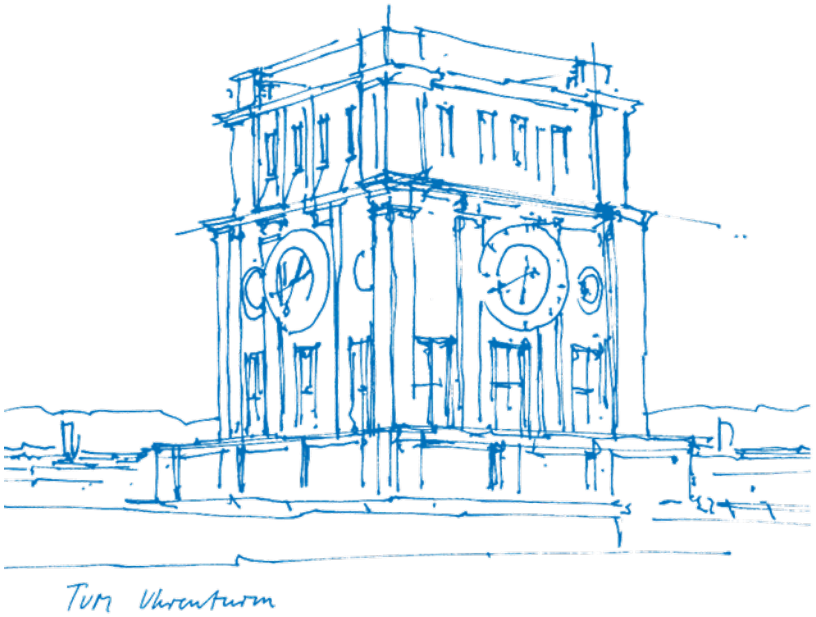
Model Selection for Graphical Continuous Lyapunov Models

Philipp Dettling

Department of Mathematics
Technical University of Munich (TUM)

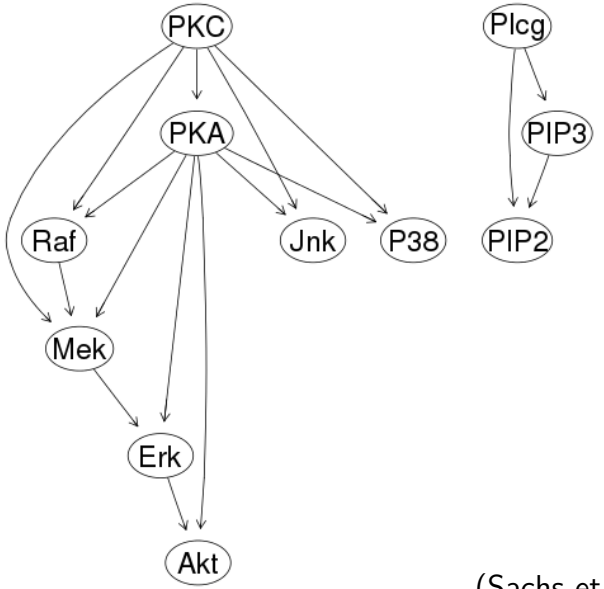
(with Mathias Drton and Mladen Kolar)



Causal Discovery

Given multivariate data, estimate underlying causal structure.

