Geometry of faithfulness, sparsest Markov representation assumption and other simplicity assumptions

Caroline Uhler (IST Austria)

Joint work with Garvesh Raskutti, Patrik Norén, Peter Bühlmann, Shaowei Lin, Bernd Sturmfels & Bin Yu

Center for Formal Epistemology
Carnegie Mellon University

June 6, 2014
Overview

- Geometry of $\lambda$-strong faithfulness assumption
  - partial correlation hypersurfaces
  - fixed $\lambda > 0$: lower bounds on tube volumes
  - $\lambda \to 0$: real log canonical threshold

- Learning the ordering of the nodes
  - Sparsest Markov representation condition
  - even permutohedron
Directed Gaussian graphical models

- $G = (V, E)$: DAG with vertices $V = \{1, \ldots, p\}$ and directed edges $E = \{(i, j) \mid i \to j\}$ with $i < j$ for all $(i, j) \in E$

- $a_{ij} \in [-1, 1]$: edge weights, $a_{ij} \neq 0 \iff (i, j) \in E$

- $X_1, \ldots, X_n \in \mathbb{R}^m$: i.i.d. sample from $\mathcal{N}(0, \Sigma)$, where

  $\Sigma^{-1} = (I - A)D^{-1}(I - A^T)$, $A = (a_{ij})$

### 3-node example: Linear structural equation model

- $\epsilon \sim \mathcal{N}(0, D)$ indep. noise $\Rightarrow X \sim \mathcal{N}(0, [(I - A)D^{-1}(I - A^T)]^{-1}$

\[
    \begin{align*}
    X_1 &= \epsilon_1 \\
    X_2 &= a_{12}X_1 + \epsilon_2 \\
    X_3 &= a_{13}X_1 + a_{23}X_2 + \epsilon_3
    \end{align*}
\]
Partial correlation and d-separation

Def: \( i \) is d-separated from \( j \) given \( S \subset V \setminus \{i, j\} \) if every path between \( i \) and \( j \) is blocked by \( S \)

\[
\begin{align*}
&\text{If } (i, j) \notin E \text{ then there exists } S \subset V \setminus \{i, j\} \text{ such that } i \text{ is d-separated from } j \text{ given } S \\
&\text{In the Gaussian setting: } |\text{corr}(i, j|S)| = \frac{\text{det}(K_{iR,jR})}{\sqrt{\text{det}(K_{iR,iR}) \cdot \text{det}(K_{jR,jR})}} \\
&\text{det}(K_{iR,jR}) \text{ is linear combination of all non-blocked paths from } i \text{ to } j \\
&\text{corr}(i, j|S) \equiv 0 \iff i \text{ is d-separated from } j \text{ given } S
\end{align*}
\]
Strong-faithfulness

Def: A multivariate Gaussian distribution $\mathbb{P}$ is $\lambda$-**strongly faithful** to a DAG $G$ for $\lambda \in (0,1)$ if for any $i, j \in V$ and any $S \subset V\setminus\{i,j\}$:

$$|\text{corr}(i,j|S)| \leq \lambda \implies i \text{ is d-separated from } j \text{ given } S.$$ 

Question: What is the proportion of distributions that don’t satisfy the strong-faithfulness condition?

- $\text{Tube}_{\text{corr}(i,j|S)}(\lambda) = \{(a_{ij}) \in [-1, +1]^E : |\text{corr}(i,j|S)| \leq \lambda\}$
- $\text{Vol}_{\text{corr}(i,j|S)}(\lambda) = \int_{\text{Tube}_{\text{corr}(i,j|S)}(\lambda)} \varphi(\omega) \, d\omega$
Unfaithful distributions: 3-node example

\[
\Sigma^{-1} = \begin{pmatrix}
1 + a_{12}^2 + a_{13}^2 & a_{13} a_{23} - a_{12} & -a_{13} \\
 a_{13} a_{23} - a_{12} & 1 + a_{23}^2 & -a_{23} \\
- a_{13} & -a_{23} & 1
\end{pmatrix}
\]

Faithfulness is **NOT** satisfied if any of the following relations hold:

- \( X_1 \perp \perp X_2 \iff a_{12} = 0 \)
- \( X_1 \perp \perp X_3 \iff a_{13} + a_{12} a_{23} = 0 \)
- \( X_2 \perp \perp X_3 \iff a_{12}^2 a_{23} + a_{12} a_{13} + a_{23} = 0 \)
- \( X_1 \perp \perp X_2 \mid X_3 \iff a_{13} a_{23} - a_{12} = 0 \)
- \( X_1 \perp \perp X_3 \mid X_2 \iff -a_{13} = 0 \)
- \( X_2 \perp \perp X_3 \mid X_1 \iff -a_{23} = 0 \)

