \text{Umbrella}$

$\Longrightarrow$ $X_2 \perp\!\!\!\perp X_3 \mid X_1$ says that given Rain, Wet Ground carries no information for Umbrella (and vice versa).

A corresponding graph that imposes $X_2 \perp\!\!\!\perp X_3 \mid X_1$ via the pairwise and global Markov properties:



Example

- Let G be the graph on the right.
- Example: This graph satisfies for example

$$\{1,2\} \perp_G \{4,5\}|\{3\},$$

but also

$$\{2\} \perp_G \{4,5\}|\{3\}, \ldots$$

- The pairwise statements are easier to list:

$$\{i \perp_G j|V \setminus \{i,j\},\ \forall ij \notin E\}.$$

$\rightarrow$The pairwise Markov property can be considered as "more practical".

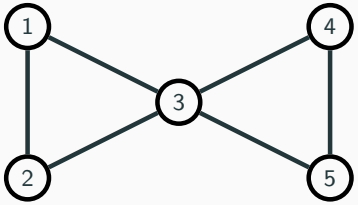


Figure 4: Butterfly graph

Hammersley–Clifford Theorem

Clique: A set of nodes $C \subset V$ such that all pairs in C are connected.

Density factorization via the Hammersley–Clifford theorem

Let $G = (V, E)$ with cliques $\mathcal{C}$ and X a random vector with positive density f. Then

$$X \text{ satisfies the global Markov property} \iff f(x) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C),$$

where $\psi_C(x_C)$ are nonnegative functions.

Remarks:

- In general, $\psi_C \neq f_C$ (not equal to the marginal density)
- If G is decomposable, then

$$f(x) = \frac{\prod_C f_C(x_C)}{\prod_S f_S(x_S)} \quad \text{(cliques } C\text{, separators } S\text{)}$$

Gaussian Graphical Models

Math grades example

- We consider examination grades of 88 students in 5 different mathematical subjects:

  Vectors, Analysis, Algebra, Mechanics and Statistics

  available in the dataset `mathmarks` from the `R` package `bnmonitor`.
- We compute the sample covariance matrix S: [1]

$$S = \begin{array}{c|ccccc} & \text{mech} & \text{vect} & \text{alg} & \text{ana} & \text{stat} \\ \hline \text{mech} & 305.77 & & & & \\ \text{vect} & 127.22 & 172.84 & & & \\ \text{alg} & 101.58 & 85.16 & 112.89 & & \\ \text{ana} & 106.27 & 94.67 & 112.11 & 220.38 & \\ \text{stat} & 117.40 & 99.01 & 121.87 & 155.54 & 297.76 \end{array}$$

[1] Note that the (sample) covariance matrix is symmetric, therefore, only its lower-triangular part is displayed for brevity.

Math grades example

We derive the sample precision matrix $\widetilde{K} = S^{-1}$:

$\widetilde{K} = 10^{-3} \times$

	mech	vect	alg	ana	stat
mech	5.24				
vect	−2.44	10.43			
alg	−2.74	−4.71	26.95		
ana	0.010.01	−0.79	−7.05	9.88	
stat	−0.14 − 0.14	−0.17 − 0.17	−4.70	−2.02	6.45

$\Rightarrow$ We observe that certain entries in the sample precision matrix are somewhat close to zero.

Gaussian model

- Let's model the data with a 5-variate Gaussian random vector $X \sim N_5(\mu, \Sigma)$.
- The small values in $\hat{K}$ may indicate conditional independence in the ground truth. What would be a reasonable model for this data?
- We saw before that $X_i \perp\!\!\!\perp X_j | X_{V \setminus ij} \iff K_{ij} = 0$. Can we use this for graphical modeling?

Let $X \sim N_d(\mu, \Sigma)$ be a d-variate Gaussian with precision matrix $K := \Sigma^{-1}$.

Definition

Let $G = (V, E)$ be an undirected graph. X is a Gaussian graphical model wrt G when

$$(i, j) \notin E \implies K_{ij} = 0.$$

Note that this is based on the pairwise Markov property.

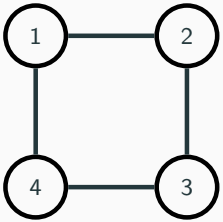


Figure 5: 4-cycle

Example:

$$K = \begin{pmatrix} K_{11} & K_{12} & 0 & K_{14} \\ K_{12} & K_{22} & K_{23} & 0 \\ 0 & K_{23} & K_{33} & K_{34} \\ K_{14} & 0 & K_{34} & K_{44} \end{pmatrix}$$

Rudimentary model selection

Problem

Even if the ground truth has zeros in K, we would not see them in a sample concentration matrix $\hat{K}$ with probability 1.