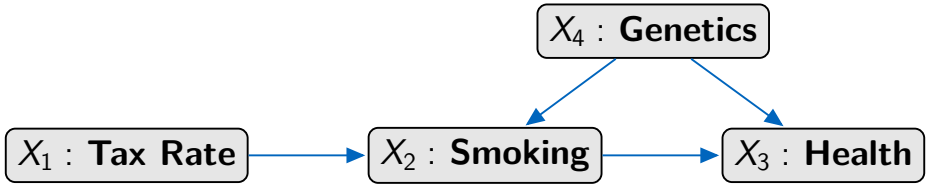
2	26.4	13.2	8.82	18.3	58.8	6.61	17	414
3	35.9	16.5	12.3	16.8	8.13	18.6	32.5	352
4	59.4	44.1	14.6	10.2	13	14.9	32.5	403
5	73	82.8	23.1	13.5	1.29	5.83	11.8	528
6	33.7	19.8	5.19	9.73	24.8	21.1	46.1	305
7	18.8	3.75	17.6	22.1	10.9	11.9	25.7	610
8	44.9	36.5	10.4	132	16.3	8.66	17.9	835
9	47.4	15	14.6	30.5	17.5	20.2	45.3	466
10	104	61.5	10.6	21.1	41.8	11.5	23.5	445
11	21.1	21.5	1.88	205	43.7	13.2	135	213
12	16.4	16.4	14.5	17	11.2	21.9	34.6	449
13	74.3	22.9	7.5	15.5	26.2	20.9	36.5	389
14	85.1	39.6	8.9	64.9	11.7	6.67	12.2	528
15	36.8	29.2	5	9.06	15.5	17.9	17.9	400



(Sachs et al., 2005)

Dominant Approach: Structural Causal/Equation Models

(the 'usual' graphical models)



Noisy functional relationships:

$$
\begin{aligned}
X_1 &= f_1(\varepsilon_1), \\
X_2 &= f_2(X_1, X_4, \varepsilon_2), \\
X_3 &= f_3(X_2, X_4, \varepsilon_3), \\
X_4 &= f_4(\varepsilon_4),
\end{aligned}
$$

Often, linear relationships:

$$
\begin{aligned}
X_1 &= \lambda_{01} + \varepsilon_1, \\
X_2 &= \lambda_{02} + \lambda_{12} X_1 + \lambda_{42} X_4 + \varepsilon_2, \\
X_3 &= \lambda_{03} + \lambda_{23} X_2 + \lambda_{43} X_4 + \varepsilon_3, \\
X_4 &= \lambda_{04} + \varepsilon_4.
\end{aligned}
$$

Noise terms are "just noise": $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2 \perp\!\!\!\perp \varepsilon_3 \perp\!\!\!\perp \varepsilon_4$.

Motivation

- DAGs (directed acyclic graphs) very well understood:
  - simple interpretation
  - scalable statistically and computationally
  - solid theory: characterization of model equivalence ($X \to Y$ versus $X \leftarrow Y$), ...
- Feedback loops $\equiv$ directed cycles:
  - far more complicated model geometry
  - statistics tricky
  - interpretation less clear
- Attempts to interpret directed cycles typically appeal to equilibria of temporal processes in post-hoc way.

Alternative: Immediately consider models derived from a temporal process in equilibrium.

Graphical Continuous Lyapunov Models

Varando & Hansen (2020, UAI) and Fitch (2019, arXiv)

- $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ i.i.d. sample with

  $$X^{(i)} \sim \text{equilibrium distribution of a multivariate Ornstein-Uhlenbeck processes}$$

- Ornstein-Uhlenbeck process $X(t)$ solves stochastic differential equation

  $$\mathrm{d}X(t) = M\,(X(t) - \mu)\,\mathrm{d}t + D\,\mathrm{d}W(t),$$

  where $W(t)$ is a Wiener process.

  Parameters: $\mu \in \mathbb{R}^p$ and $M, D \in \mathbb{R}^{p \times p}$ non-singular.

- Key object: Drift matrix M captures relations between the coordinates of $X(t)$

Continuous Lyapunov Equation

- If M is stable, $X(t)$ admits a Gaussian equilibrium distribution $N(\mu, \Sigma)$ with covariance matrix $\Sigma \in \mathrm{PD}_p$ given by the continuous Lyapunov equation:

$$M\Sigma + \Sigma M^\top = -C \qquad \text{where} \quad C = DD^\top \in \mathrm{PD}_p.$$

- We assume the volatility matrix C to be known up to a positive scalar multiple (e.g., $C = \gamma I_p$).

