Graphical models and total positivity

Graphical models, causality, and positive dependence

Piotr Zwiernik

Universitat Pompeu Fabra

December 2-3, 2019

Outline

Introduce basic concepts of total positivity. Three parts:

1. Total positivity and Markov structures.
   Total positivity in Markov structures (with S. Fallat, S. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth), Ann. Stat., 2017.
2. Gaussian graphical models.
   Maximum likelihood estimation in Gaussian models under total positivity (with S. Lauritzen, C. Uhler), Ann. Stat., 2019.
3. Binary models and beyond.
   Total positivity in structured binary distributions (with S. Lauritzen, C. Uhler), arXiv:1905.00516.

Lecture 1

Basics

Definition

- $X = (X_1, \ldots, X_m)$, $\mathcal{X} = \prod_{i=1}^{m} \mathcal{X}_i \subset \mathbb{R}^m$, density p.

- A function p is MTP_2 if:

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

  - e.g. $p(1,1,0)p(0,0,1) \leq p(0,0,0)p(1,1,1)$

- If $p > 0$ the condition simplifies.

  - e.g. $p(1,1,0)p(0,1,1) \leq p(0,1,0)p(1,1,1)$

Original motivation

- $X = (X_1, \ldots, X_m)$ is positively associated if for any two *non-decreasing* functions $\phi, \psi : \mathbb{R}^m \to \mathbb{R}$

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

Theorem[FKG inequality]

$$\mathrm{MTP}_2 \quad \Longrightarrow \quad \text{positively associated.}$$

Proof: Discrete case by Fortuin et al. (1971). General case by Sarkar (1969).

Basic properties

If $X = (X_1, \ldots, X_m)$ is MTP_2, then

(i) any **marginal** distribution is MTP_2;

(ii) any **conditional** distribution is MTP_2,

for details see (Karlin and Rinott, 1980).

Our motivation

- Useful concept in data modelling.
- Some popular models are implicitly MTP_2.
- Leads to sparsity, applies in high-dimensions.

Elementary examples

Some classical examples

Mostly from Karlin and Rinott (1980):

- Eigenvalues of a Wishart matrix W, or of $W_1 W_2^{-1}$, or $W_1(W_1 + W_2)^{-1}$, where $W_1 \perp\!\!\!\perp W_2$ (Dykstra and Hewett, 1978);
- Ferromagnetic (attractive) Ising models (Lebowitz, 1972);
- Bivariate logistic density (Gumbel, 1961);
- Gaussian free fields (random height landscapes) (Dynkin, 1980);
- Many other examples...

Example 1: Three binary variables

Example: $X = (X_1, X_2, X_3) \in \{0,1\}^3$

$$\begin{array}{lll} p_{001}p_{110} \le p_{000}p_{111} & p_{010}p_{101} \le p_{000}p_{111} & p_{100}p_{011} \le p_{000}p_{111} \\ p_{011}p_{101} \le p_{001}p_{111} & p_{011}p_{110} \le p_{010}p_{111} & p_{101}p_{110} \le p_{100}p_{111} \\ p_{001}p_{010} \le p_{000}p_{011} & p_{001}p_{100} \le p_{000}p_{101} & p_{010}p_{100} \le p_{000}p_{110} \end{array}$$

Note: If $p > 0$ then the first row is implied by the other two:

$$(p_{011}p_{101})(p_{001}p_{010}) \le (p_{001}p_{111})(p_{000}p_{011})$$

Boundary points satisfy context specific independence:

$$p_{011}p_{101} = p_{001}p_{111} \qquad \Longleftrightarrow \qquad X_1 \perp\!\!\!\perp X_2 | \{X_3 = 1\}$$

There are various useful reformulations:

All conditional covariances are nonnegative:

$$p_{01k}p_{10k} \leq p_{00k}p_{11k} \quad \iff \quad \mathrm{cov}(X_1, X_2 | \{X_3 = k\}) \geq 0.$$

Equivalently all conditional log-odds ratios are nonnegative:

$$p_{01k}p_{10k} \leq p_{00k}p_{11k} \quad \iff \quad \log\left(\frac{p_{00k}p_{11k}}{p_{01k}p_{10k}}\right) \geq 0.$$

Equivalently, $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3 | H, \quad H$ binary, $\mathrm{cov}(X_i, H) \geq 0$.

For details see Zwiernik (2015).

Example 2: Gaussian distribution

PD_m = symmetric $m \times m$ positive definite matrices

Gaussian distribution with mean $\mu \in \mathbb{R}^m$ and covariance $\Sigma \in \mathrm{PD}_m$

concentration matrix $K := \Sigma^{-1}$

$$p(x; K) \;=\; \frac{1}{(2\pi)^{m/2}}(\det K)^{1/2} \exp\{-\tfrac{1}{2}(x-\mu)^T K (x-\mu)\}$$

Gaussian X is MTP_2 if and only if $K_{ij} \leq 0$ for all $i \neq j$.

- Equivalently, K is an M-matrix (aka Stieltjes matrix).

See Bølviken (1982) and Karlin and Rinott (1983).

Recall: The partial correlations satisfy

$$\rho_{ij|V\setminus\{i,j\}} \;=\; -\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}} \;\geq\; 0.$$

For Gaussians partial and conditional correlations are equal.

Closure of the MTP_2 property under marginalization gives:

$$\mathrm{cov}(X_i, X_j | X_C) \;\geq\; 0 \qquad \text{for every } C \subseteq V \setminus \{i,j\}.$$

The set of M-matrices is convex (in K). Its boundary is given by some $K_{ij} = 0$ (or equivalently $X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}$).

Binary Ising model

The p.m.f. for $x \in \mathcal{X} = \{-1, 1\}^m$ satisfies

$$\log p(x; h, J) \;=\; h^T x \;+\; \tfrac{1}{2} x^T J x \;-\; A(h, J),$$

with $h \in \mathbb{R}^m$ and $J \in \mathbb{R}^{m \times m}$ symmetric with zeros on the diagonal.

(a log-linear model with only second order interactions)

Binary Ising model is MTP_2 if and only if $J_{ij} \geq 0$ for all $i \neq j$.

Similar to the Gaussian case, the hypothesis is convex (in J) and:

$$J_{ij} = 0 \quad \Longleftrightarrow \quad X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}}.$$

Modelling

A restrictive condition?

- MTP_2 contraints appear to be *restrictive*:
  - 3-dim Gaussian: about 5% of distributions are MTP_2,
  - 4-dim Gaussian: about 0.09% of distributions are MTP_2.
- Less restrictive under additional Markov structure.
- In the 3-dim case
  - if $1 \perp\!\!\!\perp 2|3$ then 25% are MTP_2
  - if, in addition, $1 \perp\!\!\!\perp 3|2$ then 50% are MTP_2
  - if $1 \perp\!\!\!\perp 2 \perp\!\!\!\perp 3$ then everything is MTP_2

Informally: In sparse structures MTP_2 is more likely.

(*) It is more likely to see MTP_2 distribution in sparse structures especially in settings where total positivity is expected.