Let's consider every value in $\hat{K}$ that is smaller than 10^{-3} as evidence for sparsity:

$10^{-3} \times$

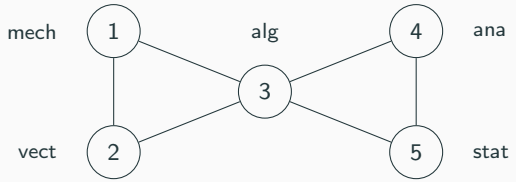
	mech	vect	alg	ana	stat
mech	5.24				
vect	−2.44	10.43			
alg	−2.74	−4.71	26.95		
ana	0.01	−0.79	−7.05	9.88	
stat	−0.14	−0.17	−4.70	−2.02	6.45

Graphical model for Math grades

We assume the ground truth precision matrix K has the form

	mech	vect	alg	ana	stat
mech	K_{11}				
vect	K_{12}	K_{22}			
alg	K_{13}	K_{23}	K_{33}		
ana	0	0	K_{34}	K_{44}	
stat	0	0	K_{35}	K_{45}	K_{55}

This is a Gaussian graphical model with 11 parameters, with respect to the butterfly graph:



Estimation in Gaussian graphical models

Estimation in Gaussian graphical models

- For a given Gaussian graphical model, we would like to estimate the precision matrix from data.
- Maximum likelihood estimation (MLE) is the most popular method for Gaussian graphical models.
- As a reminder, we first discuss maximum likelihood estimation for regular (non-graphical) multivariate Gaussians.

Reminder: Gaussian covariance MLE

Likelihood function for multivariate Gaussians

- Consider n i.i.d. centered (random) samples $X^{(1)}, \dots, X^{(n)} \sim \mathcal{N}(0, \Sigma)$[2]
- Our goal is to estimate the covariance matrix Σ or, equivalently, the precision matrix $K = \Sigma^{-1}$.

Gaussian Likelihood

The likelihood function for K is given by

$$L(K; X^{(1)}, \dots, X^{(n)}) = 2\pi^{-nd/2} \det(K)^{n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (X^{(i)})^\top K X^{(i)} \right).$$

[2] Note that the mean μ can be simply estimated by the sample mean and thus data can be centered around 0.

Gaussian log-likelihood

Up to constants, the log-likelihood function is then given by

$$\begin{aligned}\ell(K; X_1, \ldots, X_n) &= \log\det(K) - \frac{1}{n}\sum_{i=1}^{n}(X^{(i)})^\top K X^{(i)} \\ &= \log\det(K) - \mathrm{tr}(SK),\end{aligned}$$

where $S = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}(X^{(i)})^\top$ is the sample covariance matrix.

MLE for the multivariate Gaussian[3]

$$\hat{K} = \underset{K \succ 0}{\operatorname{argmax}} \quad \log\det(K) - \mathrm{tr}\,(SK)$$

[3] Here, $K \succ 0$ means that K is symmetric positive definite

Maximum likelihood estimator (MLE)

MLE for the multivariate Gaussian[4]

$$\hat{K} = \underset{K \succ 0}{\operatorname{argmax}} \quad \log\det(K) - \operatorname{tr} S \cdot K$$

First-order optimality condition

The score equation is

$$\nabla_K \ell(K; S) = K^{-1} - S.$$

Setting the gradient to zero yields

$$K^{-1} = S.$$

One obtains the MLE of the precision matrix as

$$\hat{K} = S^{-1}.$$

[4] Here, $K \succ 0$ means that K is symmetric positive definite

MLE for the covariance in Gaussian graphical models

MLE for the Gaussian graphical model

Assume a Gaussian graphical model with respect to some undirected graph $G = (V, E)$.

