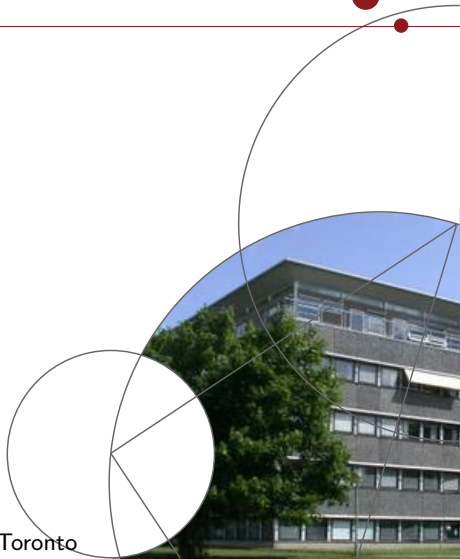
Mixed Convex Exponential Families

Steffen Lauritzen[1]



[1]Based on joint work with Piotr Zwiernik, University of Toronto

Overview

1. Mixed convex exponential families
2. Locally associated Gaussian graphical models
3. The mixed dual estimator
4. Estimating equations for laGGMs
5. The GOLAZO algorithm
6. Data example

Exponential family

Consider minimally represented and steep exponential family $\mathcal{E} = \{P_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in \Theta\}$ with canonical statistic $\boldsymbol{t} : \mathcal{X} \mapsto \mathbb{R}^k$ and canonical parameter $\boldsymbol{\theta}$:

$$p(\boldsymbol{x}; \boldsymbol{\theta}) \;=\; \exp\{\langle \boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\} \qquad \text{for } \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k,$$

where $\nu\{x : \langle \lambda, t(x) \rangle = c\} = 0$ if $\lambda \neq 0$.

The space of canonical parameters is

$$\Theta \;:= \text{int} \left\{ \boldsymbol{\theta} \in \mathbb{R}^k : \int_{\mathcal{X}} \exp\left\{ \langle \boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x}) \rangle \right\} \nu(\mathrm{d}\boldsymbol{x}) < \infty \right\}$$

and the cumulant function $A : \Theta \to \mathbb{R}$ is strictly convex and smooth with a gradient tending to ∞ at the boundary of Θ.

Mixed parametrisation

The map μ between the canonical parameter $\boldsymbol{\theta} \in \Theta$ and the mean parameter $\boldsymbol{\mu} \in M$ satisfies

$$\mu(\boldsymbol{\theta}) = \nabla A(\boldsymbol{\theta})$$

and establishes a smooth bijection between Θ and M. The inverse map is $\boldsymbol{\theta} = \theta(\boldsymbol{\mu})$.

Split canonical statistic into subvectors $\boldsymbol{t}(\boldsymbol{x}) = (\boldsymbol{u}, \boldsymbol{v})$ of dimension r, s where $r + s = k$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_v)$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_u, \boldsymbol{\mu}_v)$ corresponding splits.

The pair $(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$ forms an alternative parametrization $\mathcal{E}$ called the *mixed parametrization* with $\boldsymbol{\mu}_u$ and $\boldsymbol{\theta}_v$ *variation independent* (Barndorff-Nielsen, 1978, Theorem 8.4).

Thus we may without ambiguity write

$$\mathcal{E} = \{P_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in \Theta\} = \{P_{\boldsymbol{\mu}} \,|\, \boldsymbol{\mu} \in M\} = \{P_{(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)} \,|\, \boldsymbol{\mu}_u \in M_u, \boldsymbol{\theta}_v \in \Theta_v\}.$$

Mixed convex exponential family

Definition

Fix a mixed parametrization $(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$ of the exponential family $\mathcal{E}$. Consider submodel $\mathcal{E}' = \mathcal{E}_u \cap \mathcal{E}_v \subseteq \mathcal{E}$, where

(i) $\mathcal{E}_u = \{P_{\boldsymbol{\mu}} \,|\, \boldsymbol{\mu} \in C_u\}$ where $C_u \subseteq M$ is given by convex constraints on $\boldsymbol{\mu}_u$;

(ii) $\mathcal{E}_v = \{P_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in C_v\}$ where $C_v \subseteq \Theta$ is given by convex constraints on $\boldsymbol{\theta}_v$

Then $\mathcal{E}'$ is called a *mixed convex* submodel of $\mathcal{E}$ and a *mixed convex exponential family*.