Example 1: EPH-gestosis

- Dataset collected 45 years ago in a study on "Pregnancy and Child Development"
- EPH-gestosis (pre-eclampsia): disease **syndrome** for pregnant women; three **symptoms** (high body water retention, high amounts of urinary proteins, elevated blood pressure)
- A syndrome is a set of medical signs and symptoms that are correlated with each other and, often, with a particular disease or disorder.

The sample distribution

$$\begin{bmatrix} \hat{p}_{000} & \hat{p}_{010} & \hat{p}_{001} & \hat{p}_{011} \\ \hat{p}_{100} & \hat{p}_{110} & \hat{p}_{101} & \hat{p}_{111} \end{bmatrix} = \frac{1}{4649} \begin{bmatrix} 3299 & 107 & 1012 & 58 \\ 78 & 11 & 65 & 19 \end{bmatrix}$$

is already MTP_2. Equivalently, $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3 | H$ for a latent binary H.

Example 2: Financial time series.

Monthly correlations of global stock markets

$$S = \begin{array}{c} \textit{Nasdaq} \\ \textit{Canada} \\ \textit{Europe} \\ \textit{UK} \\ \textit{Australia} \end{array} \overset{\begin{array}{ccccc} \textit{Nasdaq} & \textit{Canada} & \textit{Europe} & \textit{UK} & \textit{Australia} \end{array}}{\begin{pmatrix} 1.000 & 0.606 & 0.731 & 0.618 & 0.613 \\ 0.606 & 1.000 & 0.550 & 0.661 & 0.598 \\ 0.731 & 0.550 & 1.000 & 0.644 & 0.569 \\ 0.618 & 0.661 & 0.644 & 1.000 & 0.615 \\ 0.613 & 0.598 & 0.569 & 0.615 & 1.000 \end{pmatrix}}$$

$$S^{-1} = \begin{array}{c} \textit{Nasdaq} \\ \textit{Canada} \\ \textit{Europe} \\ \textit{UK} \\ \textit{Australia} \end{array} \overset{\begin{array}{ccccc} \textit{Nasdaq} & \textit{Canada} & \textit{Europe} & \textit{UK} & \textit{Australia} \end{array}}{\begin{pmatrix} 2.629 & -0.480 & -1.249 & -0.202 & -0.490 \\ -0.480 & 2.109 & -0.039 & -0.790 & -0.459 \\ -1.249 & -0.039 & 2.491 & -0.675 & -0.213 \\ -0.202 & -0.790 & -0.675 & 2.378 & -0.482 \\ -0.490 & -0.459 & -0.213 & -0.482 & 1.992 \end{pmatrix}}$$

Sampled uniformly this happens with prob. $< 10^{-6}$!

Example 3: Math grades

Data: grades of 88 students in `Mechanics`, `Vectors`, `Algebra`, `Analysis`, `Statistics` (`data(math)` in package gRbase)

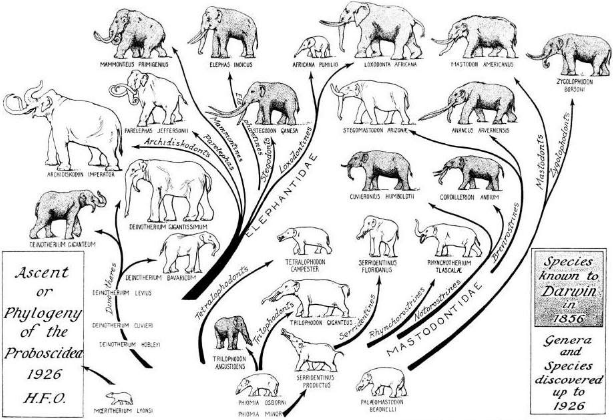
$S =$

	mechanics	vectors	algebra	analysis	statistics
mechanics	305.7680	127.2226	101.5794	106.2727	117.4049
vectors	127.2226	172.8422	85.1573	94.6729	99.0120
algebra	101.5794	85.1573	112.8860	112.1134	121.8706
analysis	106.2727	94.6729	112.1134	220.3804	155.5355
statistics	117.4049	99.0120	121.8706	155.5355	297.7554

$S^{-1} =$

	mechanics	vectors	algebra	analysis	statistics
mechanics	1	−0.329	−0.230	0.002	−0.025
vectors	−0.329	1	−0.281	−0.078	−0.020
algebra	−0.230	−0.281	1	−0.432	−0.357
analysis	0.002	−0.078	−0.432	1	−0.253
statistics	−0.025	−0.020	−0.357	−0.253	1

MTP_2 constraints are often explicit



Ascent phyla of Elephants and Mastodonts. After Osborn in the year 1926

X is MTP_2 in:

- ferromagnetic Ising models
- Markov chains with TP_2 transitions
- order statistics of *iid* variables
- Brownian motion tree model

$|X|$ is MTP_2 in (c.f. (Zwiernik, 2015)):

- Gaussian/binary tree models
- Gaussian/binary latent tree models
- binary latent class models
- single factor analysis

Signed MTP_2 distributions

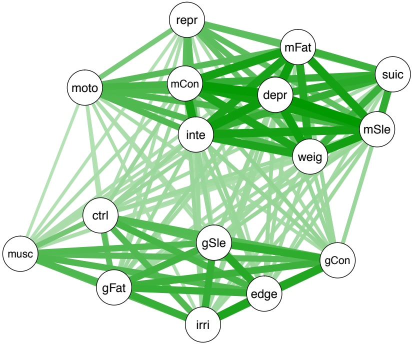
Definition: A Gaussian/discrete r.v. $X = (X_1, \dots, X_m)$ has a signed MTP_2 distribution if and only if:

- (Gaussian) there exists a diagonal matrix $D \in \{-1, +1\}^m$ such that DX has an MTP_2 distribution.
- (discrete) the distribution of X is MTP_2 up to a permutation of values in each $\mathcal{X}_i$

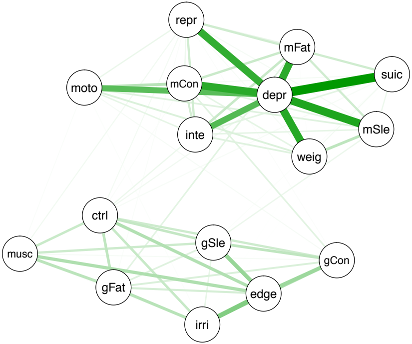
- Every binary/Gaussian pairwise interaction model on a tree is signed MTP_2.
- Signed MTP_2 property is preserved under taking margins:
  - single factor analysis models, Gaussian latent tree models, binary latent class models, and binary latent tree models are all signed MTP_2

Sparsity with no extra parameters

Correlation network



MTP_2 Ising model



Independence structure

Marginal independence

Proposition: If X is positively associated then

$$X_A \perp\!\!\!\perp X_B \qquad \Longleftrightarrow \qquad \mathrm{cov}(X_u, X_v) = 0 \text{ for all } u \in A, v \in B.$$

Proof: Shown in Lebowitz (1972).

