On Testing Marginal versus Conditional Independence

Richard Guo
ricguo@uw.edu

Nov, 2019

Department of Statistics, University of Washington, Seattle

Introduction

Motivation

Inferring causal structures usually involves model selection among directed acyclic graphs (DAGs).

While learning undirected graphical models has been relatively well-developed (e.g., graphical lasso, neighborhood selection), model selection for DAGs is less well-understood.

This poses a challenge to maintaining error guarantee in causal inference, even in large samples. In this talk, I will analyze the simplest example where such a challenge arises.

Marginal vs. conditional independence

Consider $(X_1, X_2, X_3)^\intercal \sim \mathcal{N}(0, \Sigma)$ on $\mathbb{R}^3$.

Covariance $\Sigma \in \mathbb{S}^3$, the set of 3×3 real positive definite matrices.

We want to test between

$$\begin{aligned} \mathcal{M}_0 &: X_1 \perp\!\!\!\perp X_2, \qquad && (X_1 \to X_3 \leftarrow X_2), \\ \mathcal{M}_1 &: X_1 \perp\!\!\!\perp X_2 \mid X_3, && (X_1 - X_3 - X_2), \end{aligned}$$

assuming that **at least** one of them is true.

$X_1 - X_3 - X_2$ includes the following Markov-equivalent DAGs

$$X_1 \leftarrow X_3 \leftarrow X_2, \quad X_1 \to X_3 \to X_2, \quad X_1 \leftarrow X_3 \to X_2.$$

Marginal vs. conditional independence

Testing between

$$\mathcal{M}_0 : \ X_1 \perp\!\!\!\perp X_2 \quad \text{vs.} \quad \mathcal{M}_1 : \ X_1 \perp\!\!\!\perp X_2 \mid X_3$$

is a **non-nested** model selection problem.

They correspond to equality/algebraic constraints on $\Sigma = \{\sigma_{ij}\}$:

$$\begin{aligned}
&\mathcal{M}_0 : \sigma_{12} = 0, \\
&\mathcal{M}_1 : \sigma_{12\cdot 3} = \sigma_{12} - \sigma_{13}\sigma_{33}^{-1}\sigma_{23} = 0 \ \Leftrightarrow \ \sigma_{12}\sigma_{33} = \sigma_{13}\sigma_{23}.
\end{aligned}$$

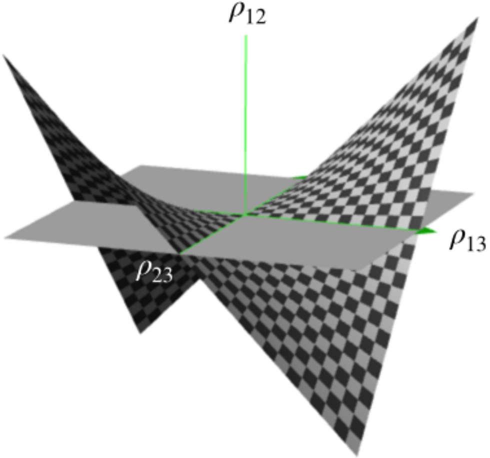
$\mathcal{M}_0$ and $\mathcal{M}_1$ intersect at **the two axes**

$$\mathcal{M}_0 \cap \mathcal{M}_1 = \{\sigma_{12} = \sigma_{13} = 0\} \cup \{\sigma_{12} = \sigma_{23} = 0\}.$$

Geometry

We visualize the parameter space in the correlation space.

$$\mathcal{M}_0 : \rho_{12} = 0, \quad \mathcal{M}_1 : \rho_{12} = \rho_{13}\rho_{23}$$



Singularity

The two axes further intersect at the origin

$$\mathcal{M}_{\text{sing}} : \{\sigma_{12} = \sigma_{13} = \sigma_{23} = 0\},$$

which is a **singularity**. $\mathcal{M}_{\text{sing}}$ corresponds to diagonal Σ.

- $\mathcal{M}_0 \cap \mathcal{M}_1$ vs. $\mathbb{S}^3$: Likelihood-ratio test (LRT) was studied by Drton (2006, 2009) and Drton and Sullivant (2007).
  - LRT has a non-standard asymptotic distribution at $\mathcal{M}_{\text{sing}}$.
- $\mathcal{M}_0$ vs. $\mathcal{M}_1$: At $\mathcal{M}_{\text{sing}}$, the tangent cones of the two models coincide.
  - They are called "1-equivalent" by Evans (2018), meaning that linear approximations to the parameter space are the same.
  - In the Euclidean $m^{-1/2}$-ball of $\mathcal{M}_{\text{sing}}$, m^2 samples are required to distinguish $\mathcal{M}_0$ and $\mathcal{M}_1$.

Difficulty

Model selection for DAGs is usually conducted by the following approaches (Drton and Maathuis, 2017).

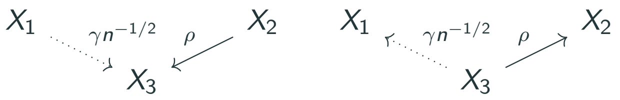
- **Score-based**: Picking the model with the highest penalized likelihood score (e.g., AIC, BIC).
  Since $\dim(\mathcal{M}_0) = \dim(\mathcal{M}_1)$, both AIC and BIC will pick the model with the higher likelihood.
- **Constraint-based**: Testing

$$\mathcal{M}_0 : \; X_1 \perp\!\!\!\perp X_2 \quad \text{vs.} \quad \mathcal{M}_1 : \; X_1 \perp\!\!\!\perp X_2 \mid X_3.$$

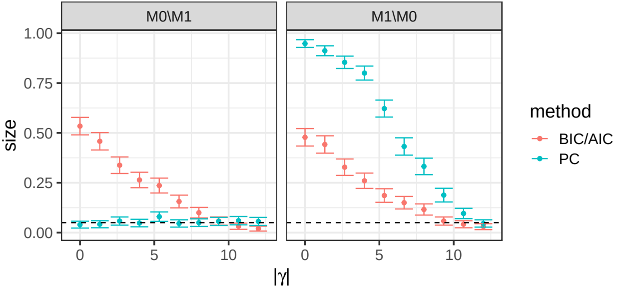
  This is adopted by the PC algorithm. For Gaussian data, Fisher's z-transformation of partial correlation is used as the test statistic.

Difficulty

Simulated with $n = 1,000$, $\rho = 0.3$ and unit variances under level $\alpha = 0.05$.



$\mathcal{M}_0 \setminus \mathcal{M}_1$ $\qquad$ $\mathcal{M}_1 \setminus \mathcal{M}_0$



Method

Likelihood ratio test for nested models

Consider a parametric family $\{P_\theta : \theta \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^d$. For $\Theta_0 \subseteq \Theta$, suppose we want to test

$$\mathcal{H}_0 : \theta \in \Theta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \theta \in \Theta.$$

Under regularity, the likelihood ratio test (LRT) statistic

$$\lambda_n = 2\left(\sup_{\theta} \ell_n(\theta) - \sup_{\theta_0} \ell_n(\theta)\right) \overset{d}{\Rightarrow} \chi^2_c,$$

where $c = d - \dim(\Theta_0)$. $\ell_n(\cdot)$ is the log-likelihood under sample size n.