\( \Rightarrow \) Faithfulness not satisfied on collection of **hypersurfaces** in \( \mathbb{R}^{|E|} \)
Unfaithful distributions: 3-node example

\[ a_{13} + a_{12}a_{23} = 0 \]

\[ a_{12}^2a_{23} + a_{12}a_{13} + a_{23} = 0 \]

\[ a_{13}a_{23} - a_{12} = 0 \]
Unfaithful distributions: 3-node example
Main question

(a) $f(x, y) = x$ 
(b) $f(x, y) = xy$ 
(c) $f(x, y) = x^2y^3$ 
(d) $f(x, y) = x^3y - xy^3$

Problem: Given a DAG $G$, what proportion of distributions does not satisfy strong-faithfulness?

- $\mathcal{P}_{j,k|S}^\lambda = \{(a_{st}) \in [-1, +1]^{|E|} : |\text{corr}(X_j, X_k | X_S)| \leq \lambda \}$
- $\mathcal{M}_{G,\lambda} = \bigcup_{j,k|S \text{ not } d\text{-separated}} \mathcal{P}_{j,k|S}^\lambda$

Problem: Bounds on $\frac{\text{vol}(\mathcal{M}_{G,\lambda})}{2|E|}$; Dependence on $G$?
Proportion of strongly unfaithful distributions increases with $\lambda$ ($\simeq 1/\sqrt{n}$)

Proportion of strongly unfaithful distributions increases with graph density and size
Lower bounds

- **Approach:** Find bounds on surface area of hypersurfaces and thickness

- Finding upper bounds using real algebraic geometry is pretty straight-forward: Crofton’s Formula, Lojasiewicz Inequality, and the Union Bound

- A similar approach using tools from real algebraic geometry does **NOT** work to find non-trivial lower bounds for general graphs

- Hypersurface could consist of complex points only, e.g.

\[ x^2 + y^2 + 1 = 0 \]

\[ \Rightarrow \quad \text{General lower bound is } 0 \]

- More specific information about polynomials is needed

\[ \Rightarrow \quad \text{Analyze different classes of graphs separately} \]
Lower bound for trees

**Theorem (U., Raskutti, Bühlmann & Yu, 2012)**

Let $T_p$ be a connected directed tree on $p$ nodes. The polynomial corresponding to the CI relation $X_i \perp \perp X_j \mid X_S$ is of the form:

$$a_{i \rightarrow j} \cdot (1 + \text{SOS}(a)),$$

where $a_{i \rightarrow j}$ denotes the value of the unique path from $i$ to $j$ and $\text{SOS}(a)$ is a sum of squares polynomial in the variables $a$. Then

$$\frac{\text{vol}(\mathcal{M}_{T_p, \lambda})}{2|E|} \geq 1 - (1 - \lambda)^{p-1}.$$

- Proportion of strongly unfaithful distributions on trees converges to 1 exponentially in $p$
- For high-dimensional consistency: $p_n = o(\sqrt{n/\log(n)})$ for bounded degree trees and $p_n = o((n/\log(n))^{1/3})$ for star-shaped trees
Theorem (U., Raskutti, Bühlmann & Yu, 2012)

Let $C_p$ be a directed cycle on $p$ nodes. The polynomial corresponding to the CI relation $X_i \perp \perp X_j \mid X_S$ is of the form:

$$a_{i \rightarrow j} \cdot \text{SOS}(a) \quad \text{or} \quad f(\bar{a})a_{i,i+1} - g(\bar{a})a_{j,j+1}$$

where $a_{i \rightarrow j}$ denotes the value of a path from $i$ to $j$ and $f(\bar{a}), g(\bar{a})$ are polynomials in $\bar{a} = \{a_{st} \mid (s, t) \notin \{(i, i+1), (j, j+1)\}\}$. Then

$$\frac{\text{vol}(\mathcal{M}_{C_p,\lambda})}{2|E|} \geq 1 - (1 - \lambda)^{p + (p-1)/2}.$$ 

- Proportion of strongly unfaithful distributions on cycles converges to 1 exponentially in $p^2$
- For high-dimensional consistency: $p_n = o((n/\log(n))^{1/4})$
Lower bound for bipartite graphs

Theorem \((U., Raskutti, Bühlmann & Yu, 2012)\)

Let \(K_{2,p-2}\) be a bipartite graph on \(p\) nodes. Besides the \(2(p-2)\) coordinate hyperplanes there are \((p-2)(2^{p-3} - 1)\) distinct hypersurfaces defined by polynomials of the form:

\[ f(\bar{a})a_{i,i+1} - g(\bar{a})a_{j,j+1} \]