Such a result is usually special for the Gaussian distribution.

Proposition Fallat et al. (2017): X can be split into independent blocks and within each block all covariances are strictly positive.

Abstract conditional independence

An independence model $\perp$ is a ternary relation over subsets of V.

It is semi-graphoid if for disjoint subsets A, B, C, D:

(S1) if $A \perp B|C$ then $B \perp A|C$ (symmetry);

(S2) if $A \perp (B \cup D)|C$ then $A \perp B|C$ and $A \perp D|C$ (decomposition);

(S3) if $A \perp (B \cup C)|D$ then $A \perp B|(C \cup D)$ (weak union);

(S4) if $A \perp B|C$ and $A \perp D|(B \cup C)$, then $A \perp (B \cup D)|C$ (contraction).

(Any probabilistic independence model $\perp\!\!\!\perp$ is a semi-graphoid)

It is a graphoid if (S1)–(S4) holds and

(S5) if $A \perp B|(C \cup D)$ and $A \perp C|(B \cup D)$ then $A \perp (B \cup C)|D$ (intersection).

(If X has a density $f > 0$ its independence model $\perp\!\!\!\perp$ is a graphoid.)

Conditional independence and total positivity

Proposition (Fallat et al., 2017): If X is MTP_2, its independence model $\perp\!\!\!\perp$ satisfies

(S6) $(A \perp\!\!\!\perp B \mid C) \wedge (A \perp\!\!\!\perp D \mid C) \;\Rightarrow\; A \perp\!\!\!\perp (B \cup D) \mid C$ (composition);

(S7) $(u \perp\!\!\!\perp v \mid C) \wedge (u \perp\!\!\!\perp v \mid (C \cup w)) \;\Rightarrow\; (u \perp\!\!\!\perp w \mid C) \vee (v \perp\!\!\!\perp w \mid C)$ (singleton transitivity)

(S8) $(A \perp\!\!\!\perp B \mid C) \;\Rightarrow\; A \perp\!\!\!\perp B \mid (C \cup D)$ (upward stability).

These are all fulfilled for separation $\perp_G$ in graphs, but not necessarily for any probabilistic independence model $\perp\!\!\!\perp$.

Upward stability is a strong property; see (Sadeghi, 2017) for a follow-up.

Independence graph and Markov properties

Let P be a probability distribution on $\mathcal{X}$. The pairwise independence graph $\mathcal{G}(P) = (V, E)$ is defined through the relation

$$uv \notin E \iff u \perp\!\!\!\perp v | V \setminus \{u, v\}.$$

We say that P is globally Markov w.r.t. a graph G if

$$A \perp_G B | S \implies A \perp\!\!\!\perp B | S$$

where $\perp_G$ is separation in G. (c.f. Hammersley-Clifford theorem)

Further, we say that P is faithful to G if

$$A \perp_G B | S \iff A \perp\!\!\!\perp B | S$$

i.e. if the independence models $\perp\!\!\!\perp$ and $\perp_G$ are identical.

A main result

Theorem (Fallat et al., 2017): Assume the distribution P of X is MTP_2 with strictly positive density $f > 0$. Then P is faithful to $\mathcal{G}(P)$.

In other words, for MTP_2 distributions, the pairwise independence graph yields a complete 'picture' of the independence relations in P.

Lecture 2

Gaussian graphical models

see e.g. (Lauritzen, 1996; Højsgaard et al., 2012)

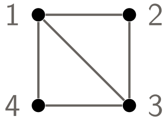
Factorization

G = an undirected graph with nodes $\{1, \ldots, m\}$ and cliques $C_1, \ldots, C_k$. We say that density $f(\boldsymbol{x})$ **factorizes according to G** if for all $\boldsymbol{x} \in \mathcal{X}$

$$f(\boldsymbol{x}) = \phi_{C_1}(\boldsymbol{x}_{C_1}) \cdots \phi_{C_k}(\boldsymbol{x}_{C_k}),$$

where $\phi_C(\boldsymbol{x}_C) \geq 0$. (a notion of simplicity)

For example



$$f(\boldsymbol{x}) = \phi_{123}(x_1, x_2, x_3)\phi_{134}(x_1, x_3, x_4).$$

This gives an alternative characterisation of $X_2 \perp\!\!\!\perp X_4 | (X_1, X_2)$.

Hammersley-Clifford theorem

Let $f > 0$ be a density function for $\boldsymbol{X} = (X_1, \ldots, X_m)$. Then the following are equivalent:

(F) f factorizes according to $G = (V, E)$.

(P) $X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}}$ if $ij \notin E$.

(G) $X_A \perp\!\!\!\perp X_B | X_C$ whenever C separates A and B in G.

If $f > 0$ then P is globally Markov to its pairwise independence graph.

$\mathcal{M}(G)$ = all distributions that factorize according to G.

The Gaussian case

For a Gaussian distribution in $\mathcal{M}(G)$:

The non-edges correspond to conditional independences $X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}$ or equivalently $K_{ij} = 0$.

- Indeed, $\rho_{ij|V\setminus\{i,j\}} = -\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$.

Two main estimation problems (Lauritzen, 1996):

Consider an *iid* sample $X^1, \ldots, X^n$ from $\mathcal{M}(G)$.

The partial correlation of the sample will have no zeros.

(i) Estimate Σ for a fixed graph G.

(ii) Estimate the graph in a statistically meaningful way.

The Gaussian likelihood function

The sample covariance matrix of the sample $X^1, \ldots, X^n$ is

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

The log-likelihood is

$$\log L(\mu, K) \;=\; \tfrac{n}{2}\log\det K - \tfrac{n}{2}\operatorname{tr}(KS) - \tfrac{n}{2}(\bar{X}-\mu)^T K(\bar{X}-\mu).$$

For fixed K we get $\hat{\mu} = \bar{X}$ giving the profile likelihood

$$\log L(\hat{\mu}, K) \;=\; \tfrac{n}{2}\log\det K - \tfrac{n}{2}\operatorname{tr}(KS).$$

Maximum likelihood over $\mathcal{M}(G)$

Fix a graph $G = (V, E)$ and the Gaussian model

$$\mathcal{M}(G) = \{K \in \mathrm{PD}_m : K_{ij} = 0 \text{ for all } ij \notin E\}.$$

The function $\log L(\hat{\mu}, K) = \frac{n}{2} \log \det K - \frac{n}{2} \operatorname{tr}(SK)$ is a concave function over the convex set $\mathcal{M}(G)$.

The MLE (if exists) is the unique point $\hat{K} = \hat{\Sigma}^{-1} \in \mathrm{PD}_m$ such that:

(i) $\hat{\Sigma}_{ij} = S_{ij}$ for all $ij \in E$,

(ii) $\hat{K}_{ij} = 0$ for all $ij \notin E$.

The MLE exists (with probability one) if $n \geq \max_C |C|$.

A block-coordinate descent approach is typically used, e.g. `ggmfit` in R.

Model selection methods

Main methods for learning the graph:

- Stepwise methods,
- Convex optimization,
- Thresholding,
- Simultaneous *p*-values.

