# Learning extremal graphical models in high dimensions

Sebastian Engelke[1] **Michaël Lalancette**[2] Stanislav Volgushev[2]

[1]Research Center for Statistics, University of Geneva

[2]Department of Statistical Sciences, University of Toronto

SSC annual meeting, June 3, 2022

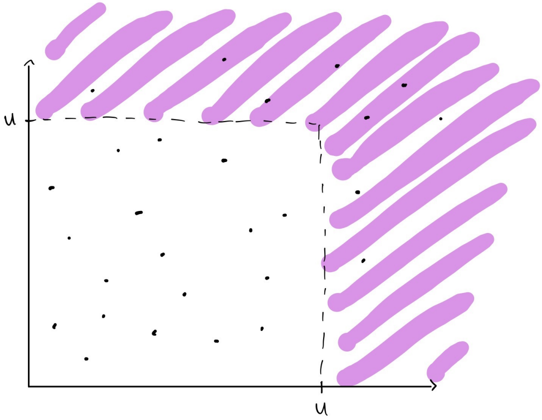
UNIVERSITY OF TORONTO

# Tail (or extremal) dependence

- Random vector $\boldsymbol{X} \in \mathbb{R}^d$
- Tail dependence can be defined as the dependence structure of $\boldsymbol{X}$ in extreme regions/conditional on an extreme event
- Extreme events:

$$\{X_1 > u\} \quad \text{or} \quad \{\max X_i > u\} \quad \text{or} \quad \{\min X_i > u\}$$

# Tail dependence: illustration



# Multivariate Pareto distributions

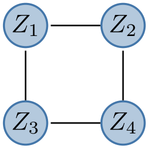
- Suppose that

$$\mathbb{P}\big(F(\boldsymbol{X}) \leq 1 - q/\boldsymbol{x} \mid \max_i F_i(X_i) > 1 - q\big) \longrightarrow \mathbb{P}(\boldsymbol{Y} \leq \boldsymbol{x}), \quad q \downarrow 0,$$

  where $F(\boldsymbol{X}) := (F_1(X_1), \dots, F_d(X_d))$
- "Given that at least one component of $\boldsymbol{X}$ exceeds it's $(1-q)$th quantile, $q/(1 - F(\boldsymbol{X})) \approx \boldsymbol{Y}$ in distribution"
- Then the random vector $\boldsymbol{Y} \in \mathbb{R}^d$ satisfies
  1. $\boldsymbol{Y} \in \mathcal{L} := \{\boldsymbol{y} \geq 0 : \|\boldsymbol{y}\|_\infty > 1\}$
  2. $\mathbb{P}(Y_1 > 1) = \cdots = \mathbb{P}(Y_d > 1)$
  3. For $A \subset \mathcal{L}$ and $t \geq 1$, $\mathbb{P}(\boldsymbol{Y} \in tA) = t^{-1}\mathbb{P}(\boldsymbol{Y} \in A)$
- $\boldsymbol{Y}$ is *multivariate Pareto* (MP)
- $\boldsymbol{X}$ is in the *domain of attraction* of $\boldsymbol{Y}$

# Graphical models

- $\boldsymbol{Z} = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ a random vector indexed by $V := \{1, \ldots, d\}$




- $G := (V, E)$ an undirected graph
- $\boldsymbol{Z}$ is a graphical model on $G$ if for each pair $(i, j)$,

$$Z_i \perp Z_j \mid \boldsymbol{Z}_{\setminus\{i,j\}} \iff (i, j) \notin E$$

- Why this is important: if $\boldsymbol{Z}$ has a positive density/mass on a product space, its density/mass can be factorized over the *cliques* of $G$
- Requires knowledge of the graph $\Longrightarrow$ Learning graphical models

# Gaussian graphical models

- If $\boldsymbol{Z} \sim \mathcal{N}(\mu, \Sigma)$, $\Theta := \Sigma^{-1}$,

$$Z_i \perp Z_j \,|\, \boldsymbol{Z}_{\setminus\{i,j\}} \iff \Theta_{ij} = 0$$

- Graph structure is entirely encoded into the zero pattern of $\Theta$
- Sparse estimation of $\Theta \implies$ Estimation of $G$