For example, in linear regression $y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$. We use χ^2_2 for testing

$$\mathcal{H}_0 : \beta_0 = \beta_1 = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \beta \in \mathbb{R}^4.$$

Likelihood ratio test

Similarly, we define the log-likelihood ratio of $\mathcal{M}_0$ versus $\mathcal{M}_1$ as

$$\begin{aligned}\lambda_n^{(0:1)} :=& 2\left(\sup_{\Sigma\in\mathcal{M}_0} \ell_n(\Sigma) - \sup_{\Sigma\in\mathcal{M}_1} \ell_n(\Sigma)\right)\\ =& 2\left(\ell_n(\hat{\Sigma}_n^{(0)}) - \ell_n(\hat{\Sigma}_n^{(1)})\right),\end{aligned}$$

where $\hat{\Sigma}_n^{(0)}$, $\hat{\Sigma}_n^{(1)}$ are MLEs within $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively.

$\ell_n(\cdot)$ is the Gaussian log-likelihood function

$$\ell_n(\Sigma) = \frac{n}{2}(-\log|\Sigma| - \mathbf{Tr}(S_n\Sigma^{-1})).$$

Likelihood ratio test

The Gaussian MLEs for DAGs take a closed form (Drton and Richardson, 2008), which yields the following expression for the LRT.

$$\lambda_n^{(0:1)} = n \log \left(\frac{\left(s_{13}^2 - s_{11}s_{33}\right)\left(s_{23}^2 - s_{22}s_{33}\right)}{s_{33}} \right) - \\ n \log \left(s_{11}s_{22} \left(\frac{s_{22}s_{13}^2 - 2s_{12}s_{23}s_{13} + s_{11}s_{23}^2}{s_{12}^2 - s_{11}s_{22}} + s_{33} \right) \right),$$

where S is the sample covariance taken with respect to mean zero.

Our plan

1. An information-theoretic analysis on how well the two models can be distinguished (by any means).
2. Look at the regimes of "effect size" $\sim n$, such that the optimal error is between 0 and 1.
   - a stable, non-degenerate asymptotic distribution of LRT.
   - We will be doing **large-*n*-small-effect asymptotics**!
3. Derive the asymptotic distributions.
   - Are they uniform?
4. Develop a model selection procedure with error guarantees.

Optimal error

We study the minimax rate of distinguishing two sequences of distributions, one within $\mathcal{M}_0$ and the other within $\mathcal{M}_1$, as they approach $\mathcal{M}_0 \cap \mathcal{M}_1$.

Lemma: testing two simple hypotheses

For testing $H_0 : X \sim P$ versus $H_1 : X \sim Q$, the minimum sum of type-I and type-II errors is $1 - \mathrm{d}_{\mathsf{TV}}(P, Q)$.

Total variation distance

$$\mathrm{d}_{\mathsf{TV}}(P, Q) = \sup_A |P(A) - Q(A)| = \frac{1}{2} \int |p - q| \, \mathrm{d}\mu.$$

Optimal error

Consider a sequence

$$P_n = P_{\Sigma_n^{(0)}}, \quad \Sigma_n^{(0)} \in \mathcal{M}_0 \setminus \mathcal{M}_1, \quad \Sigma_n^{(0)} \to \Sigma^* \in \mathcal{M}_0 \cap \mathcal{M}_1.$$

Correspondingly, let $Q_n = P_{\Sigma_n^{(1)}}$ from $\mathcal{M}_1 \setminus \mathcal{M}_0$ such that

$$\Sigma_n^{(1)} = \underset{\Sigma \in \mathcal{M}_1 \setminus \mathcal{M}_0}{\arg\min}\ \mathcal{D}_{\mathsf{KL}}(P_{\Sigma_n^{(0)}} \| P_\Sigma),$$

which is the **most difficult** to distinguish from.

With $P_n = P_{\Sigma_n^{(0)}}$ and $Q_n = P_{\Sigma_n^{(1)}}$, let us compute the total variation between the product measures (n iid samples).

The limiting optimal error can be sandwiched by the Hellinger distance $H(P, Q) := \left\{\frac{1}{2}\int(\sqrt{p} - \sqrt{q})^2 \,\mathrm{d}\mu\right\}^{1/2}$.

$$H^2(P_n^n, Q_n^n) \leq \mathrm{d}_{\mathsf{TV}}(P_n^n, Q_n^n) \leq H(P_n^n, Q_n^n)\sqrt{2 - H^2(P_n^n, Q_n^n)}.$$

Optimal error

With some algebra, we have

$$1 - \mathrm{d}_{\mathsf{TV}}(P_n^n, Q_n^n) \to \begin{cases} 0, & H(P_n, Q_n) = \omega(n^{-1/2}) \\ 1, & H(P_n, Q_n) = o(n^{-1/2}) \end{cases},$$

and when $H(P_n, Q_n) \asymp n^{-1/2}$,

$$0 < \liminf_n \{1 - \mathrm{d}_{\mathsf{TV}}(P_n^n, Q_n^n)\} \leq \limsup_n \{1 - \mathrm{d}_{\mathsf{TV}}(P_n^n, Q_n^n)\} < 1.$$

Effect size

$$H(P_n, Q_n) \asymp \rho_{13,n}\rho_{23,n},$$

where $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ is the correlation coefficient.

Optimal error

Comparing $H(P_n, Q_n)$ to $n^{-1/2}$, to stabilize the asymptotic error, there are two regimes.

Two regimes

$$\{1 - \mathrm{d}_{\mathsf{TV}}(P_n^n, Q_n^n)\} \to c \in (0,1)$$

$$\text{iff} \begin{cases} \rho_{13,n} \asymp \gamma n^{-1/2}, & \rho_{23,n} \to \rho_{23} \neq 0 \\ \rho_{23,n} \asymp \gamma n^{-1/2}, & \rho_{13,n} \to \rho_{13} \neq 0 \\ \rho_{13,n}\rho_{23,n} \asymp \delta n^{-1/2}, & \rho_{13,n}, \rho_{23,n} \to 0. \end{cases}$$

The first two cases: "**weak-strong**"

The third case: "**weak-weak**"

Asymptotics: weak-strong regime

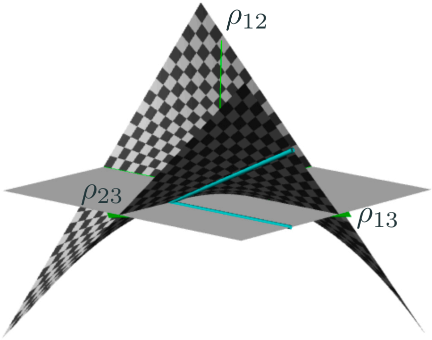
We study the (local) asymptotic distribution of $\lambda_n^{(0:1)}$.