Stepwise backward model selection

The `stepwise` function in `gRim` performs stepwise model selection based on a variety of criteria (AIC, BIC, etc)

```
sat.carc <- cmod(~.^.,data=carcass)
test.carc <- stepwise(sat.carc,details=1,"test")
plot(test.carc,"neatto")
```

Learning the graph in high-dimension

Graphical lasso (Friedman et al., 2008)

If p is large then the number of possible models is too high.

If $n < p$ the likelihood is unbounded.

Following the same idea as in the lasso regression we maximize

$$L_{\text{pen}}(K, \hat{\mu}) = \log\det(K) - \text{tr}(SK) - \lambda \|K\|_1.$$

See the package `glasso` and `EBICglasso` (finds an optimal λ).
Conveniently implemented in the package `qgraph`.

```
# S sample correlation
qgraph::qgraph(S,graph="glasso")
```

Sparsistency: if $\min_{ij \in E^*} |K_{ij}| \geq C\sqrt{\frac{\log(p)}{n}}$ for some $C > 0$.

Gaussian totally positive distributions

MLE for Gaussian models

Convex problem: $\text{maximize}\{\log\det K - \text{tr}(SK)\}$ over $K \in \text{MTP}_2$.

Theorem: The MLE exists if and only if there exists $\Sigma \succ 0$ with $\Sigma \geq S$. It is then equal to the unique element $\hat{K} = \hat{\Sigma}^{-1} \in \text{PD}_m$ that satisfies the following system of equations and inequalities

- Primal feasibility: $\hat{K}_{uv} \leq 0 \;\; \forall u \neq v$,
- Dual feasibility: $\hat{\Sigma}_{vv} - S_{vv} = 0 \;\; \forall v, \quad \hat{\Sigma}_{uv} - S_{uv} \geq 0 \;\; \forall u \neq v$
- Complimentary slackness: $(\hat{\Sigma}_{uv} - S_{uv})\,\hat{K}_{uv} = 0 \;\; \forall u \neq v$.

There are three different algorithms to find the MLE see Slawski and Hein (2015); Lauritzen et al. (2019b).

Existence of the MLE

Learning sparse structures in high dimensions was the main motivation of (Slawski and Hein, 2015): *Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields*.

Theorem (Slawski and Hein, 2015): The MLE exists with *probability one* whenever $n \geq 2$.

- our proof gives an explicit point that is both primal and dual feasible, it establishes links to Brownian motion tree models, single-linkage clustering, and ultrametrics.

The fact that a unique MLE exists for small samples suggests that the MTP_2 property adds considerable regularization for covariance matrix estimation.

Single-linkage matrix

Let S be the sample correlation matrix. Assume $S_{ij} \geq 0$ for all $i \neq j$.

Given the weighted graph of S, the single-linkage matrix Z is

$$Z_{ij} \;=\; \max_{P \in \mathcal{P}(i,j)} \min_{uv \in P} S_{uv} \qquad \text{for } i \neq j.$$

Z is both primarily and dually feasible.

The fact that $Z \geq S$ (dual feasibility) is easy to establish.

The fact that Z^{-1} is an M-matrix (primal feasibility) uses the connection to ultrametric matrices (Dellacherie et al., 2014):

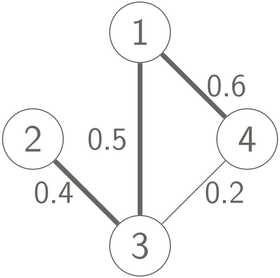
If $S_{ij} < 1$ for all $i \neq j$, then Z is non-singular and Z^{-1} is an M-matrix.

Example

Suppose that

$$S = \begin{bmatrix} 1 & -0.5 & 0.5 & 0.6 \\ -0.5 & 1 & 0.4 & -0.1 \\ 0.5 & 0.4 & 1 & 0.2 \\ 0.6 & -0.1 & 0.2 & 1 \end{bmatrix}$$

Then Z is given by



$$Z = \begin{bmatrix} 1 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1 & 0.4 & 0.4 \\ 0.5 & 0.4 & 1 & 0.5 \\ 0.6 & 0.4 & 0.5 & 1 \end{bmatrix}.$$

The maximum cost spanning tree of S plays an important role here.

Upper bound on the sparsity pattern

Compute the maximum cost spanning tree of S.

- Kruskal's algorithm takes $\mathcal{O}(m^2 \log m)$ time.

$\overline{ij}$ = the path between i and j in this tree.

Theorem (Lauritzen et al., 2019b):

$$S_{ij} < \prod_{uv \in \overline{ij}} S_{uv} \qquad \Longrightarrow \qquad \hat{K}_{ij} = 0.$$

- This allows to identify *many* non-edges of the ML-graph.
- Estimation procedure becomes more efficient.

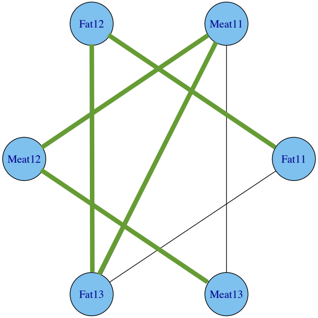
Example: Carcass data

- data: thickness of meat and fat layers at different locations on the back of a slaughter pig on each of 344 carcasses
- available in the package `gRbase` as `data(carcass)`

$$
\begin{array}{c}
\\
S =
\end{array}
\begin{array}{cccccc}
\text{Fat11} & \text{Meat11} & \text{Fat12} & \text{Meat12} & \text{Fat13} & \text{Meat13}
\end{array}
$$

$$
S = \begin{pmatrix}
1 & 0.04 & 0.84 & 0.08 & 0.82 & -0.03 \\
\cdot & 1 & 0.04 & 0.87 & 0.13 & 0.86 \\
\cdot & \cdot & 1 & 0.01 & 0.83 & -0.03 \\
\cdot & \cdot & \cdot & 1 & 0.11 & 0.90 \\
\cdot & \cdot & \cdot & \cdot & 1 & 0.02 \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1
\end{pmatrix}
\begin{array}{l}
\text{Fat11} \\ \text{Meat11} \\ \text{Fat12} \\ \text{Meat12} \\ \text{Fat13} \\ \text{Meat13}
\end{array}
$$

Example: Carcass data (2)



MTP_2 constraint

- Only one non-edge satisfies $S_{ij} \geq \prod_{uv \in \overline{ij}} S_{uv}$.

Sparsistency

Does such a procedure lead to a consistent estimation of the underlying graph? Of course not.

e.g. if the true graph has no edges, the estimated graph is not even sparse.

For regularized approaches see (Slawski and Hein, 2015; Egilmez et al., 2016; Pavez et al., 2018)

Proposition (Lauritzen et al., 2019b): The set of M-matrices is closed under thresholding.

Sparsistency without regularization

See (Wang et al., 2019) for details.

The basic observation is that, under MTP_2, if $K_{ij} < 0$ then

$$\rho_{ij|C} > 0 \quad \text{for all } C \subseteq V \setminus \{i, j\}.$$

The inequality is preserved in the sample distribution (for large n).