It is important that the restrictions concern *variation independent* components of the parameters associated with $\boldsymbol{t}$.

Locally associated Gaussian graphical models

For the multivariate Gaussian we have

$$\boldsymbol{t}(\boldsymbol{x}) = -\boldsymbol{x}\boldsymbol{x}^T/2,\ \boldsymbol{\theta} = K,\ \boldsymbol{\mu} = -\Sigma/2,\ A(K) = -\frac{1}{2}\log\det K.$$

Θ is the cone of positive definite matrices and M the cone of negative definite matrices.

Example

Fix a graph $G = (V, E)$ and consider the family $\mathcal{N}_V(0, \Sigma)$.

Split to $\boldsymbol{u} = (-x_i x_j/2)_{ij\in E}$, and $\boldsymbol{v} = (-x_i x_j/2)_{ij\notin E}$. Include $-x_i^2/2$ in $\boldsymbol{u}$.

Then $\boldsymbol{\mu}_u = (-\Sigma_{ij}/2)_{ij\in E}$ and $\boldsymbol{\theta}_v = (K_{ij})_{ij\notin E}$.

Mixed convex family given by $\boldsymbol{\mu}_u \leqslant 0$ and $\boldsymbol{\theta}_v = 0$ is a *locally associated Gaussian graphical model* (Lauritzen and Zwiernik, 2022).

Association and positivity

Definition

A random vector X with values in $\mathbb{R}^d$ is *associated* if it holds for any pair $f, g : \mathbb{R}^d \mapsto \mathbb{R}$ of functions that are non-decreasing in each coordinate that

$$\mathbf{V}(f(X), g(X)) \geqslant 0.$$

Association is easy to check in the Gaussian case:

Theorem (Pitt (1982))

Suppose X is a Gaussian vector with covariance Σ then

$$X \text{ is associated} \quad \Longleftrightarrow \quad \Sigma \geqslant 0.$$

Relaxing the positivity

Although variants of positivity occur in many applications in finance, gene expression, etc. it often appears too strong.

So consider an undirected graph $\mathcal{G} = (V, E)$. We define

Definition

A random vector X with values in $\mathbb{R}^V$ is *locally associated* w.r.t. $\mathcal{G}$ if it holds for every clique $C \in \mathcal{C}$ and any pair $f, g : \mathbb{R}^C \mapsto \mathbb{R}$ of non-decreasing functions that

$$\mathbf{V}(f(X_C), g(X_C)) \geqslant 0.$$

We then clearly have from Pitt's theorem:

Theorem

A Gaussian vector X is locally associated w.r.t. $\mathcal{G}$ if and only if $\sigma_{ij} \geqslant 0$ for all edges $ij \in E$.

Locally associated Gaussian graphical models

Now combine the local association with conditional independence restrictions. So let $A(\mathcal{G})$ denote the Gaussian distributions that are locally associated and $M(\mathcal{G})$ denote those that are Markov w.r.t. $\mathcal{G}$.

Definition

A *locally associated Gaussian graphical model* (laGGM) is determined by an undirected graph $\mathcal{G}$ and the family of Gaussian distributions

$$M_+(\mathcal{G}) = A(\mathcal{G}) \cap M(\mathcal{G}).$$

Thus clearly a mixed convex exponential family.

Another mixed convex exponential family

Example

X and Y values in $\mathcal{S} = \{0, 1, \ldots, k\}$ with $p_{xy} = P(X = x, Y = y) > 0$

May be mixed parametrized with the *marginals*

$$\mu_{x+} = p_{x+},\ x \in \mathcal{S}\backslash\{0\}, \quad \mu_{+y} = p_{+y},\ y \in \mathcal{S}\backslash\{0\}$$

and the *interactions* $\theta_{xy} = \log \frac{p_{xy} p_{00}}{p_{x0} p_{0y}},\ x, y \in \mathcal{S}\backslash\{0\}$

Consider the hypothesis of *marginal homogeneity*

$$p_{x+} = p_{+x} \text{ for all } x \in \mathcal{S} \tag{1}$$

in combination with the distribution being MTP_2:

$$\theta_{xy} + \theta_{x'y'} - \theta_{xy'} - \theta_{x'y} \geqslant 0 \text{ for all } x \geqslant x' \text{ and } y \geqslant y'. \tag{2}$$

The restriction (1) is convex (linear) in $\boldsymbol{\mu}$ and (2) is convex in $\boldsymbol{\theta}$.