MLE for the Gaussian graphical model

$$\hat{K} = \underset{K \succ 0}{\operatorname{argmax}} \quad \log\det(K) - \operatorname{tr}(SK) \text{ s.t. } K_{ij} = 0 \ \ \forall (i,j) \notin E. \tag{1}$$

- The solution of (1) is the solution to the following matrix completion problem:

Matrix completion problem

Given a graph $G = (V, E)$ and a symmetric positive definite matrix S, find a symmetric positive definite matrix K with inverse $\Sigma := K^{-1}$ such that:

$$K_{ij} = 0, \quad (i,j) \notin E \cup \{(1,1), \ldots, (d,d)\},$$
$$\Sigma_{ij} = S_{ij}, \quad (i,j) \in E \cup \{(1,1), \ldots, (d,d)\}.$$

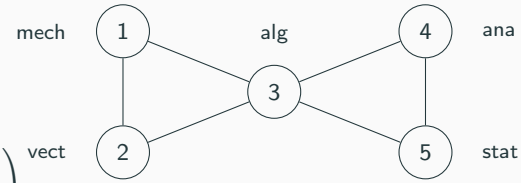
MLE for Math grades example

We obtain the MLE

$$\widehat{\Sigma} = \begin{pmatrix} 305.77 & 127.22 & 101.58 & 100.88 & 109.66 \\ 127.22 & 172.84 & 85.16 & 84.57 & 91.93 \\ 101.58 & 85.16 & 112.89 & 112.11 & 121.87 \\ 100.88 & 84.57 & 112.11 & 220.38 & 155.54 \\ 109.66 & 91.93 & 121.87 & 155.54 & 297.76 \end{pmatrix}$$

and

$$\widehat{K} = 10^{-3} * \begin{pmatrix} 5.24 & -2.44 & -2.87 & 0 & 0 \\ -2.44 & 10.35 & -5.61 & 0 & 0 \\ -2.87 & -5.61 & 28.49 & -7.55 & -4.93 \\ 0 & 0 & -7.55 & 9.82 & -2.04 \\ 0 & 0 & -4.93 & -2.04 & 6.44 \end{pmatrix}$$



Gaussian graphical models and linear regression

Suppose

$$X \sim \mathcal{N}_d(0, \Sigma), \qquad K = \Sigma^{-1}.$$

Regression view of a GGM:

For each variable X_j, regress it on all remaining variables:

$$X_j = \sum_{k \neq j} \beta_{jk} X_k + \varepsilon_j.$$

In the Gaussian case, these coefficients are determined by the precision matrix:

$$\beta_{jk} = -\frac{K_{jk}}{K_{jj}}.$$

- $K_{jk} \neq 0 \Rightarrow X_k$ contributes to predicting X_j after adjusting for all other variables.
- $K_{jk} = 0 \Rightarrow$ the coefficient of X_k is zero in the population regression.
- See `R` code for Math grades example.

Structure learning via the graphical lasso

Why do we need regularization?

- In practice, we often face high-dimensional settings (d large, n small).
- The sample covariance S will be singular if $n < d$.
- Then the regular MLE
$$\hat{K} = S^{-1}$$
does not exist.
- Even when S^{-1} exists:
  - it is typically dense,
  - small entries are difficult to interpret,
  - poor recovery of the true sparse graph.

Idea

Encourage sparsity in K directly during estimation.

ℓ_1-penalized Gaussian likelihood

The graphical lasso estimates K by solving

$$\hat{K} = \underset{K \succ 0}{\operatorname{argmax}} \quad \log\det(K) - \operatorname{tr}(SK) \; - \; \lambda \sum_{i \neq j} |K_{ij}|.$$

- The penalty term
$$\sum_{i \neq j} |K_{ij}|$$
promotes sparsity in the precision matrix.
- $\lambda \geq 0$ is a tuning parameter:
  - $\lambda = 0$: recovers the MLE S^{-1} (if it exists),
  - large λ: more zeros in $\hat{K}$.

Sparsistency of the graphical lasso

Question

When does the graphical lasso recover the correct sparsity pattern of K?

Sparsistency

Under suitable assumptions:

- the true graph is sparse,
- the nonzero entries of K are not too small,
- an incoherence / irrepresentability condition holds,
- and n is sufficiently large relative to the graph complexity,

then, with high probability,

$$\text{sign}(\hat{K}_{ij}) = \text{sign}(K_{ij}) \qquad \text{for all } i \neq j.$$

In particular,

$$\hat{K}_{ij} = 0 \quad \Longleftrightarrow \quad K_{ij} = 0,$$

so the estimated graph equals the true graph.