For $r = \gamma\sqrt{\sigma_{11}\sigma_{33}}$, we set

$$\Sigma_n^{(0)} = \begin{pmatrix} \sigma_{11} & 0 & r/\sqrt{n} \\ 0 & \sigma_{22} & \sigma_{23} \\ r/\sqrt{n} & \sigma_{23} & \sigma_{33} \end{pmatrix},$$

$$\Sigma_n^{(1)} = \begin{pmatrix} \sigma_{11} & (r/\sqrt{n})\sigma_{23}/\sigma_{33} & r/\sqrt{n} \\ (r/\sqrt{n})\sigma_{23}/\sigma_{33} & \sigma_{22} & \sigma_{23} \\ r/\sqrt{n} & \sigma_{23} & \sigma_{33} \end{pmatrix},$$

$$\Sigma_n^{(0)}, \Sigma_n^{(1)} \to \Sigma^* = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_{33} \end{pmatrix}$$



$$\begin{tikzcd} X_1 \arrow[dr, dotted, "\gamma n^{-1/2}"] & & X_2 \arrow[dl, "\rho"'] \\ & X_3 & \end{tikzcd}$$

$\Sigma_n^{(0)} \in \mathcal{M}_0 \setminus \mathcal{M}_1$

$$\begin{tikzcd} X_1 & & X_2 \\ & X_3 \arrow[ul, dotted, "\gamma n^{-1/2}"] \arrow[ur, "\rho"'] & \end{tikzcd}$$

$\Sigma_n^{(1)} \in \mathcal{M}_1 \setminus \mathcal{M}_0$

$$\begin{tikzcd} X_1 & & X_2 \\ & X_3 \arrow[ur, "\rho"'] & \end{tikzcd}$$

$\Sigma^* \in \mathcal{M}_0 \cap \mathcal{M}_1$

Asymptotics: weak-strong regime

Let Z_1, Z_2 be two independent standard normals.

LRT in the weak-strong regime

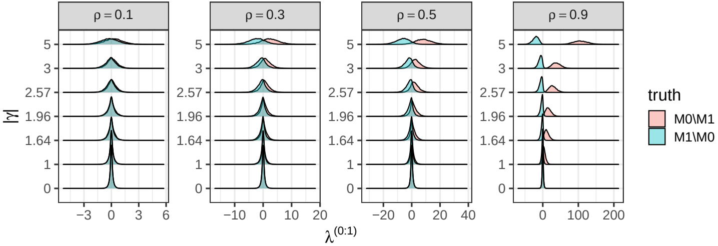
Under $\Sigma_n^{(0)}$,

$$\lambda_n^{(0:1)} \overset{d}{\Rightarrow} \rho \left[\left(Z_1 + \frac{\gamma}{\sqrt{2(1-\rho)}} \right)^2 - \left(Z_2 + \frac{\gamma}{\sqrt{2(1+\rho)}} \right)^2 \right];$$

Under $\Sigma_n^{(1)}$,

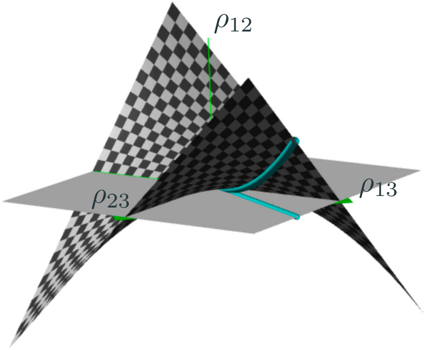
$$\lambda_n^{(0:1)} \overset{d}{\Rightarrow} \rho \left[\left(Z_1 + \gamma \sqrt{\frac{1-\rho}{2}} \right)^2 - \left(Z_2 + \gamma \sqrt{\frac{1+\rho}{2}} \right)^2 \right].$$

Asymptotics: weak-strong regime



The asymptotic distribution is a scaled difference between two independent non-central χ_1^2 variables.

- No simple analytic form for PDF/CDF.
- Adding an $n^{-1/2}$ shift to other elements in Σ_n does not change the distribution (regularity).
- Can be derived from local asymptotic normality (LAN) or Le Cam's 3rd Lemma (change of measure under contiguity).



$$\begin{tikzcd} X_1 \arrow[dr, dotted] & & X_2 \arrow[dl, dotted] \\ & X_3 & \end{tikzcd}$$

$$\Sigma_n^{(0)} \in \mathcal{M}_0 \setminus \mathcal{M}_1$$

$$\begin{tikzcd} X_1 & & X_2 \\ & X_3 \arrow[ul, dotted] \arrow[ur, dotted] & \end{tikzcd}$$

$$\Sigma_n^{(1)} \in \mathcal{M}_1 \setminus \mathcal{M}_0$$

$$\begin{tikzcd} X_1 & & X_2 \\ & X_3 & \end{tikzcd}$$

$$\Sigma^* \in \mathcal{M}_0 \cap \mathcal{M}_1$$

Asymptotics: weak-weak regime

Under the weak-weak regime $\rho_{13,n}\rho_{23,n} = \delta n^{-1/2}$, e.g., $\rho_{13,n} = \sqrt{\delta} n^{-1/3}$ and $\rho_{23,n} = \sqrt{\delta} n^{-1/6}$, the usual tactics fail due to irregularity: (i) $\mathcal{M}_0$ and $\mathcal{M}_1$ cannot be embedded into the same LAN family; (ii) contiguity to an iid static law no longer holds.

P_n^n, Q_n^n contiguous to each other, but neither contiguous to $P_{\Sigma^*}^n$.

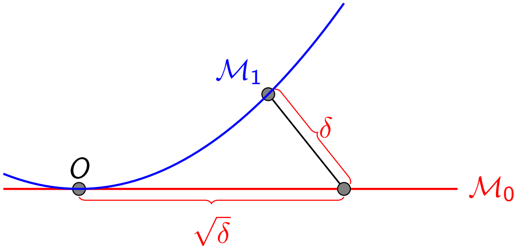


Figure 1: $\mathcal{M}_0$ and $\mathcal{M}_1$ are $\sqrt{\delta}$ away from origin; but they are δ away from each other (Evans, 2018).

Asymptotics: weak-weak regime

Thanks to the closed form of $\lambda_n^{(0:1)}$, by a manual "change of measure" (relating the distribution of sample covariance under $\Sigma_n^{(i)}$ to that under $\Sigma = I$), we obtain a **Gaussian limit**.

LRT in the weak-weak regime

For $\rho_{13,n}\rho_{23,n} = \delta n^{-1/2} + o(n^{-1/2})$,

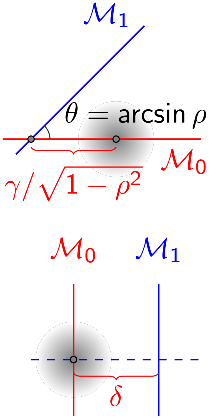
$$\lambda_n^{(0:1)} \overset{d}{\Rightarrow} \begin{cases} \delta(2Z + \delta) =_d \mathcal{N}(\delta^2, (2\delta)^2), & \text{under } \Sigma_n^{(0)} \\ \delta(2Z - \delta) =_d \mathcal{N}(-\delta^2, (2\delta)^2), & \text{under } \Sigma_n^{(1)} \end{cases}.$$

The limit only depends on δ. It does **not** depend on how $\rho_{13,n}$ and $\rho_{23,n}$ approach zero individually.

