Chair of Mathematical Statistics
Department of Mathematics
Technical University of Munich

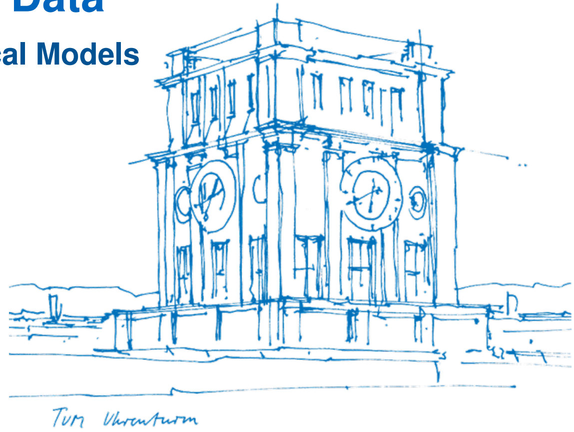
High-Dimensional Undirected Graphical Models for Arbitrary Mixed Data

ETH-UCPH-TUM Workshop on Graphical Models

K. Göbler, M. Drton, S. Mukherjee, A. Miloschewski

Chair of Mathematical Statistics
Department of Mathematics
Technical University of Munich

October 12, 2022



Outline

Motivation

- Long tradition of estimating undirected GMs for discrete or continuous data.

Motivation

- Long tradition of estimating undirected GMs for discrete or continuous data.
- Mixed graphs have seen less attention.
- Seminal work by Lauritzen and Wermuth [1989], Lauritzen [1996] on the conditional Gaussian distribution and its Markov properties $\rightarrow$ later adopted to the high-dimensional setting by Cheng et al. [2017].

Motivation

- Long tradition of estimating undirected GMs for discrete or continuous data.
- Mixed graphs have seen less attention.
- Seminal work by Lauritzen and Wermuth [1989], Lauritzen [1996] on the conditional Gaussian distribution and its Markov properties $\rightarrow$ later adopted to the high-dimensional setting by Cheng et al. [2017].
- Fan et al. [2017] proposed a latent generative model for mixed data $\rightarrow$ only binary-continuous mix.

Workhorse: The nonparanormal family

- According to Liu et al. [2009], a random vector $\boldsymbol{Z} \in \mathbb{R}^d$ has a nonparanormal distribution if there exist functions $\{f_j\}_{j=1}^d$ such that $f(\boldsymbol{Z}) \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Workhorse: The nonparanormal family

- According to Liu et al. [2009], a random vector $\boldsymbol{Z} \in \mathbb{R}^d$ has a nonparanormal distribution if there exist functions $\{f_j\}_{j=1}^d$ such that $f(\boldsymbol{Z}) \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- If the f_j's are differentiable and monotone then nonparanormal distribution $\iff$ Gaussian copula

Workhorse: The nonparanormal family

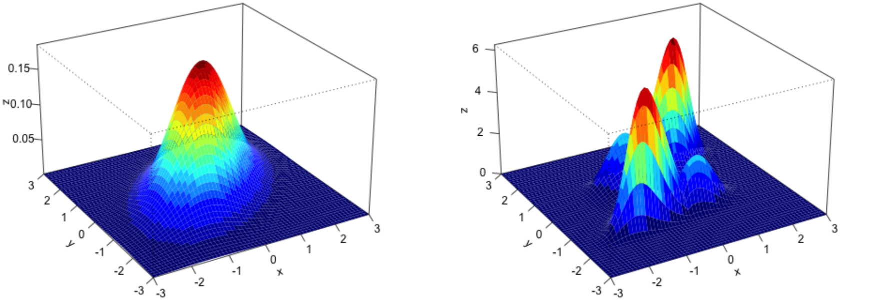
- According to Liu et al. [2009], a random vector $\boldsymbol{Z} \in \mathbb{R}^d$ has a nonparanormal distribution if there exist functions $\{f_j\}_{j=1}^d$ such that $f(\boldsymbol{Z}) \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- If the f_j's are differentiable and monotone then nonparanormal distribution $\iff$ Gaussian copula

- The independence graph of the nonparanormal is encoded in $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

- $\Omega_{jk} = 0 \iff Z_j \perp\!\!\!\perp Z_k \mid Z_{\{\setminus j,k\}}$

Example: The Normal and the Nonparanormal



Comparison between a 2-dimensional Gaussian and a 2-dimensional nonparanormal with $\boldsymbol{\mu} = (0, 0)$, $\Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$, and $f_j(x) =$ sign$(x)|x|^{\alpha_j}$ and $\alpha_1 = 1.5$ and $\alpha_2 = 2.5$.