Then

\[
\frac{\text{vol}(\mathcal{M}_{K_{2,p-2},\lambda})}{2|E|} \geq 1 - (1 - \lambda)^{(p-2)(2^{p-3}+1)}.
\]

- Proportion of strongly unfaithful distributions on bipartite graphs converges to 1 doubly exponential in \(p\)
- For high-dimensional consistency: \(p_n = o(\log(n))\)
Infinite sample size

(a) $f(x, y) = x$
(b) $f(x, y) = xy$
(c) $f(x, y) = x^2y^3$
(d) $f(x,y) = x^3y − xy^3$

(a) $\text{Vol}(\lambda) = \lambda$

(b) $\text{Vol}(\lambda) = 4 \left( \lambda + \int_{\lambda}^{1} \frac{\lambda}{x} \, dx \right) \frac{1}{4} = \lambda(- \ln \lambda) + \lambda$

(c) $\text{Vol}(\lambda) = 4 \left( \lambda^{1/2} + \int_{\lambda^{1/2}}^{1} \lambda^{1/3}x^{-2/3} \, dx \right) \frac{1}{4} = 3\lambda^{1/3} - 2\lambda^{1/2}$

Watanabe (2009): $\text{Vol}(\lambda) \approx C \lambda^{\ell}(- \ln \lambda)^{m-1}$ asymptotically as $\lambda \to 0$

(a) $(\ell, m) = (1, 1)$, $C = 1$,   (b) $(\ell, m) = (1, 2)$, $C = 1$,
(c) $(\ell, m) = (1/3, 1)$, $C = 3$,   (d) $(\ell, m) = (1/2, 1)$
Real log canonical threshold

Volume of by $\lambda$ fattened-up hypersurface defined by partial correlation is

$$\text{Vol}_{\text{corr}}(i,j\mid S)(\lambda) = \int_{\text{Tube}_{\text{corr}}(i,j\mid S)(\lambda)} \varphi(\omega) d\omega$$

Asymptotically (as $\lambda \to 0$) we have

$$\text{Vol}_{\text{corr}}(i,j\mid S)(\lambda) \approx C \cdot \lambda^{\ell} \cdot (-\ln \lambda)^{m-1}$$

**Def:** The real log canonical threshold (RLCT) of $f = \text{corr}(i,j\mid S)$ is

$$\text{RLCT}_\Omega(f, \varphi) = (\ell, m) \in \mathbb{Q}_+ \times \mathbb{Z}_+,$$

where $(\ell_1, m_1) < (\ell_2, m_2) \iff \ell_1 < \ell_2$ or $\ell_1 = \ell_2$ and $m_1 > m_2$. 
Sparsest Permutation (SP) algorithm

Idea: DAG defined by ordering of vertices (permutation) and skeleton

- For each permutation $\pi$ construct a DAG $G_{\pi} = (V, E_{\pi})$ by
  $$(\pi(i), \pi(j)) \in E_{\pi} \iff X_{\pi(i)} \not\perp \not\| X_{\pi(j)} \mid X_{\{\pi(1), \ldots, \pi(i-1), \pi(i+1), \ldots, \pi(j-1)\}}$$

- Choose permutation $\pi^*$ leading to the sparsest DAG.

- Most parsimonious DAG representation satisfying Markov property

- SP algorithm is cautious in deleting edges:
  $$(i, j) \in E \iff X_i \not\perp \not\| X_j \mid X_S \text{ for all } S \subset V \setminus \{i, j\}$$
Consistency of SP algorithm

**Def:** \((G^*, \mathbb{P})\) satisfies the **sparsest Markov representation (SMR)** assumption if

(i) \((G^*, \mathbb{P})\) satisfies the Markov property,

(ii) \(|S(G)| > |S(G^*)|\) for every DAG \(G\) such that \((G, \mathbb{P})\) satisfies the Markov property and \(G \notin \mathcal{M}(G^*)\).

**Theorem (Raskutti & U., 2013)**

- **SP algorithm is consistent** ⇔ **SMR assumption holds**
- **Faithfulness** ⇒ **SMR assumption**

**Ex:** \(G^*:\)

\[ X_1 \perp X_3 \mid X_2, \ X_2 \perp X_4 \mid X_1, X_3, X_1 \perp X_2 \mid X_4 \]
Relation to other minimality assumptions

**Def:** $(G^*, P)$ satisfies the **SGS-minimality assumption** if there exists no proper sub-DAG of $G^*$ that satisfies the Markov assumption with respect to $P$.