On the other hand:

- If $\rho_{ij|A} = 0$ then $\rho_{ij|B} = 0$ for all $B \supseteq A$ (upward stability).
- If true G is sparse such minimal A is small and so there are many B's satisfying $B \supseteq A$.
- Probability that all corresponding $\hat{\rho}_{ij|B}$ are nonnegative is very low.

Application in Portfolio Selection

See (Agrawal et al., 2019).

Optimal Markowitz Portfolio

Global minimum variance portfolio:

minimize $w^T \Sigma_t^* w$ subject to $w^T \mathbf{1} = 1$.

Replacing the unknown true covariance matrix of returns Σ_t^* by some estimator $\hat{\Sigma}_t$ yields the following analytical solution

$$\hat{w} \ = \ \frac{\hat{\Sigma}_t \mathbf{1}}{\mathbf{1}^T \hat{\Sigma}_t \mathbf{1}}$$

In this setting, estimating Σ_t^* becomes the main problem.

Covariance matrix estimators

Sample covariance is typically a bad estimator of Σ_t^*.

Structural assumptions give lower variance (higher bias).

- Dynamic factor models: Returns for day t are given by a linear combination of a (small) collection of latent factors.
- Static factor models: As above but Σ_t does not depend on t.
- Shrinkage of eigenvalues: see e.g. (Ledoit and Wolf, 2004, 2012).
- Regularization of the precision matrix: graphical lasso (Friedman et al., 2008; Ravikumar et al., 2011), CLIME (Liu et al., 2012).

Exploiting total positivity

MTP_2 constraint gives a natural regularizer.

Particularly so when applied to finance:

- capital asset pricing model (CAPM) (one factor model with positive loadings) is MTP_2.
- latent tree models used for unsupervised learning tasks (e.g. clustering similar stocks).

Data analysis suggests that MTP_2 regularization performs well where CAPM underfits. It outperforms other methods (Agrawal et al., 2019).

Extensions to heavy-tailed distributions

Typically the stocks data are log-transformed.

The log-transformed data may still be heavy-tailed.

With small modifications, similar approach works for elliptical distributions, or more generally, for for trans-elliptical distributions.

Lecture 3

Beyond the Gaussian case and beyond MTP_2

Transelliptical distributions

$X = (X_1, \ldots, X_m)$ has an elliptical distribution if its density function can be expressed as

$$g((x-\mu)^T \Sigma^{-1} (x-\mu)).$$

X has transelliptical distribution if $(f_1(X_1), \ldots, f_m(X_m))$ has elliptical distribution for some monotone functions $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}$:

$$X \sim TE(\Sigma_f; f_1, \ldots, f_m).$$

Theorem (Agrawal et al., 2019): If X is MTP_2 and transelliptical (so that $f(X)$ is elliptical) then Σ_f^{-1} is an M-matrix.

The other direction not true, e.g. no t-distribution is MTP_2.

M-matrix relaxation

Assuming Σ_f^{-1} is an M-matrix we relax the MTP_2 assumption.

- Compute Kendall's tau coefficients

$$\hat{\tau}_{ij} = \frac{1}{\binom{n}{2}} \sum_{1 \le t \le t' \le n} \mathrm{sign}(X_{it} - X_{it'})\mathrm{sign}(X_{jt} - X_{jt'})$$

- Use the relation between the Kendall's tau and the correlation coefficient (c.f. (Lindskog et al., 2003)): $(S_\tau)_{ij} = \sin(\frac{\pi}{2}\hat{\tau}_{ij})$.

- Follow a similar procedure as before with S_τ replacing S.

- This results in a consistent estimator an a relatively small efficiency loss (Liu et al., 2012; Barber and Kolar, 2018).

Exponential families

MTP_2 exponential families

- $p(x;\theta) = g(x)\exp(\langle\theta, T(x)\rangle - A(\theta)), \quad \theta \in \mathcal{K} \subseteq \mathbb{R}^d,\, x \in \mathcal{X}.$
  - (sufficient statistics) $T : \mathcal{X} \to \mathbb{R}^d$
  - (canonical parameters) $\mathcal{K} = \{\theta : \int \exp(\langle\theta, T(x)\rangle)\nu(\mathrm{d}x) < \infty\}$
- $\mathcal{K}$ is convex and $A(\theta)$ is strictly convex in $\mathcal{K}$

Theorem: The set of all $\theta \in \mathcal{K}$ such that $p(x;\theta)$ is MTP_2 is an intersection of $\mathcal{K}$ with a closed convex set $\mathcal{C}$.
(In many interesting cases $\mathcal{C}$ is a polyhedral cone)

Proof: Define $\Delta_{x,y}(\theta) = \log\left(\frac{p(x\vee y;\theta)p(x\wedge y;\theta)}{p(x;\theta)p(y;\theta)}\right)$, then

$$\Delta_{x,y}(\theta) = \langle\theta, T(x\vee y) + T(x\wedge y) - T(x) - T(y)\rangle + \mathrm{const}.$$

So $\Delta_{x,y}(\theta) \geq 0$ for all x, y defines a convex subset of $\mathcal{K}$.

Dependence on $g(x)$

Proposition (Lauritzen et al., 2019a): The MTP_2 property does not depend on the base measure. The counting/Lebesgue measure can be replaced by any other product measure.

Remark: If the exponential family contains a product distribution then it can be chosen to be the base measure.

- $\theta \in \mathcal{K}$ is MTP_2 if and only if $F(x) = -\langle \theta, T(x) \rangle$ is submodular.
- also $\mathcal{C}$ is a cone.

If F is twice differenciable then equivalently:

$$\left\langle \theta, \frac{\partial^2 T}{\partial x_i \partial x_j} \right\rangle \geq 0 \text{ and all } i \neq j, x \in \mathcal{X}.$$

Maximum likelihood estimation

$\log p(x;\theta) \;=\; \langle\theta, T(x)\rangle - A(\theta), \quad \theta \in \mathcal{K} \subseteq \mathbb{R}^d,\, x \in \mathcal{X}$

$x^{(1)}, \ldots, x^{(n)} \in \mathcal{X}$ independent sample; $\quad \bar{T} := \frac{1}{n}\sum_i T(x^{(i)})$

the log-likelihood function: $\quad \ell(\theta) = n\langle\theta, \bar{T}\rangle - nA(\theta)$

ℓ strictly concave in $\mathcal{K}$ and so the MLE is unique (if exists)

- Hence, MTP_2 distributions also admit a unique maximizer.
- The MLE usually exists under *much* weaker conditions on n.