Outline

Latent Gaussian Copula Model (LGCM)

- Assume we have a mix of (ordered) discrete and continuous variables, i.e. $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ of size $d_1 + d_2 = d$.

Latent Gaussian Copula Model (LGCM)

- Assume we have a mix of (ordered) discrete and continuous variables, i.e. $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ of size $d_1 + d_2 = d$.
- Lets assume there exists $\boldsymbol{Z}_1 = (Z_1, \ldots, Z_{d_1})^T$ s.t. $\boldsymbol{Z} := (\boldsymbol{Z}_1, \boldsymbol{X}_2) \sim \text{NPN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, f)$ where $\boldsymbol{\mu} = (\mu_j)_{j=1,\ldots,d}$ is the mean vector and $\boldsymbol{\Sigma}^* = (\Sigma^*_{jk})_{1\leq j,k\leq d}$ the correlation matrix and $f = \{f_1, \ldots, f_d\}$ a set of monotone differentiable univariate functions.

Latent Gaussian Copula Model (LGCM)

- Assume we have a mix of (ordered) discrete and continuous variables, i.e. $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ of size $d_1 + d_2 = d$.
- Lets assume there exists $\boldsymbol{Z}_1 = (Z_1, \ldots, Z_{d_1})^T$ s.t. $\boldsymbol{Z} := (\boldsymbol{Z}_1, \boldsymbol{X}_2) \sim \text{NPN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, f)$ where $\boldsymbol{\mu} = (\mu_j)_{j=1,\ldots,d}$ is the mean vector and $\boldsymbol{\Sigma}^* = (\Sigma^*_{jk})_{1 \le j,k \le d}$ the correlation matrix and $f = \{f_1, \ldots, f_d\}$ a set of monotone differentiable univariate functions.
- Relationship between observed discrete variables $\boldsymbol{X}_1$ and latent continuous variables $\boldsymbol{Z}_1$ is given by:

$$X_j = x_r^j \quad if \quad \gamma_{r-1}^j \le Z_j < \gamma_r^j$$

for $j = 1, \ldots, d_1$ and $r = 1, \ldots, l_{X_j}$ and $\gamma_0^j = -\infty$ and $\gamma_{l_{X_j}}^j = +\infty$.

Latent Gaussian Copula Model (LGCM)

- Assume we have a mix of (ordered) discrete and continuous variables, i.e. $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ of size $d_1 + d_2 = d$.
- Lets assume there exists $\boldsymbol{Z}_1 = (Z_1, \dots, Z_{d_1})^T$ s.t. $\boldsymbol{Z} := (\boldsymbol{Z}_1, \boldsymbol{X}_2) \sim \text{NPN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, f)$ where $\boldsymbol{\mu} = (\mu_j)_{j=1,\dots,d}$ is the mean vector and $\boldsymbol{\Sigma}^* = (\Sigma^*_{jk})_{1 \le j,k \le d}$ the correlation matrix and $f = \{f_1, \dots, f_d\}$ a set of monotone differentiable univariate functions.
- Relationship between observed discrete variables $\boldsymbol{X}_1$ and latent continuous variables $\boldsymbol{Z}_1$ is given by:

$$X_j = x_r^j \quad if \quad \gamma_{r-1}^j \le Z_j < \gamma_r^j$$

  for $j = 1, \dots, d_1$ and $r = 1, \dots, l_{X_j}$ and $\gamma_0^j = -\infty$ and $\gamma_{l_{X_j}}^j = +\infty$.
- In short we write $\boldsymbol{X} \sim \text{LNPN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, f, \boldsymbol{\Gamma})$ where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}^1, \dots, \boldsymbol{\gamma}^{d_1})$ is a collection of thresholds.