- For $\gamma > 0$, we have

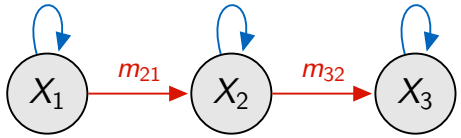
$$M\Sigma + \Sigma M^\top = -C \iff \gamma M\Sigma + \Sigma \gamma M^\top = -\gamma C.$$

  Hence, (M, C) and $(\gamma M, \gamma C)$ define the same covariance matrix Σ.

- Going forward, we treat C as known.

Graphical Continuous Lyapunov Models

- Support of drift matrix M corresponds to a directed graph:



$$M = \begin{pmatrix} m_{11} & 0 & 0 \\ m_{21} & m_{22} & 0 \\ 0 & m_{32} & m_{33} \end{pmatrix}.$$

  Formally, $G = (V, E)$ with $V = \{1, \ldots, p\}$ and $i \to j \in E$ when $M_{ji} \neq 0$.

- Sparse stable matrices:

$$\mathrm{Stab}_p(E) \;=\; \left\{M \in \mathbb{R}^{p \times p} : M \text{ stable},\; M_{ji} = 0 \text{ if } i \to j \notin E\right\}$$

- Associated normal distributions form the graphical continuous Lyapunov model of G, which corresponds to the cone

$$\mathcal{M}_G \;=\; \{\Sigma \in \mathrm{PD}_p : \Sigma \text{ solves Lyapunov equation for some } M \in \mathrm{Stab}_p(E)\}$$

Problems We Studied So Far ...

1. Parameter Identifiability (not today)

- Is M uniquely determined by the covariance matrix Σ?
- Is M uniquely determined by the cov. matrix Σ if we know $M \in \mathrm{Stab}_p(E)$ for graph $G = (V, E)$ and for any (diagonal) $C \in \mathrm{PD}_p$?

  (Mapping $M \mapsto \Sigma$ injective on $\mathrm{Stab}_p(E)$?)
- Identifiability in Continuous Lyapunov Models; arXiv preprint 2022; D, Homs, Améndola, Drton, Hansen.

2. Estimation/Model selection

- Direct Lasso:

$$\min_{M \in \mathbb{R}^{p \times p}} \frac{1}{2} \| M\hat{\Sigma} + \hat{\Sigma} M^T + C \|_F^2 + \lambda \| M \|_1$$

- On the Lasso for Graphical Continuous Lyapunov Models, arXiv preprint 2022, D, Drton, Kolar.

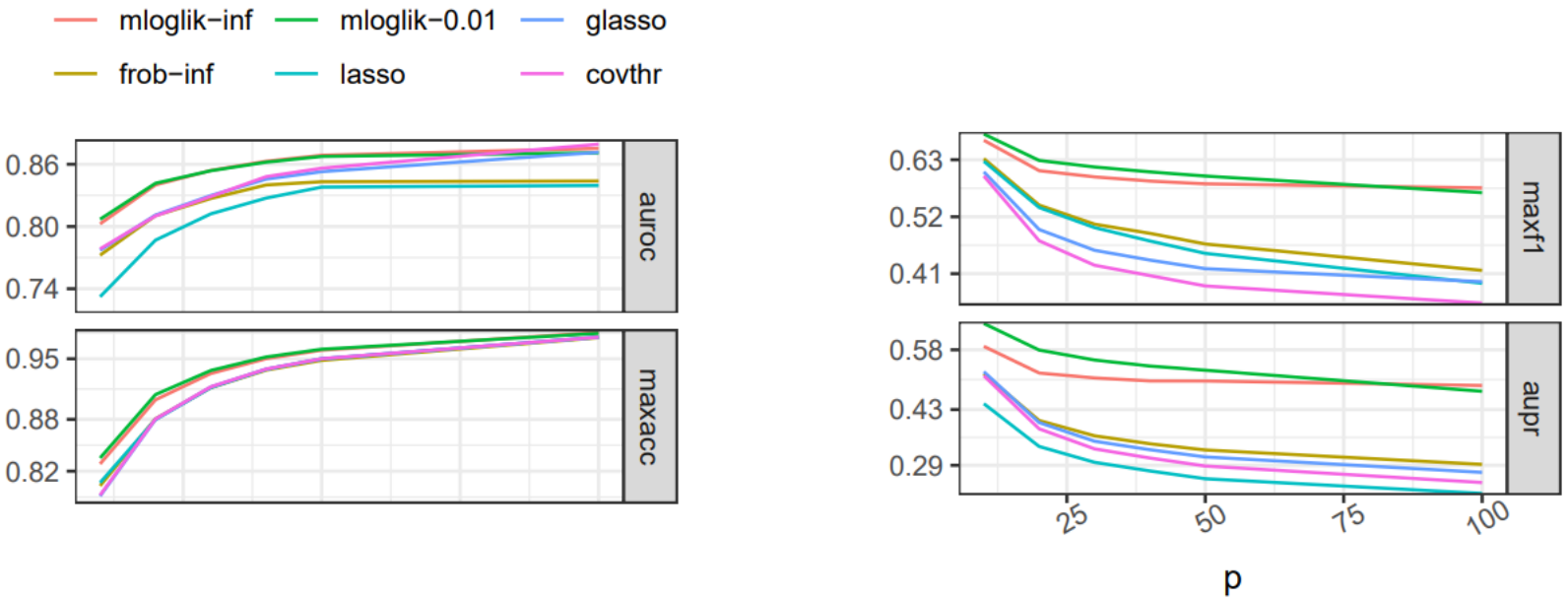
Direct Lasso for Estimation/Model Selection