Estimation under positivity constraints

Example

Let's take another look at the math grades example. The sample covariance and precision matrices were

$S =$

	mech	vect	alg	ana	stat
mech	305.77				
vect	127.22	172.84			
alg	101.58	85.16	112.89		
ana	106.27	94.67	112.11	220.38	
stat	117.40	99.01	121.87	155.54	297.76

$\widetilde{K} = 10^{-3} \times$

	mech	vect	alg	ana	stat
mech	5.24				
vect	−2.44	10.43			
alg	−2.74	−4.71	26.95		
ana	0.01	−0.79	−7.05	9.88	
stat	−0.14	−0.17	−4.70	−2.02	6.45

Reminder

$\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$ and $\mathrm{Cov}(X_i, X_j \mid X_{[d]\setminus\{i,j\}}) = -K_{ij}/\det\begin{pmatrix} K_{ii} & K_{ij} \\ K_{ij} & K_{jj} \end{pmatrix}$. What do you observe?

Example for total positivity

- We observe that
  - all covariances are positive and
  - all conditional covariances are positive or close to zero.
- If you would randomly generate a positive definite matrix (or a Gaussian graphical model), would you expect such a structure?
- There seems to be some intrinsic positive dependence
  $\Rightarrow$ Multivariate total positivity of order two (MTP_2).

Definition

A probability density f on $\mathbb{R}^d$ is MTP_2 if

$$f(x \wedge y)\, f(x \vee y) \;\geq\; f(x)\, f(y), \qquad \forall x, y \in \mathbb{R}^d,$$

where $(x \wedge y)_i = \min(x_i, y_i)$ and $(x \vee y)_i = \max(x_i, y_i)$.

MTP_2 for Gaussians

Gaussian MTP_2

If $X \sim \mathcal{N}_d(\mu, \Sigma)$ with precision matrix $K = \Sigma^{-1}$, then

$$X \text{ is } \text{MTP}_2 \quad \iff \quad K_{ij} \leq 0 \text{ for all } i \neq j.$$

- Thus, in a Gaussian model, MTP_2 means that K is a positive definite M-matrix.
- Equivalently, all conditional covariances are nonnegative.

Gaussian MLE under MTP_2

Given the sample covariance matrix S, estimate K by

$$\hat{K} = \underset{K \succ 0}{\operatorname{argmax}} \quad \log\det(K) - \operatorname{tr}(SK) \quad \text{s.t.} \quad K_{ij} \leq 0, \qquad \forall i \neq j.$$

MTP_2 acts as an implicit regularizer, the MLE exists for sample size 2!

The Golazo constraint

The Golazo constraint

MTP_2, the graphical lasso, and other penalized/ constrained methods can be generalized via the Golazo penalty:

$$\|K\|_{L,U} = \sum_{i \neq j} \max\{L_{ij}K_{ij},\ U_{ij}K_{ij}\}$$

Here, $L, U \in (\mathbb{R} \cup \{-\infty, \infty\})^{p \times p}$ with $L_{ij} \leq 0 \leq U_{ij}$ for all i, j and $\operatorname{diag}(L) = \operatorname{diag}(U) = 0$.

Optimization problem:

$$\hat{K} = \arg\max_{K \succeq 0} \left\{ \ell(K; S) - \|K\|_{L,U} \right\}$$

Examples:

- **Graphical lasso:** $L_{ij} = -\lambda, \quad U_{ij} = \lambda \ \Rightarrow\ \|K\|_{L,U} = \lambda \sum_{i \neq j} |K_{ij}|$
- **MTP_2:** $L_{ij} = 0, \quad U_{ij} = \infty \ \Rightarrow\ K_{ij} \leq 0$
- **Asymmetric / adaptive penalties:** $L_{ij} = \ell_{ij} < 0, \quad U_{ij} = u_{ij} > 0$

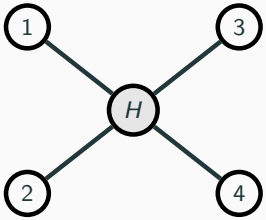
Latent variables in undirected Gaussian models

Why do latent variables matter?

- In many applications, not all relevant variables are observed.
- Examples:
  - an unobserved common factor,
  - missing measurements.
- If we ignore latent variables, the observed graph can look more dense than the true underlying structure.

Key idea

A latent variable can induce dependence between observed variables, even if these observed variables would be conditionally independent given the latent one.