Limit experiments

Asymptotically, testing between $\mathcal{M}_0$ and $\mathcal{M}_1$ is equivalent to testing **the location of a normal between two lines**, from a single Gaussian observation.

It is characterized by an **angle** and an **intercept**.



Decision, error and power

Due to non-nestedness, we refrain from choosing either as the "null". Instead, we consider a **three-way** decision rule

$$\phi_n : \Sigma_n \to \{\mathcal{M}_0,\ \mathcal{M}_1,\ \mathcal{M}_0 \cup \mathcal{M}_1\}.$$

Size

For all $\Sigma_n \to \Sigma^*$ on $\mathcal{M}_i \setminus \mathcal{M}_{1-i}$ for $i = 0, 1$, control

$$\limsup_{n\to\infty} P_{\Sigma_n}(\phi_n = \mathcal{M}_{1-i}) \leq \alpha.$$

The limit Σ^* could be in $\mathcal{M}_0 \cap \mathcal{M}_1$ or $\mathcal{M}_i \setminus \mathcal{M}_{1-i}$.

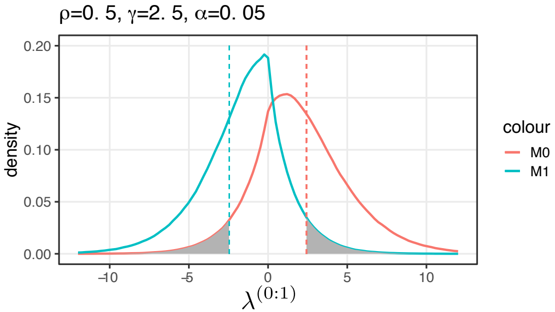
Power

Under $\Sigma_n \to \Sigma^*$ from $\mathcal{M}_i \setminus \mathcal{M}_{1-i}$, power is defined as

$$\liminf_{n\to\infty} P_{\Sigma_n}(\phi_n = \mathcal{M}_i).$$

Decision boundaries from asymptotics

Given the (1) regime, (2) ρ and (3) the local parameter (γ or δ), a three-way decision can be constructed from asymptotic quantiles.



$$\phi_n = \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > F_1^{-1}(1-\alpha) \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < F_0^{-1}(\alpha) \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{otherwise} \end{cases} .$$

Non-uniform asymptotics :(

But this is impossible.

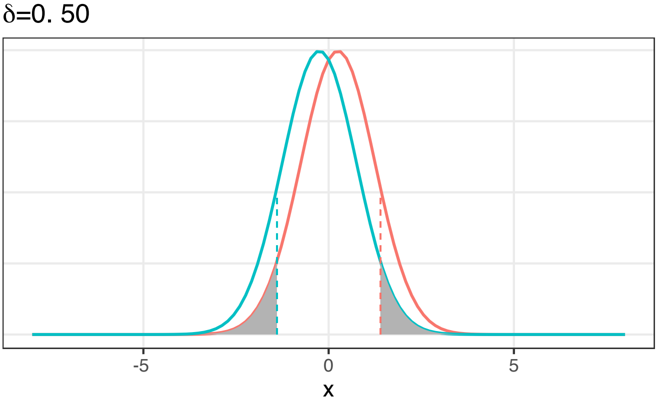
- Depends on the **regime** ("where"): weak-strong or weak-weak.
  - **Discontinuity** across regimes: the law under weak-strong does **not** converge to that of weak-weak when $\rho \to 0$.
- Depends on the **local parameter** γ or δ ("how").
  - Local parameter has scale $n^{-1/2}$, not point-identified.
  - Impossible to judge if an edge is weak based on whether its confidence interval contains zero without further assumptions.
- Further, a procedure that tries to first estimate "where" and "how" before applying the decision rule is susceptible to irregularity issues.

Envelope!

Extremal quantile

Let us look at the weak-weak Gaussian asymptotic as an example.

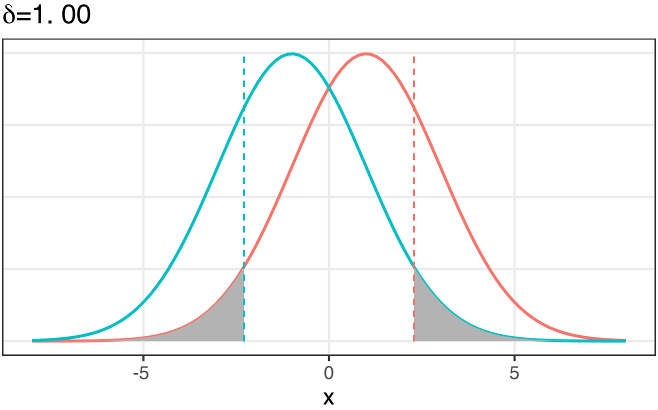
$$F_0^{-1}(\alpha) = (\delta + \Phi^{-1}(\alpha))^2 - \Phi^{-1}(\alpha)^2.$$



Extremal quantile

Let us look at the weak-weak Gaussian asymptotic as an example.

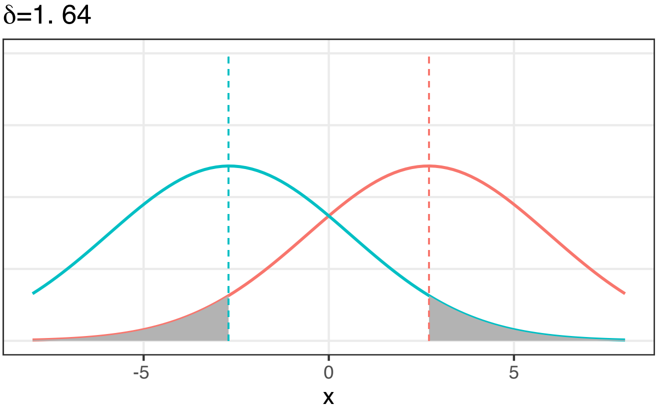
$$F_0^{-1}(\alpha) = (\delta + \Phi^{-1}(\alpha))^2 - \Phi^{-1}(\alpha)^2.$$



Extremal quantile

Let us look at the weak-weak Gaussian asymptotic as an example.

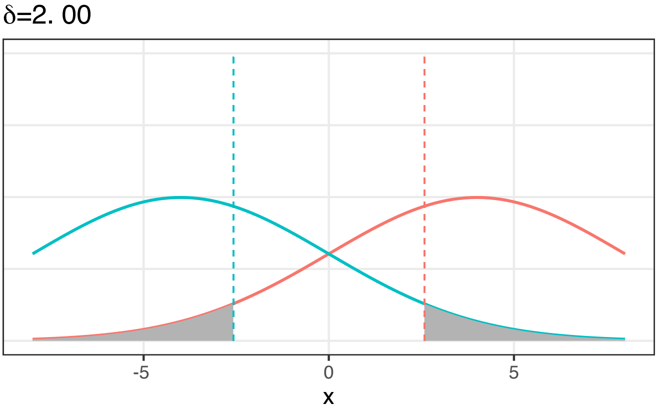
$$F_0^{-1}(\alpha) = (\delta + \Phi^{-1}(\alpha))^2 - \Phi^{-1}(\alpha)^2.$$



Extremal quantile

Let us look at the weak-weak Gaussian asymptotic as an example.