- Sample covariance matrix $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X^{(i)} (X^{(i)})^\top$
- Direct Lasso (Fitch, 2019):

$$\min_{M \in \mathbb{R}^{p \times p}} \frac{1}{2} \| M \hat{\Sigma} + \hat{\Sigma} M^\top + C \|_F^2 + \lambda \| M \|_1.$$

- Goal (for now): Support recovery
  - True signal M^* with support $S \equiv S(M^*) = \{(j,k) : M^*_{jk} \neq 0\}$.
  - Estimate $\hat{M}$ with support $\hat{S} \equiv S(\hat{M}) = \{(j,k) : \hat{M}_{jk} \neq 0\}$

Some Simulations from Hansen and Varando (2020)



Just another Lasso Problem

- Vectorized version:

$$\min_{M \in \mathbb{R}^{p \times p}} \frac{1}{2} \|A(\hat{\Sigma})\mathrm{vec}(M) + \mathrm{vec}(C)\|_2^2 + \lambda \|\mathrm{vec}(M)\|_1,$$

  with 'design matrix'

$$A(\hat{\Sigma}) \in \mathbb{R}^{p^2 \times p^2}.$$

- Quadratic form written out:

$$\min_{M \in \mathbb{R}^{p \times p}} \frac{1}{2} \mathrm{vec}(M)^\top \Gamma(\hat{\Sigma})\mathrm{vec}(M) - g(\hat{\Sigma})^\top \mathrm{vec}(M) + \lambda \|\mathrm{vec}(M)\|_1.$$

  with

$$\text{Gram matrix: } \Gamma(\Sigma) := A(\Sigma)^\top A(\Sigma) \quad \text{and} \quad g(\Sigma) := -A(\Sigma)\mathrm{vec}(C).$$

- Computationally a lasso problem
- Analysis of support recovery via Primal-Dual-Witness method

Support Recovery via PDW

- Support recovery succeeds if Γ^*_{SS} invertible + irrepresentability + beta-min.
- For a probabilistic guarantee, need to bound in particular

$$\mathbb{P}(\|(\Gamma(\hat{\Sigma}) - \Gamma(\Sigma^*))_{\cdot S}\|_\infty \geq \epsilon_1).$$

- Standard analysis: Give a concentration inequality for each entry of the Gram matrix $\hat{\Gamma}$ + union bound.

  Unfortunately, our Gram matrix here has p^2 of its entries of the type:

$$\Gamma_{(1,1),(2,1)} = 4 \cdot \Sigma_{11} \cdot \Sigma_{21} + \sum_{i=2}^{p} \Sigma_{1i} \cdot \Sigma_{2i}.$$

  Estimates of these entries do not concentrate well ('$p^2 < n$').