A small modification

Example

Another alternative would exploit that categories are ordered and for example specify that p_{i+} is stochastically smaller than p_{+i} i.e.

$$\sum_{x=0}^{j} p_{x+} \leqslant \sum_{y=0}^{j} p_{+y} \text{ for all } j \in \mathcal{S},$$

yielding a convex restriction also on the mean parameters; see Agresti (1983) and Agresti (2003) for further details of this model.

Note that without the result on variation independence of the mixed parametrization, it is not so clear that MTP_2 and the stochastic ordering are variation independent restrictions.

Likelihood and conjugate functions

Problem: Likelihood function may have multiple local maxima over $\mathcal{E}'$ as the restrictions are not convex in $\boldsymbol{\theta}$ and we consider another estimator.

Given a random sample $X^{(1)}, \ldots, X^{(n)}$ of size n denote

$$\boldsymbol{t} = \boldsymbol{t}_n = \sum_{i=1}^{n} \boldsymbol{t}(X^{(i)})/n$$

The log-likelihood function is strictly convex in $\boldsymbol{\theta}$ and

$$\ell(\boldsymbol{\theta}; \boldsymbol{t}) \;=\; \langle \boldsymbol{\theta}, \boldsymbol{t} \rangle - A(\boldsymbol{\theta}).$$

Since $\nabla \ell(\boldsymbol{\theta}; \boldsymbol{t}) = \boldsymbol{t} - \nabla A(\boldsymbol{\theta}) = \boldsymbol{t} - \mu(\boldsymbol{\theta})$, the unique optimizer is $\boldsymbol{\theta} = \theta(\boldsymbol{t})$.

The *Fenchel conjugate* of A is the strictly convex function

$$A^*(\boldsymbol{\mu}) \;=\; \sup\{\ell(\boldsymbol{\theta}; \boldsymbol{\mu}) : \boldsymbol{\theta} \in \mathbb{R}^k\}.$$

Dual likelihood function

For any fixed $\boldsymbol{\theta}$, the function

$$\breve{\ell}(\boldsymbol{\mu};\boldsymbol{\theta}) \;:=\; \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu})$$

is strictly concave in $\boldsymbol{\mu}$ and called the *dual log-likelihood function*.

Analogously to the log-likelihood function, $\breve{\ell}$ satisfies

$$\nabla_{\boldsymbol{\mu}} \breve{\ell}(\boldsymbol{\mu};\boldsymbol{\theta}) \;=\; \boldsymbol{\theta} - \nabla_{\boldsymbol{\mu}} A^*(\boldsymbol{\mu}) \;=\; \boldsymbol{\theta} - \theta(\boldsymbol{\mu}).$$

Kullback–Leibler divergence

Consider two distributions in $\mathcal{E}$, one with the mean parameter $\boldsymbol{\mu}_1$ and the other with canonical parameter $\boldsymbol{\theta}_2$.

The *Kullback–Leibler divergence* $D(f \,|\, g) = \int \log(f(x)/g(x)) f(x)\, dx$ between these is

$$\mathbf{K}(\boldsymbol{\mu}_1, \boldsymbol{\theta}_2) \;=\; -\langle \boldsymbol{\mu}_1, \boldsymbol{\theta}_2 \rangle + A^*(\boldsymbol{\mu}_1) + A(\boldsymbol{\theta}_2).$$