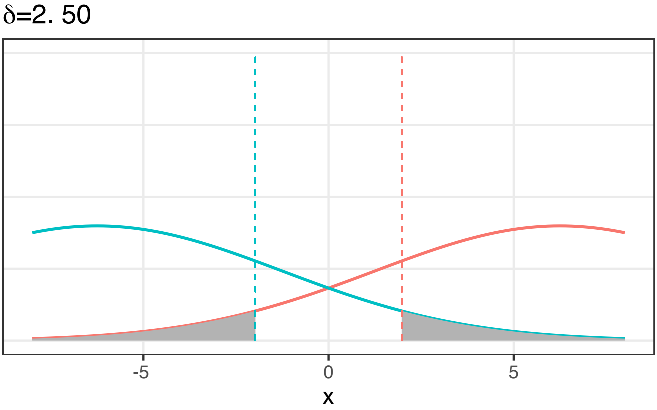
$$F_0^{-1}(\alpha) = (\delta + \Phi^{-1}(\alpha))^2 - \Phi^{-1}(\alpha)^2.$$



Extremal quantile

Let us look at the weak-weak Gaussian asymptotic as an example.

$$F_0^{-1}(\alpha) = (\delta + \Phi^{-1}(\alpha))^2 - \Phi^{-1}(\alpha)^2.$$



Envelope distribution

Taking extremal quantiles for every α is equivalent to taking pointwise supremum of CDF over the local parameter γ or δ.

Envelope distribution

Given a family of distribution functions $\{F_h : h \in \mathcal{H}\}$ on $\mathbb{R}$, define

$$\bar{F}^*(x) := \sup_{h \in \mathcal{H}} F_h(x),$$

and

$$\bar{F}(x) := \begin{cases} \bar{F}^*(x), & \bar{F}^* \text{ is continuous at } x \\ \lim_{y \to x^+} \bar{F}^*(y), & \bar{F}^* \text{ is discontinuous at } x \end{cases}.$$

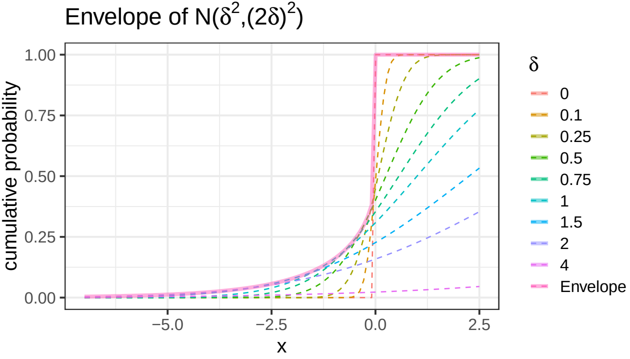
We call $\bar{F}$ the envelope distribution of $\{F_h : h \in \mathcal{H}\}$ if $\bar{F}$ is a valid distribution function.

Envelope distribution

Envelope distribution function

Lemma: If $\bar{F}^*(x) \to 0$ as $x \to -\infty$, then $\bar{F}(x)$ is a valid distribution function.

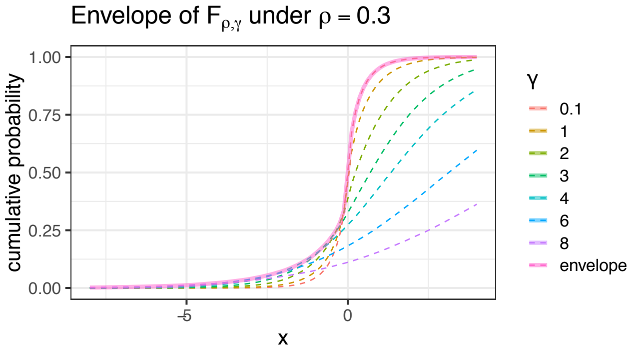
For the weak-weak regime, it can be shown $\bar{F} = \frac{1}{2}(-\chi_1^2) + \frac{1}{2}\delta_0$.



Envelope distribution

The same phenomenon occurs for the weak-strong regime!

We can verify that $\bar{F}_\rho^*(x) \to 0$ as $x \to -\infty$ for every $|\rho| \in (0, 1]$. Therefore, $\bar{F}_\rho$, the envelope of $\{F_{\rho,\gamma} : \gamma \in \mathbb{R}\}$, is a valid distribution function.



Envelope distribution

Continuity of envelope!

Proposition: $\bar{F}_\rho \overset{d}{\Rightarrow} \bar{F}$ as $\rho \to 0$, where $\bar{F}$ is the envelope distribution for the **weak-weak** regime.

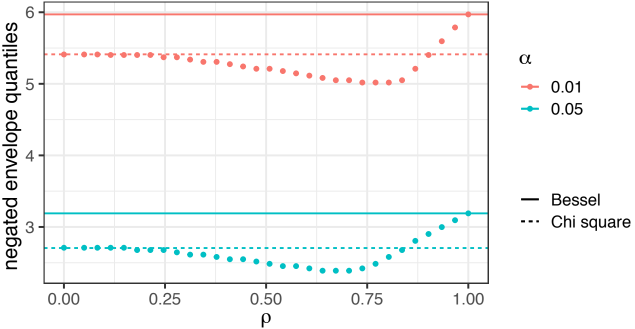
Further, we show the following properties for $\{F_\rho : -1 \leq \rho \leq 1\}$.

- $\bar{F}_\rho = \bar{F}_{|\rho|}$.
- $\bar{F}_\rho$ under $\mathcal{M}_0 \setminus \mathcal{M}_1$ and $\mathcal{M}_1 \setminus \mathcal{M}_0$ have the same form.
- The positive part of $\bar{F}_\rho$ for $|\rho| \in (0, 1]$ is distributed as the positive part of $\rho(Z_1^2 - Z_2^2)$ for two independent standard normals.
- Only the negative part of $\bar{F}_\rho$ is relevant for decision making.
- We do not have an analytic form for the negative part of $\bar{F}_\rho$, except for $\rho \in \{-1, 0, 1\}$.

Envelope quantiles

Quantiles of $\bar{F}_\rho$ can be evaluated by Monte Carlo on a grid of values for ρ and interpolating.

It is interesting to notice that $\bar{F}_\rho^{-1}(\alpha)$ is not monotonic in $|\rho|$.