- Better through a spectral perspective.

Sample Covariance Matrix

We use a standard result on the spectral norm of the estimation error of the sample covariance matrix.

Theorem

Suppose that $(X^{(i)})_{i=1}^n$ *are* σ *sub-Gaussian random variables. Then the sample covariance matrix* $\hat{\Sigma}$ *satisfies*

$$\mathbb{P}\left(\frac{\|\hat{\Sigma} - \Sigma^*\|_2}{\sigma^2} \geq c_1\left\{\sqrt{\frac{p}{n}} + \frac{p}{n}\right\} + \delta\right) \leq c_2 \exp(-c_3 n \min\{\delta, \delta^2\}) \quad \text{for all } \delta \geq 0,$$

where $\{c_j\}_{j=0}^3$ *are universal constants.*

Spectrum of Gram Matrix

- Gram matrix:

$$\Gamma(\Sigma) = 2(\Sigma^2 \otimes I_p) + (\Sigma \otimes \Sigma)K^{(p,p)} + K^{(p,p)}(\Sigma \otimes \Sigma)$$

- Separately consider

$$\Gamma_1(\Sigma) = 2(\Sigma^2 \otimes I_p) \quad \text{and} \quad \Gamma_2(\Sigma) = (\Sigma \otimes \Sigma)K^{(p,p)} + K^{(p,p)}(\Sigma \otimes \Sigma).$$

- Let $\Delta_\Sigma = \hat{\Sigma} - \Sigma^*$, and for illustration consider Γ_2. Using that
  - commutation matrix $K^{(p,p)}$ is orthogonal with $\|K^{(p,p)}\|_2 = 1$,
  - Kronecker product is bilinear,
  - eigenvalues of Kronecker product are products of eigenvalues, we obtain that

$$\begin{aligned}
\|\Gamma_2(\hat{\Sigma}) - \Gamma_2(\Sigma^*)\|_2 &\leq 2\|\hat{\Sigma} \otimes \hat{\Sigma} - \Sigma^* \otimes \Sigma^*\|_2 \\
&= 2\|\Delta_\Sigma \otimes \Delta_\Sigma + \Delta_\Sigma \otimes \Sigma^* + \Sigma^* \otimes \Delta_\Sigma + \Sigma^* \otimes \Sigma^* - \Sigma^* \otimes \Sigma^*\|_2 \\
&\leq 2\|\Delta_\Sigma \otimes \Delta_\Sigma\|_2 + 2\|\Delta_\Sigma \otimes \Sigma^*\|_2 + 2\|\Sigma^* \otimes \Delta_\Sigma\|_2 \\
&\leq 2\|\Delta_\Sigma\|_2^2 + 4\|\Sigma^*\|_2 \|\Delta_\Sigma\|_2,
\end{aligned}$$

Probabilistic Guarantee

Theorem

Suppose the sample is drawn from a p-dimensional Ornstein-Uhlenbeck process in equilibrium.
Process defined by a true stable drift matrix M^* with support set S of size $d = |S|$.
If Γ^*_{SS} is invertible and the irrepresentability condition holds for $\alpha \in (0, 1]$,
then . . .

In short:

$$n > C\, d \log p$$

ensures that with high prob $(1 - p^\tau)$, for tuning parameter $\lambda = C' \sqrt{d \log p / n}$, the direct lasso

- *has a unique solution $\hat{M}$,*
- *with support $S(\hat{M}) \subseteq S$,*
- *and $\|\hat{M} - M^*\|_\infty \leq C'' \sqrt{(d \log p)/n}$.*

What About Irrepresentability?

- Irrepresentability condition: For $\alpha > 0$,

$$\|\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}\mathrm{sign}(\mathrm{vec}(M^*_S))\|_\infty < 1 - \alpha.$$

- Standard lasso for linear regression, with Gram matrix X^TX:
  Irrepresentability holds in particular if X^TX is close to diagonal ('orthogonal design')
- Natural guess for the Lyapunov problem:
  Irrepresentability holds, in particular, for M^* close to diagonal (when Σ^* is close to diagonal).