Latent generative scheme: Example

- Let us consider an example with an ordinal variable X_1 that can take 3 different values, say $\{1, 2, 3\}$.
- We assume there exists a latent continuous variable Z_1 with the following relation:

$$
\begin{tikzcd}
& \boxed{Z_1 \in (-\infty, \infty)} \arrow[dl, squiggly] \arrow[d, squiggly] \arrow[dr, squiggly] & \\
(-\infty, \gamma_1^1) \arrow[dr, squiggly] & [\gamma_1^1, \gamma_2^1) \arrow[d, squiggly] & [\gamma_2^1, \infty) \arrow[dl, squiggly] \\
& \boxed{\{1,\quad 2,\quad 3\} \ni X_1} &
\end{tikzcd}
$$

Outline

Mode of action

1. Find estimate of the sample correlation matrix $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}_{jk}^{(n)})_{1\leq j,k\leq d}$ of $\mathbf{\Sigma}^*$.

Mode of action

1. Find estimate of the sample correlation matrix $\hat{\boldsymbol{\Sigma}}^{(n)} = (\hat{\Sigma}_{jk}^{(n)})_{1 \le j,k \le d}$ of $\boldsymbol{\Sigma}^*$.

2. Plug estimate of the sample correlation matrix into existing routines for estimating $\boldsymbol{\Omega}^*$, e.g. glasso

$$\hat{\boldsymbol{\Omega}} = \arg\min_{\Omega \succeq 0} \left[\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{(n)} \boldsymbol{\Omega}) - \log|\boldsymbol{\Omega}| + \lambda \sum_{j \neq k} |\Omega_{jk}| \right].$$

Mode of action

1. Find estimate of the sample correlation matrix $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}_{jk}^{(n)})_{1 \leq j,k \leq d}$ of $\mathbf{\Sigma}^*$.

2. Plug estimate of the sample correlation matrix into existing routines for estimating $\mathbf{\Omega}^*$, e.g. glasso
$$\hat{\mathbf{\Omega}} = \underset{\Omega \succeq 0}{\arg\min} \left[\mathrm{tr}(\hat{\mathbf{\Sigma}}^{(n)} \mathbf{\Omega}) - \log|\mathbf{\Omega}| + \lambda \sum_{j \neq k} |\Omega_{jk}| \right].$$

3. Choose graph that minimizes some information criterion, e.g. extended BIC that additionally accounts for dimensionality of the problem [Foygel and Drton, 2010].

Estimating Σ^*

- We have to take care of three different cases for the couple (X_j, X_k).

Estimating Σ^*

- We have to take care of three different cases for the couple (X_j, X_k).
  1. Case I: both X_j and X_k are continuous,

Estimating Σ^*

- We have to take care of three different cases for the couple (X_j, X_k).
  1. Case I: both X_j and X_k are continuous,
  2. Case II: X_j is discrete and X_k is continuous (or vice versa),

Estimating Σ^*

- We have to take care of three different cases for the couple (X_j, X_k).
  1. Case I: both X_j and X_k are continuous,
  2. Case II: X_j is discrete and X_k is continuous (or vice versa),
  3. Case III: both X_j and X_k are discrete.

Estimating Σ^*

- We have to take care of three different cases for the couple (X_j, X_k).
  1. Case I: both X_j and X_k are continuous,
  2. Case II: X_j is discrete and X_k is continuous (or vice versa),
  3. Case III: both X_j and X_k are discrete.

- The product moment correlation between the latent continuous and the observed discrete variable (Case II) is called point polyserial correlation [Pearson, 1909, Bedrick, 1992].

Estimating Σ^*

- We have to take care of three different cases for the couple (X_j, X_k).
  1. Case I: both X_j and X_k are continuous,
  2. Case II: X_j is discrete and X_k is continuous (or vice versa),
  3. Case III: both X_j and X_k are discrete.

- The product moment correlation between the latent continuous and the observed discrete variable (Case II) is called point polyserial correlation [Pearson, 1909, Bedrick, 1992].

- Between both latent continuous variables (Case III) it's called point polychoric correlation [Pearson, 1900, Olsson, 1979].

Case I

Definition 1 (Estimator $\hat{\Sigma}^{(n)}$ of Σ^*; Case I nonparanormal)**.** *The estimator* $\hat{\Sigma}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1 \leq j,k \leq d}$ *of the covariance matrix* Σ^* *is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = 2 \sin \frac{\pi}{6} \hat{\rho}^{Sp}_{jk}$$

for all $d_1 < j < k \leq d_2$*.*

Case II

- Rank estimators are no longer available in general.