Model selection procedure: adaptive rule

Note that $\bar{F}_\rho$ is continuous in ρ. Recall that $\rho = \rho_{\text{strong}}$ in the weak-strong regime, and $\rho = 0$ in the weak-weak regime. $|\rho|$ can be consistently estimated by

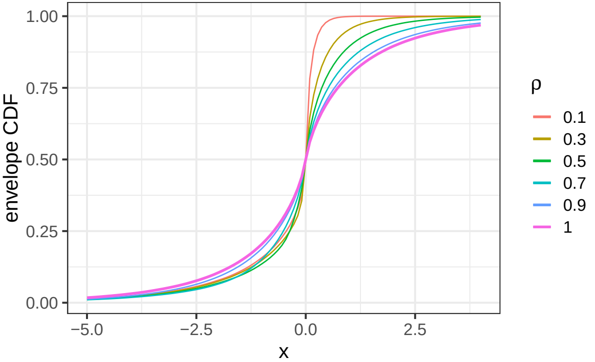
$$\hat{\rho}_n = |\hat{\rho}_{13,n}| \vee |\hat{\rho}_{23,n}|.$$

Adaptive rule

$$\phi_n^{\text{ada}} := \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > -\bar{F}_{\hat{\rho}_n}^{-1}(\alpha) \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < \bar{F}_{\hat{\rho}_n}^{-1}(\alpha) \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{otherwise} \end{cases}.$$

Envelope of envelopes

The negative parts of $\{\bar{F}_\rho : \rho \in [-1, 1]\}$ are dominated by that of $\bar{F}_{\rho=1}$.



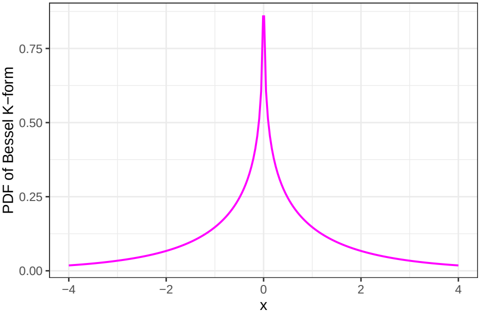
Envelope of envelopes

Bessel envelope

$\bar{F}_{\rho=1}$ is distributed as the difference between two independent χ_1^2 variables.

It has density involving modified Bessel function of the 2nd kind

$$p_B(u) = \frac{1}{2\pi} K_0(|u|/2).$$



Model selection procedure: uniform rule

Uniform rule

$$\phi_n^{\text{unif}} := \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > -\bar{F}_{\rho=1}^{-1}(\alpha) \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < \bar{F}_{\rho=1}^{-1}(\alpha) \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{otherwise} \end{cases} .$$

The quantile is 3.19 for $\alpha = 0.05$ and 5.97 for $\alpha = 0.01$.

Error guarantee

Error guarantee (rate-free)

Theorem: The adaptive rule ϕ_n^{ada} controls asymptotic error uniformly below α for $0 < \alpha < 1/2$.

- This holds for the local model sequences $\rho_{13,n}\rho_{23,n} \asymp n^{-1/2}$ such that the asymptotic error is between 0 and 1.
- This also holds for $\rho_{13,n}\rho_{23,n} = o(n^{-1/2})$ since $\lambda_n^{(0:1)} \to_p 0$ and $\Pr(\phi_n = \mathcal{M}_0 \cup \mathcal{M}_1) \to 1$.
- And also holds for $\rho_{13,n}\rho_{23,n} = \omega(n^{-1/2})$ where $\lambda_n^{(0:1)}$ goes to $\pm\infty$.

Hence, our guarantee holds under $P^n_{\Sigma_n}$ for **any converging sequence Σ_n**. An assumption on the rate of signal strength is not required.

Corollary: ϕ_n^{unif} has the same guarantee.

p-value

When it is desired to report a p-value, the rules can be restated as

$$\phi_n = \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > 0 \text{ and } p\text{-value} < \alpha \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < 0 \text{ and } p\text{-value} < \alpha \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{otherwise} \end{cases},$$

where a potentially conservative p-value is defined as

$$p\text{-value} := \bar{F}_\rho(-|\lambda_n^{(0:1)}|)$$

for $\rho = 1$ (uniform) or $\rho = \hat{\rho}_n$ (adaptive).

Numerical results

Methods for comparison

Naive Simply choosing the model with highest likelihood/AIC/BIC

$$\phi_n^{\text{naive}} := \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > 0 \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < 0 \end{cases}.$$

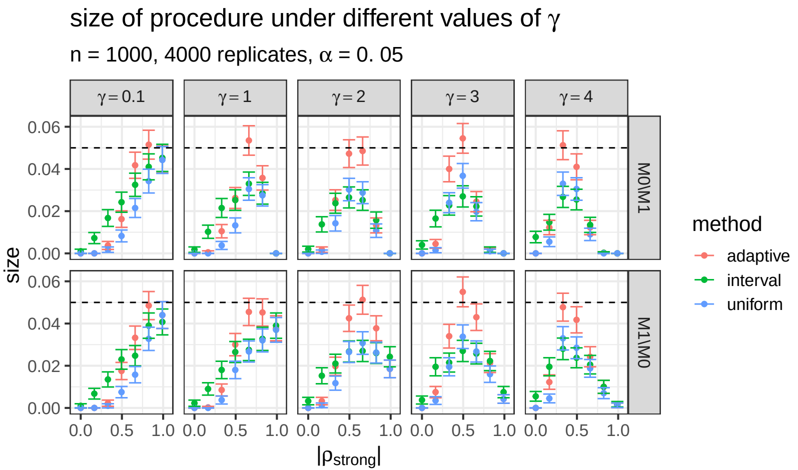
Interval selection This is based on Drton and Perlman (2004). Construct (marginally) $(1-\alpha)$-level confidence intervals for ρ_{12} and $\rho_{12\cdot 3}$, and let

$$\phi_n^{\text{interval}} := \begin{cases} \mathcal{M}_0, & \text{only C.I. for } \rho_{12} \text{ contains } 0 \\ \mathcal{M}_1, & \text{only C.I. for } \rho_{12\cdot 3} \text{ contains } 0 \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{both C.I.'s contain } 0 \end{cases}.$$

ϕ_n^{interval} guarantees asymptotic size below α (suppose $\mathcal{M}_0$ is true, then one only makes an error when the C.I. for ρ_{12} does not contain zero).

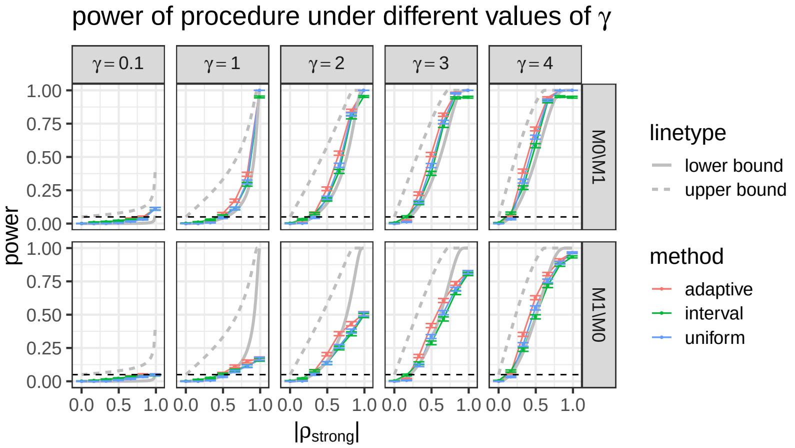
Weak-strong regime: size under $\mathcal{M}_0$ and $\mathcal{M}_1$

Models are simulated as in the weak-strong regime.