**Def:** $(G^*, P)$ satisfies the **P-minimality assumption** if every DAG $G \neq G^*$ that satisfies the Markov property with respect to $P$ entails a strict super-set of CI statements compared to the DAG $G^*$.

**Theorem (Raskutti & U., 2013)**

- **SMR assumption** $\implies$ **SGS-minimality**
- **SMR assumption** $\implies$ **P-minimality**

**Hence:** Faithfulness $\implies$ SMR assumption $\implies$ P-minimality $\implies$ SGS-minimality
Uniform consistency of SP algorithm

**Inferred CI relations:** \( \Omega_\lambda(P) = \{(i, j, S) \mid |\text{corr}(i, j \mid S)| \leq \lambda\} \)

**Def:** \((G^*, P)\) satisfies the \(\lambda\)-strong faithfulness assumption if \(G^*\) satisfies the faithfulness assumption w.r.t. \(\Omega_\lambda(P)\).

**Def:** \((G^*, P)\) satisfies the \(\lambda\)-strong SMR assumption if

1. \(G^*\) satisfies the Markov property w.r.t. \(\Omega_\lambda(P)\),
2. \(|S(G)| > |S(G^*)|\) for every DAG \(G\) such that \(G\) satisfies the Markov property w.r.t. \(\Omega_\lambda(P)\) and \(G \notin \mathcal{M}(G^*)\).

**Theorem (Raskutti & U., 2013)**

*SP algorithm is uniformly consistent under the \(\lambda\)-strong SMR assumption, which is strictly weaker than the \(\lambda\)-strong faithfulness assumption.*
\( \lambda \)-strong faithfulness for DAGs

- \( \alpha \) is the significance level used for the \( z \)-test
- \( \alpha \) was chosen to optimize the performance of the PC-algorithm
SP algorithm in Gaussian setting

Model: \( X \sim \mathcal{N}(0, \Sigma) \), where \( \Sigma^{-1} = (I - A)D(I - A^T) \) with \( D \) diagonal, \( U = (I - A) \) upper triangular

\[ \implies \Sigma^{-1} = UDU^T \text{ upper Cholesky decomposition} \]

Theorem (Pourahmadi, 1999)

Let \( (\Sigma^{-1})^\pi = U^\pi D^\pi U^\pi T \) denote the upper Cholesky decomposition of the permuted inverse covariance matrix \( (\Sigma^{-1})^\pi \). Then

\[ U^\pi_{ij} = 0 \iff X^\pi(i) \perp \perp X^\pi(j) \mid X^{\{\pi(1), \ldots, \pi(i-1), \pi(i+1), \ldots, \pi(j-1)\}}. \]

\[ \implies \text{SP algorithm in Gaussian setting boils down to finding \textit{sparsest Cholesky decomposition}:} \]

\[ \arg\min_{\pi} \| U^\pi \|_0 \quad \text{subject to} \quad (\Sigma^{-1})^\pi = U^\pi D^\pi U^\pi T. \]
Markov equivalence relation on permutations

Given a set of CI relations, two permutations $\pi$ and $\tau$ have the same sparsity pattern in the Cholesky factor if

- $\tau = (\pi(2), \pi(1), \pi(3), \ldots, \pi(p))$

- $\tau = (\pi(1), \ldots \pi(i - 1), \pi(i + 1), \pi(i), \pi(i + 2), \ldots, \pi(p))$ and $X_{\pi(i)} \perp \perp X_{\pi(j)} \mid X_{\{\pi(1), \ldots, \pi(i - 1), \pi(i + 1), \ldots \pi(j - 1)\}}$

- $\tau = (\pi(3), \pi(1), \pi(2), \pi(4), \ldots, \pi(p))$ and $X_{\pi(1)} \perp \perp X_{\pi(3)} \mid X_{\pi(2)}$
Edges connect neighboring transpositions and represent conditional independence relations, e.g.

$$(3, 1, 4, 2) - (3, 4, 1, 2) : \quad X_1 \perp \perp X_4 \mid X_3.$$
Even permutohedron
Open problems

- How restrictive is strong-SMR assumption?
- Can search be restricted to subset of permutations? How well do heuristics for sparse Cholesky decomposition work (e.g. minimum degree algorithm)?
- Relationship between hierarchies of assumptions for consistency and computational complexity; e.g. polynomial time algorithms? Convex relaxations?
References

- Lin, U., Sturmfels, Bühlmann: Hypersurfaces and their singularities in partial correlation testing (to appear in *Foundations of Computational Mathematics*

Thank you!