# Sparse estimation of precision matrices

- Easy to estimate the covariance matrix $\Sigma$ by the sample covariance $\widehat{\Sigma}$
- But if $n < d$, $\widehat{\Sigma}^{-1}$ does not exist (certainly not sparse)
- Many algorithms turn an estimate of $\Sigma$ into an estimate of the zero pattern of $\Sigma^{-1}$:

$$\mathcal{A}(\widehat{\Sigma}) = \widehat{\mathbb{1}}\{\Theta \neq 0\}$$

- Call $\mathcal{A}$ a *base learner*
- Examples:
  - Neighborhood selection (Meinshausen & Bühlmann, 2006, Ann. Stat.)
  - Graphical lasso (Yuan & Lin, 2007, Biometrika)

# This talk

1. Do graphical models make sense for MP distributions?

   Yes, but need a different notion of conditional independence

2. Given data from $\boldsymbol{X}$ in the domain of attraction $\boldsymbol{Y}$, can we learn the graph structure of $\boldsymbol{Y}$?

   Yes, for a certain parametric model

# Extremal graphical models

- $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ a MP indexed by $V := \{1, \ldots, d\}$ with positive density
- Support $\neq$ product space
- We say that $Y_i \perp_e Y_j \mid \boldsymbol{Y}_{\setminus\{i,j\}}$ if for some $m \notin \{i,j\}$,

$$Y_i \perp Y_j \mid \{\boldsymbol{Y}_{\setminus\{i,j\}}, Y_m > 1\}$$

- $G := (V, E)$ an undirected graph
- $\boldsymbol{Y}$ is an *extremal graphical model* on $G$ if for each pair $(i,j)$,

$$Y_i \perp_e Y_j \mid \boldsymbol{Y}_{\setminus\{i,j\}} \iff (i,j) \notin E$$

- Engelke & Hitz (2020, JRSSB) show that this definition leads to density factorization

# Hüsler–Reiss distributions

- A family of MP distributions, parametrized by an extremal variogram matrix $\Gamma \in \mathbb{R}^{d \times d}$
- If $\boldsymbol{Y} \sim \mathrm{HR}(\Gamma)$,

$$\Gamma_{ij} = \mathbb{V}\mathrm{ar}(\log Y_i - \log Y_j \mid Y_m > 1)$$

- Density: complicated function of $\Gamma$

# Estimating Hüsler–Reiss distributions: the empirical variogram

- $\boldsymbol{X}$ in the domain of attraction of $\boldsymbol{Y} \sim \text{HR}(\Gamma)$, iid data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim \boldsymbol{X}$
- For $m \in V$, estimate $\Gamma_{ij}$ by

$$\widehat{\Gamma}_{ij}^{(m)} := \widehat{\mathbb{V}\text{ar}}\Big( \log(1 - \widetilde{F}_i(X_{ti})) - \log(1 - \widetilde{F}_j(X_{tj})) \mid \widetilde{F}_m(X_{tm}) > 1 - k/n \Big),$$

where $k$ large, $k/n$ small, $\widetilde{F}_i$ are empirical df

- $\widehat{\Gamma} := d^{-1} \sum_{m=1}^{d} \widehat{\Gamma}^{(m)}$

**Theorem (Engelke, L. & Volgushev, 2021)**

*Under (mild) assumptions, with probability at least* $1 - \delta$,

$$\big\| \widehat{\Gamma} - \Gamma \big\|_\infty \lesssim \left( \frac{k}{n} \right)^{\xi} (\log(n/k))^2 + \sqrt{\frac{\log d + \log \frac{1}{\delta}}{k}}.$$