Case II

- Rank estimators are no longer available in general.
- Since $f(\boldsymbol{Z}) \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we have the following conditional expectation

$$E[f(X_k) \mid f(Z_j)] = \mu_{f(X_k)} + \Sigma^*_{jk} \sigma_{f(X_k)} f(Z_j), \quad \text{for } 1 \leq j \leq d_1 < k \leq d_2,$$

where we can assume w.l.o.g. that $\mu_{f(X_k)} = 0$.

Case II

- Rank estimators are no longer available in general.
- Since $f(\boldsymbol{Z}) \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we have the following conditional expectation

$$E[f(X_k) \mid f(Z_j)] = \mu_{f(X_k)} + \Sigma^*_{jk}\sigma_{f(X_k)} f(Z_j), \quad \text{for } 1 \leq j \leq d_1 < k \leq d_2,$$

  where we can assume w.l.o.g. that $\mu_{f(X_k)} = 0$.
- Multiplying both sides by X_j and dragging it into the expectation (function of $f(Z_j)$) we have

$$E[f(X_k)X_j \mid f(Z_j)] = \Sigma^*_{jk}\sigma_{f(X_k)} f(Z_j) X_j.$$

Case II

- Rank estimators are no longer available in general.
- Since $f(\boldsymbol{Z}) \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we have the following conditional expectation

$$E[f(X_k) \mid f(Z_j)] = \mu_{f(X_k)} + \Sigma^*_{jk}\sigma_{f(X_k)} f(Z_j), \quad \text{for } 1 \le j \le d_1 < k \le d_2,$$

  where we can assume w.l.o.g. that $\mu_{f(X_k)} = 0$.
- Multiplying both sides by X_j and dragging it into the expectation (function of $f(Z_j)$) we have

$$E[f(X_k)X_j \mid f(Z_j)] = \Sigma^*_{jk}\sigma_{f(X_k)} f(Z_j) X_j.$$

- Apply LIE, rearrange, and expand by σ_{X_j}, then

$$\Sigma^*_{jk} = \frac{E[f(X_k)X_j]}{\sigma_{f(X_k)} E[f(Z_j)X_j]} = \frac{r_{f(X_k)X_j}\sigma_{X_j}}{E[f(Z_j)X_j]}.$$

Case II

Definition 2 (Estimator $\hat{\Sigma}^{(n)}$ of Σ^*; Case II nonparanormal)**.** *The estimator $\hat{\Sigma}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1\le j,k\le d}$ of the covariance matrix Σ^* is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = \frac{r^{(n)}_{\hat{f}(X_k),X_j}\,\sigma^{(n)}_{X_j}}{\sum_{r=1}^{l_{X_j}-1}\phi(\hat{\gamma}^j_r)(x^j_{r+1}-x^j_r)}$$

for all $1 < j \le d_1 < k \le d_2$.

This is a double two-step estimator where first the thresholds and the unknown transformation functions f are estimated and then the expression above.

Case III

Definition 3 (Estimator $\hat{\Sigma}^{(n)}$ of Σ^*; Case III nonparanormal)**.** *The estimator $\hat{\Sigma}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1 \leq j,k \leq d}$ of the covariance matrix Σ^* is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = \arg\max_{|\Sigma_{jk}| \leq 1} \frac{1}{n} \ell^{(n)}(\Sigma_{jk}, x^j_r, x^k_s)$$

for all $1 < j < k \leq d_1$.

Outline

Concentration results

- Concentration case I (no latent variables) can be found in Liu et al. [2012]

Concentration results

- Concentration case I (no latent variables) can be found in Liu et al. [2012]
- Concentration case II was challenging.