The Kullback–Leibler divergence $\mathbf{K}(\boldsymbol{\mu}_1, \boldsymbol{\theta}_2)$ is strictly convex both in $\boldsymbol{\mu}_1$ and in $\boldsymbol{\theta}_2$. Also

$$\nabla_{\boldsymbol{\mu}} K(\boldsymbol{\mu}_1, \boldsymbol{\theta}_2) = -\boldsymbol{\theta}_2 + \nabla A^*(\boldsymbol{\mu}_1) = \theta(\boldsymbol{\mu}_1) - \boldsymbol{\theta}_2$$

and

$$\nabla_{\boldsymbol{\theta}} K(\boldsymbol{\mu}_1, \boldsymbol{\theta}_2) = -\boldsymbol{\mu}_1 + \nabla A(\boldsymbol{\theta}_2) = \mu(\boldsymbol{\theta}_2) - \boldsymbol{\mu}_1$$

The mixed dual estimator

We propose a two-step procedure to estimate the mixed parameter $(\boldsymbol{\mu}_u, \boldsymbol{\theta}_v)$ in the mixed convex family from data $\boldsymbol{t}$:

(S1) First minimize $\mathbf{K}(\boldsymbol{t}, \boldsymbol{\theta})$ over $\boldsymbol{\theta} \in C_v \subseteq \Theta$. Denote the unique optimum, assuming it exists, by $\widehat{\boldsymbol{\theta}}$.

(S2) Then minimize $\mathbf{K}(\boldsymbol{\mu}, \widehat{\boldsymbol{\theta}})$ subject to $\boldsymbol{\mu} \in C_u \subseteq M$. Denote the unique optimum by $\breve{\boldsymbol{\mu}}$.

The resulting $\breve{\boldsymbol{\mu}}$ is the *mixed dual estimator* (MDE) of $\boldsymbol{\mu}$.

Both steps (S1) and (S2) are convex optimization problems.

The optimum in (S1) is the MLE under the convex exponential family given by $\boldsymbol{\theta} \in C_v$.

Fundamental properties

Proposition

If the optimum $\widehat{\boldsymbol{\theta}}$ in (S1) exists then it is unique and the optimum $\breve{\boldsymbol{\mu}}$ in (S2) exists and is unique too.

Theorem

Let $\boldsymbol{t} = (\boldsymbol{u}, \boldsymbol{v})$ and suppose that $\widehat{\boldsymbol{\theta}}$ in step (S1) exists. Then, $\mu_u(\widehat{\boldsymbol{\theta}}) = \boldsymbol{u} \in M_u$ and in step (S2) we get that $\theta_v(\breve{\boldsymbol{\mu}}) = \widehat{\boldsymbol{\theta}}_v$.

In particular, after steps (S1) and (S2), the optimum $\breve{\boldsymbol{\mu}}$ lies in the mixed convex family $\mathcal{E}'$.

Thus (S2) preserves constraints of (S1).

Asymptotics

Let $\widetilde{\boldsymbol{\mu}}_n$ be the MLE for n observations and $\breve{\boldsymbol{\mu}}_n$ the MDE.

Theorem

The MDE and MLE are asymptotically equivalent, i.e. $\sqrt{n}(\breve{\boldsymbol{\mu}}_n - \widetilde{\boldsymbol{\mu}}_n) \to 0$ *in probability,* $\sqrt{n}(\breve{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_0)$ *converges to the same limiting distribution as* $\sqrt{n}(\widetilde{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_0)$.

The proof is technical but relies on the fact that $\ell(\boldsymbol{\psi}, \boldsymbol{t})$ and $\breve{\ell}(\boldsymbol{\psi}, \boldsymbol{t})$ have the same Hessian at the MLE and the components of the MLE for the mixed parameter are asymptotically independent.

Note that the asymptotic distribution involved may be quite complicated, obtained by projecting the mixed parameters of the usual asymptotic normal distribution onto the relevant convex sets.

Estimation in a laGGM

Theorem

If the MDE $\breve{\Sigma}$ of Σ under $M_+(G) = A(G) \cap M(G)$ exists, it is the unique positive definite solution to the following where $\breve{K} = \breve{\Sigma}^{-1}$ and $\widehat{K} = \widehat{\Sigma}^{-1}$:

i) $\widehat{\Sigma}_{ij} = S_{ij}, \quad ij \in E(G);$

ii) $\widehat{\Sigma}_{ii} = S_{ii}, \quad i \in V(G);$

iii) $\widehat{K}_{ij} = 0, \quad ij \notin E(G); \quad 0 = \breve{K}_{ij}, \quad ij \notin E(G);$

iv) $\breve{\Sigma}_{ij} \geqslant 0, \quad ij \in E(G);$

v) $\breve{K}_{ij} \leqslant \widehat{K}_{ij}, \quad ij \in E(G);$

vi) $\breve{K}_{ii} = \widehat{K}_{ii}, \quad i \in V(G);$

vii) $\breve{\Sigma}_{ij}(\widehat{K}_{ij} - \breve{K}_{ij}) = 0, \quad ij \in E(G).$

Algorithms

Generally, the MDE leads to solving two convex optimization problems, and good methods for solving these may have to be developed case by case.