# HR graphical models

- If $\boldsymbol{Y} \sim \mathrm{HR}(\Gamma)$, Engelke & Hitz (2020, JRSSB) find that for $m \notin \{i, j\}$,

$$Y_i \perp_e Y_j \mid \boldsymbol{Y}_{\setminus\{i,j\}} \iff \Theta^{(m)}_{ij} = 0,$$

  where $\Theta^{(m)}$ is the (pseudo)inverse of

$$\Sigma^{(m)} := (\Gamma_{im} + \Gamma_{jm} - \Gamma_{ij})_{i,j \in V}, \quad m \in V$$

- Extremal graph structure is encoded into the zero pattern of the matrices $\Theta^{(m)}$
- Estimate the sparsity pattern of the $\Theta^{(m)}$ and combine them through majority voting

# $\texttt{EGlearn}$: learning HR graphical models

- For $m \in V$,
  1. Compute
  $$\widehat{\Sigma}^{(m)} := (\widehat{\Gamma}_{im} + \widehat{\Gamma}_{jm} - \widehat{\Gamma}_{ij})_{i,j \in V}, \quad m \in V$$
  2. Throw $\widehat{\Sigma}^{(m)}$ into a base learner $\mathcal{A}$ to obtain a sparse estimate $\widehat{\mathbb{1}}\{\Theta^{(m)} \neq 0\}$
- For each pair $(i, j)$, add an edge to $\widehat{E}$ if and only if
$$\frac{1}{d-2} \# \Big\{ m \in V \setminus \{i, j\} : \widehat{\mathbb{1}}\{\Theta_{ij}^{(m)} \neq 0\} = 1 \Big\} > \frac{1}{2}$$
- Graph estimate $\widehat{G} := (V, \widehat{E})$

# EGlearn: illustration

$$
\begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & 1 \\ \cdot & 0 & \cdot & 1 \\ \cdot & 1 & 1 & \cdot \end{pmatrix}
\quad
\begin{pmatrix} \cdot & \cdot & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & 1 \\ 1 & \cdot & 1 & \cdot \end{pmatrix}
\quad
\begin{pmatrix} \cdot & 1 & \cdot & 0 \\ 1 & \cdot & \cdot & 1 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 1 & \cdot & \cdot \end{pmatrix}
\quad
\begin{pmatrix} \cdot & 1 & 1 & \cdot \\ 1 & \cdot & 1 & \cdot \\ 1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}
$$

Figure: Estimated sparsity pattern of $\Theta^{(m)}$, $m = 1, 2, 3, 4$

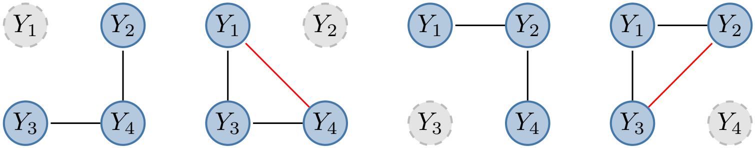



Figure: Corresponding votes

# EGlearn: model selection consistency

**Theorem (Engelke, L. & Volgushev, 2022+)**

*If $\mathcal{A}$ is neighborhood selection or graphical lasso, under assumptions,*

$$\mathbb{P}(\widehat{G} = G) \longrightarrow 1$$

*as long as* $\log d = o(k/(\log k)^8)$.

# Selected references

## Extremal graphical models

Engelke, S. and A. S. Hitz (2020). Graphical models for extremes (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol. 82*, 871–932.

Engelke, S. and S. Volgushev (2021). Structure learning for extremal tree models. *arXiv preprint arXiv:2012.06179*.

Engelke, S., M. Lalancette and S. Volgushev (2022+). Learning extremal graphical models in high dimensions. *In preparation.*

## Gaussian graphical models and sparse precision matrix estimation

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34(3)*, 1436–1462.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika 94(1)*, 19–35.

# Summary

- Extremal graphical models allow lower dimensional representation of extremal dependence structure
- In the HR parametric family, they can be learned from data even in exponentially high dimension
- We do so using majority voting combined with Gaussian graphical modeling tools
- Preprint out very soon
  - Complete methodology + extensions
  - Theoretical justifications + proofs
  - Simulation studies
  - Application
- mic-lalancette.github.io

**Thank you for your attention!**