Geometry of the MLE

- Let $\mathcal{S}$ denote the interior of $\mathrm{conv}(T(\mathcal{X}))$.
- The MLE exists in the EF if and only if $\bar{T} \in \mathcal{S}$.
- Let $\mathcal{K}_2 = \mathcal{K} \cap \mathcal{C}$ and define $\mathcal{S}_2 = \mathcal{S} + \mathcal{C}^\vee$, where

$$\mathcal{C}^\vee = \{\sigma : \langle \theta, \sigma \rangle \geq 0 \text{ for all } \theta \in \mathcal{C}\}.$$

Theorem (Lauritzen et al., 2019a): The MLE of θ based on $\bar{T}$ exists in the MTP_2 model if and only if $\bar{T} \in \mathcal{S}_2$. It is then equal to the unique element $\hat{\theta} = \nabla A(\hat{\sigma})$ that satisfies

(a) Primal feasibility: $\hat{\theta} \in \mathcal{K}_2$

(b) Dual feasibility: $\hat{\sigma} \in \mathcal{S}$ with $\bar{T} - \hat{\sigma} \in \mathcal{C}^\vee$,

(c) Complimentary slackness: $\langle \bar{T} - \hat{\sigma}, \hat{\theta} \rangle = 0$.

Binary distributions*

Support of MTP_2 distributions

State space: $\mathcal{X} = \{-1, 1\}^d$.

$\mathcal{P}_2$ = the set of all MTP_2 binary distributions.

Note: If $p \in \mathcal{P}_2$ then $\mathrm{supp}(p)$ is a sublattice of $\mathcal{X}$.

Proof: If $x, y \in \mathrm{supp}(p)$ then

$$0 < p(x)p(y) \leq p(x \wedge y)p(x \vee y).$$

Existence and uniqueness of the MLE

Sample $U = \{x_1, \ldots, x_n\}$, likelihood $L(p) = \prod_{i=1}^{n} p(x_i)$.

Theorem (Lauritzen et al., 2019a):

(i) There exists a unique maximum $\hat{p}$ of L over $\mathcal{P}_2$.

(ii) $\mathrm{supp}(\hat{p})$ is equal to the lattice generated by U.

Proof of (i): Continuity and compactness gives existence.

If $p, q \in \mathcal{P}_2$ then $c^{-1}\sqrt{pq} \in \mathcal{P}_2$ (geometric convexity)

By Cauchy-Schwarz $c = \sum_x \sqrt{p(x)q(x)} \leq 1$ (ineq. strict if $p \neq q$)

If $p \neq q$ both maximize L then

$$L(c^{-1}\sqrt{pq}) = c^{-n}\sqrt{L(p)L(q)} = c^{-n}L(p) > L(p) \quad (\textit{contradiction})$$

Binary exponential families

Sufficient statistics: $T : \mathcal{X} \to \{0,1\}^{\mathcal{X}} \quad (x \mapsto T_x)$.
$T_x(y) = 1$ if $x = y$ and $T_x(y) = 0$ otherwise.

Canonical parameter: $\theta \in \mathbb{R}^{\mathcal{X}}$.
$\theta(x) = \log p(x) - \log p(-\mathbf{1}) \quad$ (formally ignore $\theta(-\mathbf{1})$)

Inner product: $\langle \theta, T \rangle = \sum_{y \in \mathcal{X}} T(y)\theta(y)$.

Then we get

$$\log p(x) = \langle \theta, T_x \rangle - A(\theta)$$

with $A(\theta) = \log \left(\sum_{y \in \mathcal{X}} \exp(\theta(y)) \right)$.

Uniqueness and existence of the MLE

The binary exponential family is no longer compact.

Theorem: The MLE over this set exists if and only if the lattice generated by the sample U is equal to $\mathcal{X}$.

The MTP_2 distributions: for all $x, y \in \mathcal{X}$

$$\theta(x \vee y) + \theta(x \wedge y) - \theta(x) - \theta(y) \geq 0.$$

(convex condition!)

Reformulation of the MTP_2 condition

$\mathcal{S}$ semi-elementary imsets: for $x, y \in \mathcal{X}$

$$u_{x,y} \quad = \quad T_{x \wedge y} + T_{x \vee y} - T_x - T_y .$$

$\mathcal{E}$ elementary imsets: x, y differ in two entries

e.g. $x = (1, 1, -1, -1)$, $y = (1, -1, 1, -1)$
denote $u_{ij|A}$ where A indicates 1's in both x and y e.g. $u_{23|1}$

$p(\cdot;\theta)$ is MTP_2 if and only if

$$\langle \theta, v \rangle \ \geq \ 0 \qquad \forall v \in \mathcal{E}.$$

Optimality conditions

mean parameter: $\sigma = \mathbb{E}_\theta T_X$.

sample statistics: $\bar{T} = \frac{1}{n}\sum_{i=1}^n T_{x_i}$.

$(\hat{\theta},\hat{\sigma})$ is optimal if and only if:

primal feasibility: $\langle \hat{\theta}, v \rangle \geq 0$ for all $v \in \mathcal{E}$.

dual feasibility:

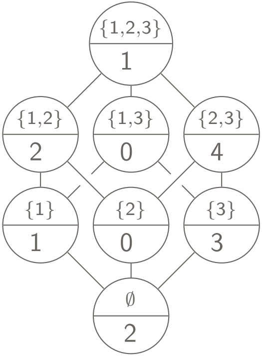
(i) $\hat{\sigma}(x) > 0$ for all $x \in \mathcal{X}$, and

(ii) $\hat{\sigma} - \bar{T}$ lies in the cone generated by $\mathcal{E}$.

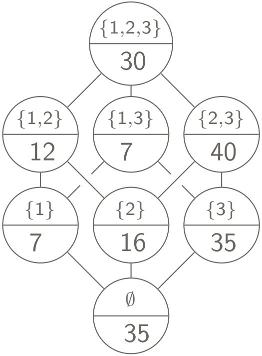
complementary slackness: $\langle \hat{\theta}, \hat{\sigma} - \bar{T} \rangle = 0$.

Example: $d = 3$

$\bar{T} = \frac{1}{13} \cdot$



$\hat{\sigma} = \frac{1}{182} \cdot$



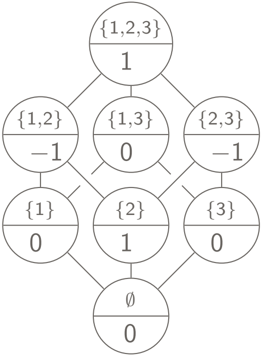
.

$$\begin{aligned}
\{1\},\{2\}: &\quad 12\cdot 35 - 7\cdot 16 > 0 \\
\{1\},\{3\}: &\quad \mathbf{7\cdot 35 - 7\cdot 35 = 0} \\
\{2\},\{3\}: &\quad 40\cdot 35 - 16\cdot 35 > 0
\end{aligned}
\qquad
\begin{aligned}
\{1,3\},\{2,3\}: &\quad 30\cdot 35 - 7\cdot 40 > 0 \\
\{1,2\},\{2,3\}: &\quad \mathbf{30\cdot 16 - 12\cdot 40 = 0} \\
\{1,2\},\{1,3\}: &\quad 30\cdot 7 - 12\cdot 7 > 0
\end{aligned},$$

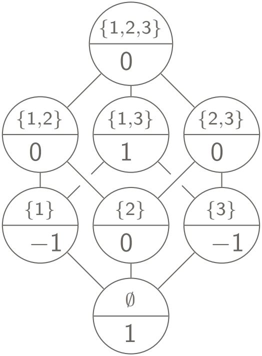
which assures primal feasibility.