Theorem 2. *Suppose that . . . some mild requirements . . . Then the following probability bound for case II holds*

$$P\left(\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk}-\Sigma^*_{jk}\right|\geq\epsilon\right)\leq 8\exp\left(2\log d-\frac{\sqrt{n}\epsilon^2}{(64\ L\ C_\gamma\ l_{\max}\ \pi)^2\log n}\right)$$
$$+8\exp\left(2\log d-\frac{n\epsilon^2}{(4L\ C_\gamma)^2\ 128(1+4c^2)^2}\right)$$
$$+8\exp\left(2\log d-\frac{\sqrt{n}}{8\pi\log n}\right)+4\exp\left(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\right)+\frac{2}{\sqrt{\pi\log(nd_2)}}.$$

Concentration results

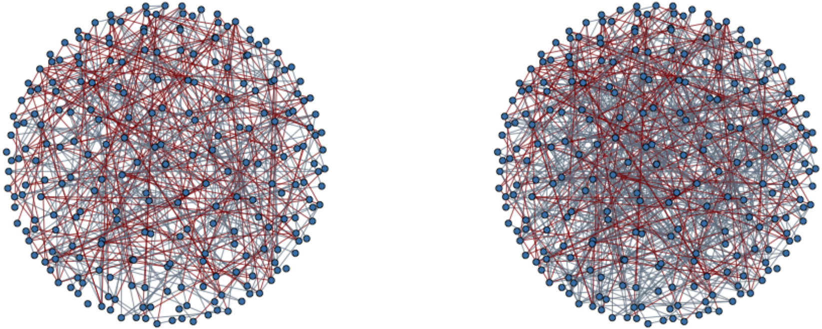
- Concentration case I (no latent variables) can be found in Liu et al. [2012]
- Concentration case II was challenging.
- Concentration case III requires the likelihood functions to behave nicely.

Theorem 3. *Suppose that . . . some mild requirements . . . Then the following probability bound for case II holds*

$$P\left(\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk}\right| \geq \epsilon\right) \leq 8\exp\left(2\log d - \frac{\sqrt{n}\epsilon^2}{(64\ L\ C_\gamma\ l_{\max}\ \pi)^2\log n}\right)$$
$$+ 8\exp\left(2\log d - \frac{n\epsilon^2}{(4L\ C_\gamma)^2\ 128(1+4c^2)^2}\right)$$
$$+ 8\exp\left(2\log d - \frac{\sqrt{n}}{8\pi\log n}\right) + 4\exp\left(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\right) + \frac{2}{\sqrt{\pi\log(nd_2)}}.$$

Outline

Illustration



Difference graph between the true underlying graph, the latent oracle (left) and `hume` (right). Red indicates false negtives and gray false positives

References I

E. J. Bedrick. A comparison of generalized and modified sample biserial correlation estimators. *Psychometrika*, 57(2):183–201, 1992. ISSN 0033-3123. doi: 10.1007/BF02294504. URL `https://doi-org.eaccess.ub.tum.de/10.1007/BF02294504`.

J. Cheng, T. Li, E. Levina, and J. Zhu. High-dimensional mixed graphical models. *J. Comput. Graph. Statist.*, 26(2):367–378, 2017. ISSN 1061-8600. doi: 10.1080/10618600.2016.1237362. URL `https://doi-org.eaccess.ub.tum.de/10.1080/10618600.2016.1237362`.

J. Fan, H. Liu, Y. Ning, and H. Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(2):405–421, 2017. ISSN 13697412, 14679868. URL `http://www.jstor.org/stable/44682518`.

References II

R. Foygel and M. Drton. Extended bayesian information criteria for Gaussian graphical models. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 604–612. Curran Associates, Inc., 2010. URL `https://proceedings.neurips.cc/paper/2010/file/072b030ba126b2f4b2374f342be9ed44-Paper.pdf`.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. ISBN 0-19-852219-3. Oxford Science Publications.

S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17(1):31–57, 1989. ISSN 0090-5364. doi: 10.1214/aos/1176347003. URL `https://doi-org.eaccess.ub.tum.de/10.1214/aos/1176347003`.

References III

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009. ISSN 1532-4435.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1037. URL `https://doi-org.eaccess.ub.tum.de/10.1214/12-AOS1037`.

U. Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979. ISSN 0033-3123. doi: 10.1007/BF02296207. URL `https://doi-org.eaccess.ub.tum.de/10.1007/BF02296207`.

K. Pearson. I. mathematical contributions to the theory of evolution.—vii. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*,

References IV

195(262-273):1–47, 1900.

K. Pearson. On a new method of determining correlation between a measured character a, and a character b, of which only the percentage of cases wherein b exceeds (or falls short of) a given intensity is recorded for each grade of a. *Biometrika*, 7(1/2):96, 1909. doi: 10.2307/2345365. URL `https://doi.org/10.2307/2345365`.