Hidden common factor

Latent variables in a Gaussian graphical model

Suppose the full random vector X is Gaussian and splits into observed and hidden variables:

$$X = (X_O^\top, X_H^\top)^\top \sim \mathcal{N}_{d+h}(0, \Sigma), \qquad K = \Sigma^{-1}.$$

Write the full precision matrix as

$$K = \begin{pmatrix} K_{OO} & K_{OH} \\ K_{HO} & K_{HH} \end{pmatrix}.$$

Marginal model for the observed variables

The observed vector X_O is again Gaussian:

$$X_O \sim \mathcal{N}_d(0, \Sigma_{OO}).$$

Its precision matrix is

$$\Sigma_{OO}^{-1} = K_{OO} - K_{OH} K_{HH}^{-1} K_{HO}.$$

- K_{OO} describes the conditional graph among the observed variables, given the hidden ones.
- The correction term

  $$K_{OH} K_{HH}^{-1} K_{HO}$$

  comes from marginalizing over the latent variables.

Sparse + low-rank decomposition

Define

$$K^{\text{obs}} := \Sigma_{OO}^{-1}.$$

From the previous slide,

$$K^{\text{obs}} = K_{OO} - K_{OH} K_{HH}^{-1} K_{HO}.$$

Sparse + low-rank form

If the conditional graph among the observed variables is sparse, then K_{OO} is sparse. If there are only few hidden variables, then

$$L := K_{OH} K_{HH}^{-1} K_{HO}$$

has low rank, with

$$L \succeq 0, \qquad \text{rk}(L) \leq h.$$

Hence

$$K^{\text{obs}} = K^{\text{sp}} - L,$$

where $K^{\text{sp}} := K_{OO}$ is sparse and L is low-rank.

- Sparse part K^{sp}: conditional relationships among observed variables.
- Low-rank part L: effect of a few latent factors.

Suppose we only observe $X_O^{(1)}, \ldots, X_O^{(n)}$ and compute the sample covariance matrix S.

Latent-variable graphical model estimator

A popular approach estimates a sparse part and a low-rank part jointly:

$$(\hat{K}^{\text{sp}}, \hat{L}) = \underset{K^{\text{sp}} - L \succ 0,\ L \succeq 0}{\operatorname{argmax}} \Big\{ \log\det(K^{\text{sp}} - L) - \operatorname{tr}\left(S(K^{\text{sp}} - L)\right) - \lambda \sum_{i \neq j} \left|K_{ij}^{\text{sp}}\right| - \gamma \operatorname{tr}(L) \Big\}.$$

- The ℓ_1 penalty promotes sparsity in K^{sp}.
- The trace penalty $\operatorname{tr}(L)$ promotes low rank because $L \succeq 0$.
- The estimate of the observed precision matrix is

$$\hat{K}^{\text{obs}} = \hat{K}^{\text{sp}} - \hat{L}.$$

Summary Part 1: Undirected graphical models

Summary Part 1: Undirected graphical models

- **Graphical models:** $(i,j) \notin E \implies X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$
- **Gaussian case:** $K = \Sigma^{-1}, \quad K_{ij} = 0 \iff$ conditional independence
- **Parameter estimation:**
  - Given a (Gaussian) graphical model, estimate the parameters.
  - MLE: $\hat{K} = S^{-1}$ (if $n \geq d$).
- **Structure learning:**
  - Learn the graphical model from observational data.
  - Sparsity $\Rightarrow$ interpretable graphs.
  - High-dimensional consistency (sparsistency).

Summary Part 1: Undirected graphical models

- **Further topics:**
  - MTP_2 provides joint parameter and structure learning, under strong assumptions.
  - Golazo allows a unified formulation for many structure learning methods.
  - Latent variables: sparse + low-rank decomposition
- **Things we did not discuss:**
  - Restriction to particular graph structures (trees, decomposable graphs,...)
  - Popular discrete models (e.g., Ising models)
  - Alternative structure learning approaches (e.g., score matching)
  - Algebraic statistics of graphical models

References

Drton, M. and M. H. Maathuis (2017).
Structure learning in graphical modeling.
Annual Review of Statistics and Its Application 4(1), 365–393.

Lauritzen, S. L. (1996).
***Graphical models*, Volume 17 of *Oxford Statistical Science Series*.**
The Clarendon Press, Oxford University Press, New York.
Oxford Science Publications.

Messerli, F. H. (2012).
Chocolate consumption, cognitive function, and nobel laureates.
New England Journal of Medicine 367(16), 1562–1564.

Peters, J., D. Janzing, and B. Schölkopf (2017).
***Elements of causal inference: foundations and learning algorithms*.**
The MIT Press.

Thank You!