Irrepresentability versus Correlation

Theorem

Let $G = (V, E)$ be a simple graph on $V = \{1, \ldots, p\}$. Let M^* be a stable diagonal matrix:

$$M^* = diag(-d_1, \ldots, -d_p).$$

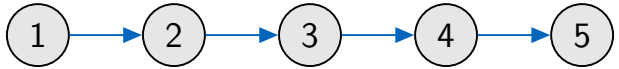
Then irrepresentability holds uniformly in a neighborhood of M^* if and only if

$$d_i < d_j \quad \text{for every edge} \quad i \to j \in E.$$

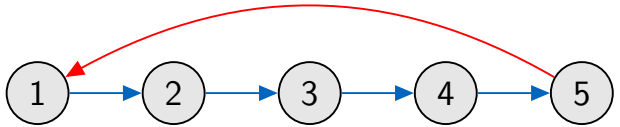
For the condition to hold it is nessecary that the graph is a DAG.

- We do not have a general recipe to construct examples of M^* that satisfy irrepresentability for simple graphs with directed cycles. (Random sampling produces rare cases.)
- Non-simple graphs only trickier as identifiability problems arise at diagonal M^*.

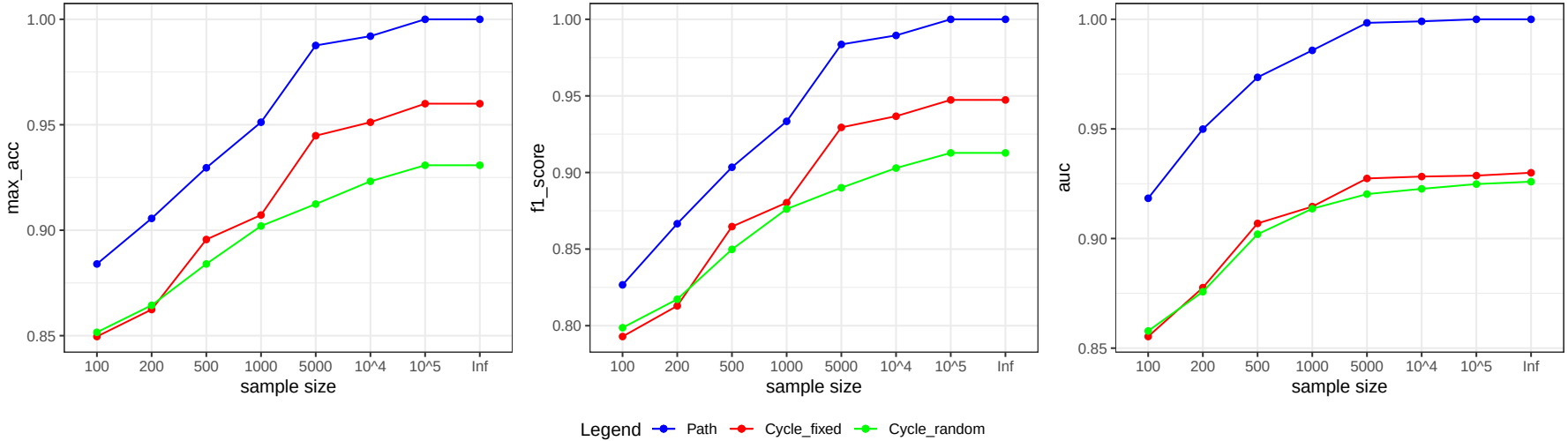
Example



(a) G_1 : a path 1 to 5.



(b) G_2 : the 5-cycle.



Conclusion/Outlook

- Direct lasso is useful, but what about irrepresentability (for cyclic graphs)? ℓ_0?
- Mixed Integer Programming, Trimmed Lasso, promising results
- Better (convex) loss?
- ...