$\hat{\sigma} > 0$ and the vector $\hat{\sigma} - \bar{T}$ can be written as

$\frac{16}{182} \cdot$



$+ \ \frac{7}{182} \cdot$

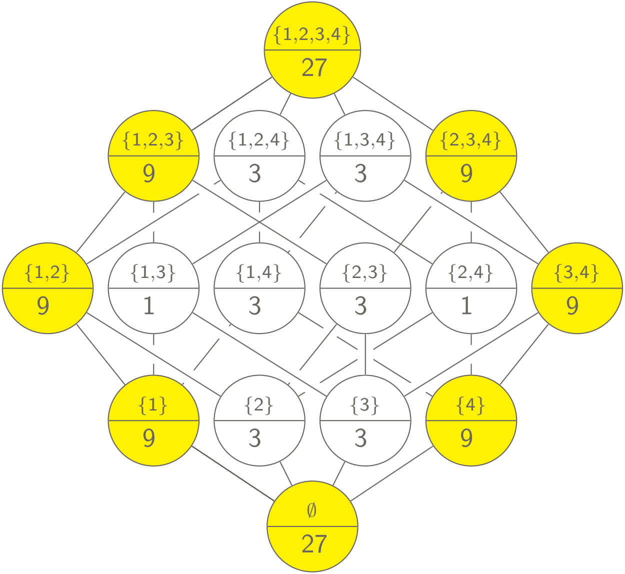


proving dual feasibility.

Complementary slackness follows by direct calculations.
(note two generators and two equalities)

Moussouris' example, $d = 4$

$\hat{\sigma} = \frac{1}{128} \cdot$



Primal feasibility:

$$
\begin{array}{llll}
\{1\},\{2\}: & 9\cdot 27 - 9\cdot 3 > 0 & \{1,3\},\{2,3\}: & 9\cdot 3 - 1\cdot 3 > 0 \\
\{1,4\},\{2,4\}: & 3\cdot 9 - 3\cdot 1 > 0 & \{1,3,4\},\{2,3,4\}: & 27\cdot 9 - 3\cdot 9 > 0 \\
\{1\},\{3\}: & \mathbf{1\cdot 27 - 9\cdot 3 = 0} & \{1,2\},\{2,3\}: & \mathbf{9\cdot 3 - 9\cdot 3 = 0} \\
\{1,4\},\{3,4\}: & \mathbf{3\cdot 9 - 3\cdot 9 = 0} & \{1,2,4\},\{2,3,4\}: & \mathbf{27\cdot 1 - 3\cdot 9 = 0} \\
\{1\},\{4\}: & \mathbf{3\cdot 27 - 9\cdot 9 = 0} & \{1,2\},\{2,4\}: & \mathbf{3\cdot 3 - 9\cdot 1 = 0} \\
\{1,3\},\{3,4\}: & \mathbf{3\cdot 3 - 1\cdot 9 = 0} & \{1,2,3\},\{2,3,4\}: & \mathbf{27\cdot 3 - 9\cdot 9 = 0} \\
\{2\},\{3\}: & 3\cdot 27 - 3\cdot 3 > 0 & \{1,2\},\{1,3\}: & 9\cdot 9 - 9\cdot 1 > 0 \\
\{2,4\},\{3,4\}: & 9\cdot 9 - 1\cdot 9 > 0 & \{1,2,4\},\{1,3,4\}: & 27\cdot 3 - 3\cdot 3 > 0 \\
\{2\},\{4\}: & \mathbf{1\cdot 27 - 3\cdot 9 = 0} & \{1,2\},\{1,4\}: & \mathbf{3\cdot 9 - 9\cdot 3 = 0} \\
\{2,3\},\{3,4\}: & \mathbf{9\cdot 3 - 3\cdot 9 = 0} & \{1,2,3\},\{1,3,4\}: & \mathbf{27\cdot 1 - 9\cdot 3 = 0} \\
\{3\},\{4\}: & 9\cdot 27 - 3\cdot 9 > 0 & \{1,3\},\{1,4\}: & 3\cdot 9 - 1\cdot 3 > 0 \\
\{2,3\},\{2,4\}: & 9\cdot 3 - 3\cdot 1 > 0 & \{1,2,3\},\{1,2,4\}: & 27\cdot 9 - 9\cdot 3 > 0
\end{array}
$$

(the MLE still Markov to the four-cycle)

Dual feasibility:

$$\begin{aligned}\hat{\sigma} - \bar{T} &= \tfrac{3}{128} \cdot u_{1,3|\emptyset} + \tfrac{1}{128} \cdot u_{1,3|2} + \tfrac{1}{128} \cdot u_{1,3|4} + \tfrac{3}{128} \cdot u_{1,3|2,4} + \\ &+ \tfrac{3}{128} \cdot u_{2,4|\emptyset} + \tfrac{1}{128} \cdot u_{2,4|1} + \tfrac{1}{128} \cdot u_{2,4|3} + \tfrac{3}{128} \cdot u_{2,4|1,3} + \\ &+ \tfrac{5}{128} \cdot u_{1,4|\emptyset} + \tfrac{5}{128} \cdot u_{1,4|2} + \tfrac{5}{128} \cdot u_{1,4|3} + \tfrac{5}{128} \cdot u_{1,4|2,3}.\end{aligned}$$

Complementary slackness can again be checked by hand.

Note that $\hat{\sigma} - \bar{T}$ is a positive combination of bold-faced rows from the previous slide.

Binary Ising model

The binary Ising model

The p.m.f. for $x \in \mathcal{X} = \{-1, 1\}^m$ satisfies

$$\log p(x; h, J) \;=\; h^T x \,+\, \tfrac{1}{2} x^T J x \,-\, A(h, J),$$

with $h \in \mathbb{R}^m$ and J symmetric with zeros on the diagonal.

This is a special subclass of:

- exponential families,
- pairwise interaction models, and
- graphical models.

The binary Ising model has dimension $\binom{m+1}{2} \ll 2^m - 1$.

Conditional odds-ratios

Fix i, j. If $x, y \in \mathcal{X}$ satisfy $x_{ij} = (-1, 1)$, $y_{ij} = (1, -1)$ and are equal otherwise then

$$\log\left(\frac{p(x \vee y)p(x \wedge y)}{p(x)p(y)}\right) \;=\; 4J_{ij}.$$

Some remarks:

- p is MTP_2 if and only if $J_{ij} \geq 0$.
- Conditional odds ratio does not depend on the condition.
- No direct link to M-matrices.

IPS algorithm for the MLE

Fix a graph $G = (V, E)$.

• Standard IPS algorithm for computing the MLE:

cycles through all pairs $ij \in E$ and optimizes the likelihood function with respect to h_i, h_j, J_{ij} keeping other parameters fixed.

• We initialize at any point. The update is:

$$p(x) \;\leftarrow\; p(x)\frac{e_{ij}(x_i,x_j)}{p_{ij}(x_i,x_j)}.$$

This affects only J_{ij}, h_i, h_j.