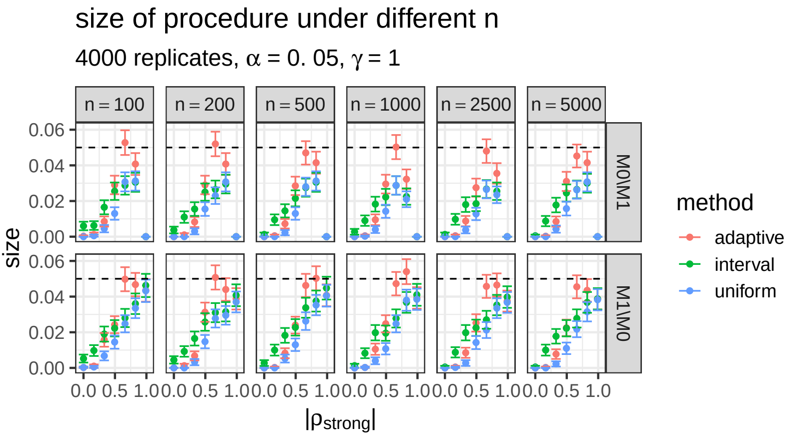
Weak-strong regime: power to select $\mathcal{M}_0$ or $\mathcal{M}_1$

Grey curves are bounds on the theoretically optimal power.



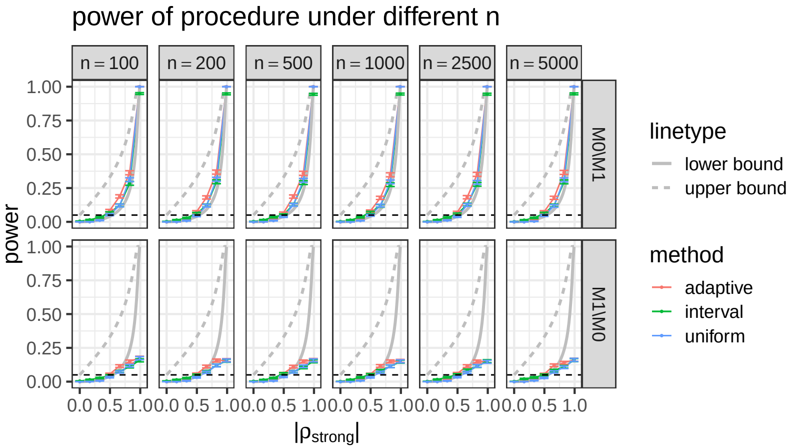
Weak-strong regime: varying sample sizes

Fix $\gamma = 1$ and vary n.



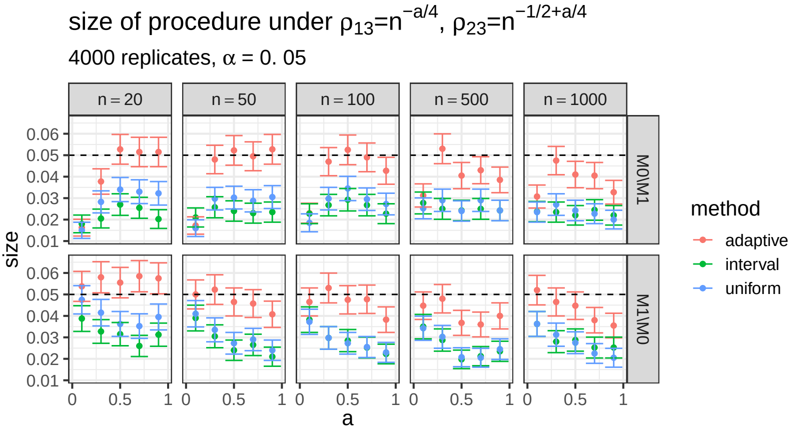
Weak-strong regime: varying sample sizes

Grey curves are bounds on the theoretically optimal power.



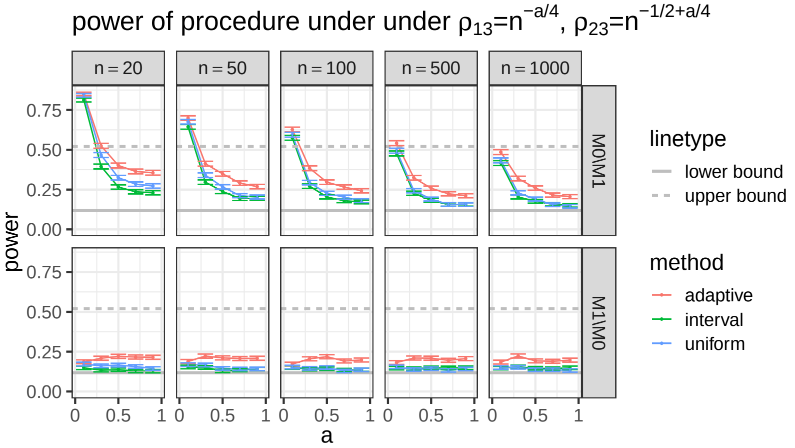
Weak-weak regime: size under $\mathcal{M}_0$ and $\mathcal{M}_1$

The weak-weak regime.



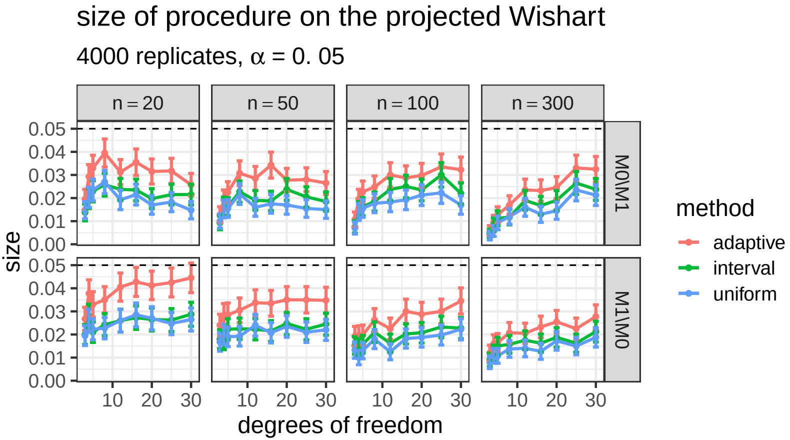
Weak-weak regime: power to select $\mathcal{M}_0$ or $\mathcal{M}_1$

Grey curves are bounds on the theoretically optimal power.



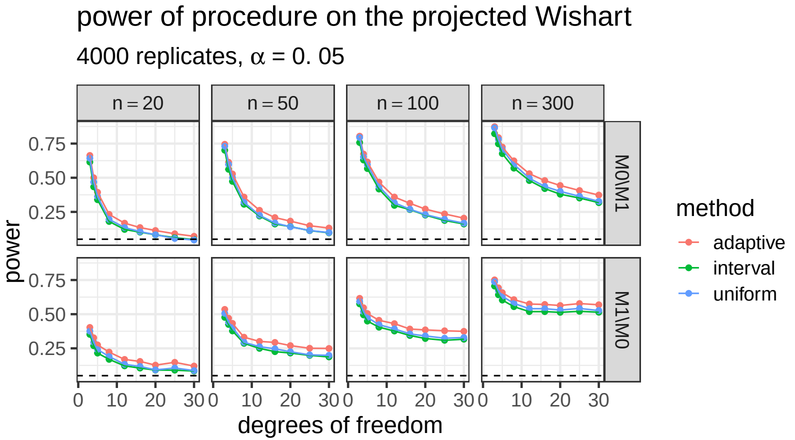
Projected Wishart

Draw $\Sigma \sim \text{Wishart}\left(\nu, (\sigma_{ij})_{3\times 3} = (-\frac{1}{2})^{|i-j|}\right)$ and then projected Σ to $\mathcal{M}_0$ and $\mathcal{M}_1$ by MLE.