For the case of laGGMs, Lauritzen and Zwiernik (2022) developed a simple generic algorithm—the GOLAZO—for solving that particular problem and a range of other relevant problems associated with Gaussian graphical models.

The GOLAZO

Let L, U be two $d \times d$ matrices with entries in $\mathbb{R} \cup \{-\infty, +\infty\}$ satisfying

$$L_{ij} \leqslant 0 \leqslant U_{ij} \text{ for all } i \neq j.$$

Denote

$$\|K\|_{LU} := \sum_{i \neq j} \max\{L_{ij}K_{ij}, U_{ij}K_{ij}\}.$$

The function $\|K\|_{LU}$ is convex, positively homogeneous, continuous, and non-negative.

Although it is sublinear, that is

$$\|K + K'\|_{LU} \leqslant \|K\|_{LU} + \|K'\|_{LU},$$

it does not define a norm unless $|L_{ij}| = |U_{ij}|$ for all $i \neq j$. We aim at solving the following problem

$$\text{minimize} \quad -\ell_n(K) + \|K\|_{LU},$$

GOLAZO instances

$$\text{minimize} \quad -\ell_n(K) + \|K\|_{LU},$$

where

$$\|K\|_{LU} := \sum_{i \neq j} \max\{L_{ij}K_{ij}, U_{ij}K_{ij}\}.$$

Graphical lasso $|L_{ij}| = |U_{ij}| = \rho > 0$ for all $i \neq j$

Positive graphical lasso $L = 0$ and $U_{ij} = \rho$ for all $i \neq j$

MTP$_2$ Gaussian $L = 0$ and $U_{ij} = +\infty$ for all $i \neq j$ ($\pm\infty \cdot 0 = 0$)

Gaussian graphical models To ensure $K_{ij} = 0$, let $L_{ij} = -\infty$, $U_{ij} = +\infty$

Mixed dual estimate Replace K with Σ, S with $\hat{K}$, let $L_{ij} = -\infty$ and $U_{ij} = 0$ for all $i \neq j$

Exploiting duality

Note that

$$\max\{L_{ij}K_{ij}, U_{ij}K_{ij}\} = \sup_{L_{ij}\leqslant\Gamma_{ij}\leqslant U_{ij}} \Gamma_{ij}K_{ij}$$

and so

$$\|K\|_{LU} = \sup_{L\leqslant\Gamma\leqslant U} \mathrm{tr}(\Gamma K)$$

whereby the optimization may be written as

$$\inf_{K>0} \sup_{L\leqslant\Gamma\leqslant U} \big\{ -\log\det K + \mathrm{tr}((S+\Gamma)K) \big\}.$$

Swapping inf with sup and using that the infimum with respect to K of the expression is attained as $K = (S+\Gamma)^{-1}$, we obtain the dual problem by letting $\Sigma = S + \Gamma$:

$$\text{maximize} \quad \log\det\Sigma + d \quad \text{subject to } S + L \leqslant \Sigma \leqslant S + U.$$

Description of the algorithm

For the j-th row we consider $\log\det\Sigma$ as the function of $\Sigma_{j,\setminus j}$ keeping the other entries of Σ fixed. We have

$$\log\det\Sigma \;=\; \log\det\Sigma_{\setminus j} + \log\det(\Sigma_{jj} - \Sigma_{j,\setminus j}(\Sigma_{\setminus j})^{-1}\Sigma_{\setminus j,j}).$$

Thus maximizing $\log\det\Sigma$ with respect to $y := \Sigma_{j,\setminus j}$ is equivalent to minimizing $y^T(\Sigma_{\setminus j})^{-1}y$ under the linear conditions that $S_{ij} + L_{ij} \leqslant y_i \leqslant S_{ij} + U_{ij}$ for every $i \in V\setminus\{j\}$. This is an instance of a quadratic program.