• If J_{ij} updates to a negative number set $J_{ij} \leftarrow 0$ and (h_i, h_j) to match sample means.

Application in psychology

Two psychological disorders

About the study, see e.g. (Borsboom and Cramer, 2013):

National Comorbidity Survey Replication (NCS-R data)

9282 observations of 18 binary variables such as:
`depr` (Depressed mood), `inte` (Loss of interest), etc

These are symptoms related to two disorders:
major depression and generalized anxiety disorder.

Bridge variables: sleep problems, fatigue, and concentration problems.

Two psychological disorders, continued

About the data:

Sparse contingency table: $872/65536$ nonzero cells.

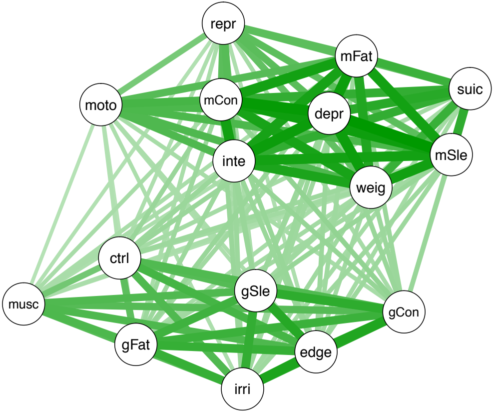
5667 out of 9282 respondents recorded no symptoms.

Positive sample correlations but not MTP_2.

Two variables perfectly correlated with each other and other seven variables (the MLE does not exist).

Two psychological disorders, continued

The sample correlation:

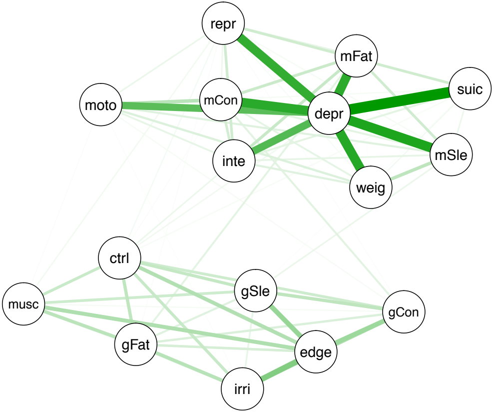


.

This network was reported by (Borsboom and Cramer, 2013).

Two psychological disorders, continued

The $\hat{J}$ matrix:



(Borsboom and Cramer, 2013) report a similar picture obtained after asking 12 Dutch clinicians for causal relationships!

Some research directions

Relaxations useful for statistical modelling

M-matrices offer a convenient relaxation for transelliptical distributions.

What about asymetric distributions, like skew normal?

Testing total positivity

In the Gaussian case, how can we test total positivity? Can we improve on (Bartolucci and Forcina, 2000) in the binary case?

Total positivity and hidden variables

In the Gaussian setting (Chandrasekaran et al., 2012) proposed a computationally efficient model selection technique for large sparse Gaussian graphical models with hidden variables. Would be interesting to see this in connection with the MTP_2 constraint.

Total positivity for noneuclidean spaces

Can we for example define a useful version of total positivity for Wishart matrices or for the Dirichlet distribution?

Thank you!

References I

Agrawal, R., Roy, U., and Uhler, C. (2019). Covariance matrix estimation under total positivity for portfolio selection. *arXiv preprint arXiv:1909.04222*.

Barber, R. F. and Kolar, M. (2018). Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. *The Annals of Statistics*, 46(6B):3422–3450.

Bartolucci, F. and Forcina, A. (2000). A likelihood ratio test for MTP_2 within binary variables. *Ann. Statist.*, 28(4):1206–1218.

Bølviken, E. (1982). Probability inequalities for the multivariate normal with non-negative partial correlations. *Scand. J. Statist.*, 9:49–58.

Borsboom, D. and Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, 9:91–121.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.*, 40(4):1935–1967.

References II

Dellacherie, C., Martinez, S., and San Martin, J. (2014). *Inverse M-matrices and ultrametric matrices*, volume 2118. Springer.

Dykstra, R. L. and Hewett, J. E. (1978). Positive dependence of the roots of a Wishart matrix. *The Annals of Statistics*, 6(1):235–238.

Dynkin, E. (1980). Markov processes and random fields. *Bulletin of the American Mathematical Society*, 3(3):975–999.

Egilmez, H. E., Pavez, E., and Ortega, A. (2016). Graph learning from data under structural and laplacian constraints. *arXiv preprint arXiv:1611.05181*.

Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N., and Zwiernik, P. (2017). Total positivity in Markov structures. *The Annals of Statistics*, 45(3):1152–1184.

Fortuin, C. M., Kasteleyn, P. W., and Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.*, 22(2):89–103.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gumbel, E. J. (1961). Bivariate logistic distributions. 56(294):335–349.

References III

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical models with R*. Springer Science & Business Media.

Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multiv. Anal.*, 10(4):467–498.

Karlin, S. and Rinott, Y. (1983). M-matrices as covariance matrices of multinormal distributions. *Linear Algebra Appl.*, 52:419 – 438.

Lauritzen, S., Uhler, C., and Zwiernik, P. (2019a). Total positivity in structured binary distributions. *arXiv preprint arXiv:1905.00516*.

Lauritzen, S., Uhler, C., Zwiernik, P., et al. (2019b). Maximum likelihood estimation in gaussian models under total positivity. *The Annals of Statistics*, 47(4):1835–1863.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford, United Kingdom.

Lebowitz, J. L. (1972). Bounds on the correlations and analyticity properties of ferromagnetic Ising spin systems. *Comm. Math. Phys.*, 28(4):313–321.

References IV

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Lindskog, F., McNeil, A., and Schmock, U. (2003). Kendall's tau for elliptical distributions. In *Credit Risk*, pages 149–156. Springer.

Liu, H., Han, F., and Zhang, C.-h. (2012). Transelliptical graphical models. In *Advances in neural information processing systems*, pages 800–808.

Pavez, E., Egilmez, H. E., and Ortega, A. (2018). Learning graphs with monotone topology properties and multiple connected components. *IEEE Transactions on Signal Processing*, 66(9):2399–2413.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Sadeghi, K. (2017). Faithfulness of probability distributions and graphs. *The Journal of Machine Learning Research*, 18(1):5429–5457.

References V

Sarkar, T. K. (1969). Some lower bounds of reliability. Tech. Report, No. 124, Department of Operations Research and Department of Statistics, Stanford University.

Slawski, M. and Hein, M. (2015). Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random field. *Linear Algebra Appl.*, 473:145–179.

Wang, Y., Roy, U., and Uhler, C. (2019). Learning high-dimensional gaussian graphical models under total positivity without tuning parameters. *arXiv preprint arXiv:1906.05159*.

Zwiernik, P. (2015). *Semialgebraic Statistics and Latent Tree Models*. Number 146 in Monographs on Statistics and Applied Probability. Chapman & Hall.