The starting point Σ^0 of the algorithm needs to be chosen carefully so that Σ^0 is dually feasible. In this case, each iterate of the algorithm is guaranteed to be dually feasible.

To solve the quadratic program in each iteration we use the `quadprog` package in `R`.

The GOLAZO algorithm

Data: Positive semidefinite matrix S, penalty matrices $L \leqslant 0 \leqslant U$.
Result: A maximizer of the dual problem.
Initialize: $\Sigma = \Sigma^0$ (a dually feasible point);
while *no convergence* **do**
  for $j = 1, \ldots, d$ **do**
    Update $\Sigma_{j,\setminus j} \leftarrow \hat{y}$, where

$$\hat{y} = \arg\min_y \left\{ y^T (\Sigma_{\setminus j})^{-1} y : \; S_{j,\setminus j} + L_{j,\setminus j} \leqslant y \leqslant S_{j,\setminus j} + U_{j,\setminus j} \right\}.$$

  end
end

Convergence and starting point

To establish convergence we track the duality gap

$$\mathrm{tr}(SK) - d + \|K\|_{LU},$$

which is non-negative for each step of the algorithm, decreases at each iteration, and is zero at the optimum.

We stop the algorithm once this positive gap becomes sufficiently close to zero.

We have methods for identifying a dually feasible starting point.

This implies that *for the graphical lasso, the optimum always exists* even when the diagonal is not penalized. Generally this holds for the GOLAZO if $L_{ij} < 0 < U_{ij}$.

Positive co-expression gene network

Microarray expression data profiling umbilical cord (UC) tissue; cf. Costa and Castelo (2016). Obtained from Robert Castelo in a normalized and filtered version.

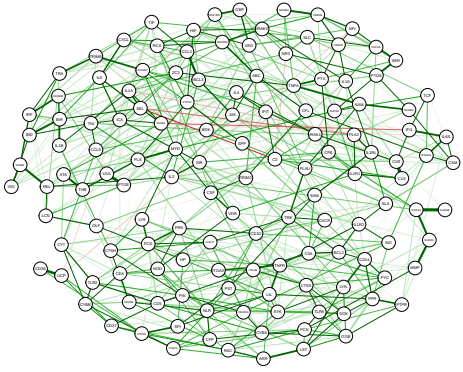
Set of 12,093 genes reduced to 704 with a role in the innate immune response. Then further reduced by focusing on a subset of 136 upregulated genes.

Vanilla implementation of GOLAZO (available on GitHub in package `golazo`) less than a minute on a standard laptop.

Resulting positive graphical lasso estimate $\hat{K}^\rho$ for $\rho = .5$ (optimal using EBIC) is very sparse with edge density 0.067. Still diameter of this graph is very small, just 5.

Optimum $\hat{K}^\rho$ not an M-matrix. However, estimated distribution is locally associated so second step of procedure is redundant.

Partial correlations in gene network



Positive partial correlations are indicated with green color and negative partial correlations with red color. The thickness of edges is proportional to their absolute size.

References

Agresti, A. (1983). Testing marginal homogeneity for ordinal categorical variables. *Biometrics*, 39(2):505–510.

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. John Wiley and Sons, New York.

Costa, D. and Castelo, R. (2016). Umbilical cord gene expression reveals the molecular architecture of the fetal inflammatory response in extremely preterm newborns. *Pediatric research*, 79(3):473–481.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28:157–175.

Grone, R., Johnson, C. R., Sá, E. M., and Wolkowicz, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra and its Applications*, 58:109–124.

Højsgaard, S. and Lauritzen, S. L. (2008). Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society, Series B*, 70:1005–1027.

Lauritzen, S. and Zwiernik, P. (2022). Locally associated graphical models and mixed convex exponential families. *arXiv preprint arXiv:2008.04688*. To appear in *The Annals of Statistics*.

Pitt, L. D. (1982). Positively correlated normal variables are associated. *The Annals of Probability*, 10:496–499.