Projected Wishart

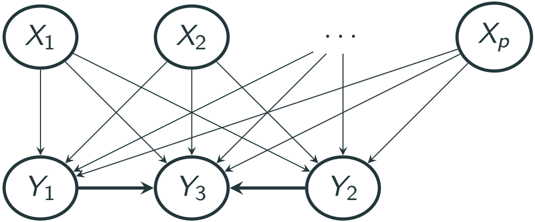
Draw $\Sigma \sim \text{Wishart}\left(\nu, (\sigma_{ij})_{3\times 3} = (-\frac{1}{2})^{|i-j|}\right)$ and then projected Σ to $\mathcal{M}_0$ and $\mathcal{M}_1$ by MLE.



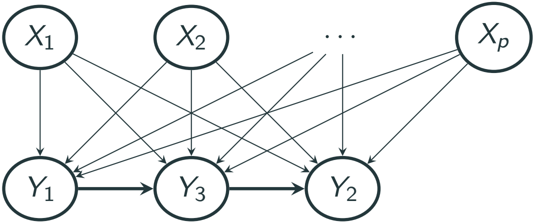
Linear regression

$(Y_1, Y_2, Y_3) = X^\intercal(\beta_1, \beta_2, \beta_3) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma^{(i)})$. $\Sigma^{(i)}$ is drawn from the projected Wishart.

$Y_1 \perp\!\!\!\perp Y_2 \mid X_1, \cdots X_p$

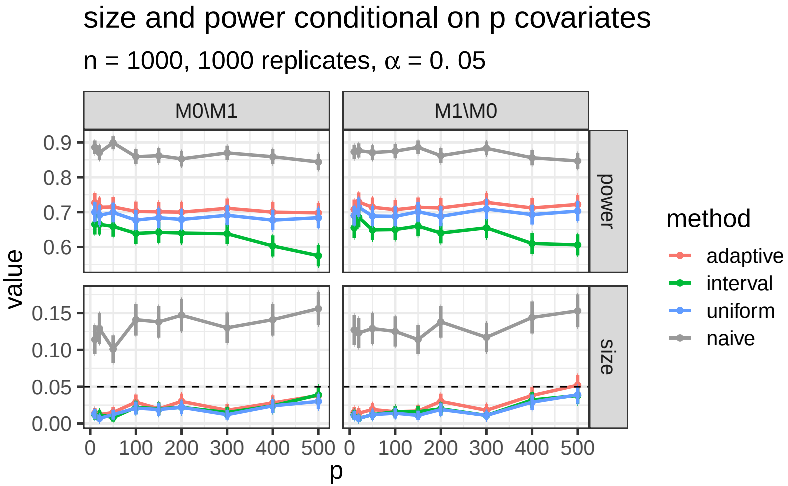


$Y_1 \perp\!\!\!\perp Y_2 \mid Y_3, X_1, \cdots X_p$



Linear regression

$(Y_1, Y_2, Y_3) = X^\intercal(\beta_1, \beta_2, \beta_3) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma^{(i)})$. $\Sigma^{(i)}$ is drawn from the projected Wishart.



Real-data example: American occupational structure

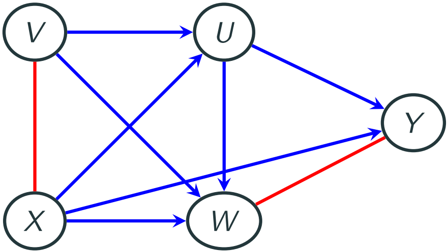
Blau and Duncan (1967) measured the following covariates on $n = 20,700$ subjects:

- V: father's educational attainment,
- X: father's occupational status,
- U: educational attainment,
- W: status of the first job,
- Y: status of occupation in 1962.

Blau and Duncan summarized the data as a correlation matrix.

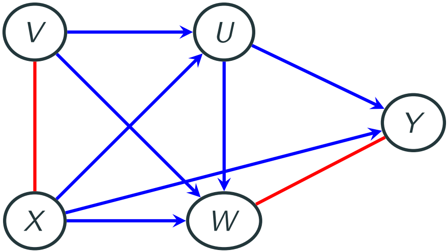
Real-data example: structure learning

We run PC algorithm at level $\alpha = 0.01$. It first identifies the skeleton by d-separation, which only removes the edge between V and Y based on $Y \perp\!\!\!\perp V \mid U, X$.



The blue edges are oriented based on a common-sense temporal ordering $\{V, X\} < U < \{W, Y\}$.

Real-data example: structure learning



Next, the PC algorithm orients edges based on V-structures. The orientation of $V - X$ is statistically unidentifiable (no V-structure).

However, the orientation of $W - Y$ raises the question of testing

$$\mathcal{M}_0\,(Y \to W) : V \perp\!\!\!\perp Y \mid U, X, \quad \mathcal{M}_1\,(Y \leftarrow W) : V \perp\!\!\!\perp Y \mid W, U, X.$$

We have $\lambda_n^{(0:1)} = 3.72$ and p-value $= 0.026$ under the envelope distribution $\bar{F}_{\hat{\rho}_n}$. Hence, under $\alpha = 0.01$ we would leave the edge **unoriented** (even though $n = 20,700$!).

Future work

Can we generalize the method as an off-the-shelf tool for non-nested model selection with error guarantees?

- $\mathcal{M}_i$ as a manifold defined on some ambient Θ. Models can have different dimensions.
- The simplest case is to select between two models. Dealing with more than two models involves multiplicity correction.
- Need a characterization of all possible stable laws of $\lambda^{(0:1)}$.
  - Take any $\theta \in \mathcal{M}_0 \cap \mathcal{M}_1$ and consider $\theta_n^{(0)}, \theta_n^{(1)} \to \theta$ in respective neighborhoods. $\theta_n^{(0)}$ and $\theta_n^{(1)}$ are "closest" to each other in the KL sense.
  - Recall that $\rho_{13}\rho_{23}$ is effectively the parameter that determines the distribution of $\lambda^{(0:1)}$.
  - Can we always introduce a **reparametrization** such that the asymptotic at every neighborhood is equivalent to something simple, even under high-order equivalence (Evans, 2018)?
  - Take an envelope over all these laws.

Thanks!

For details: `https://arxiv.org/abs/1906.01850`

Additional slides

Blau and Duncan dataset

Data collected during the March, 1962 Current Population Survey, on a nationwide sample of about 20,000 American men aged 20-64.

- Occupational statuses are measured by some index.
- Educational attainment is measured by some coding for the number of years of schooling completed.

$$S_n = \begin{pmatrix} 1.000 & 0.516 & 0.453 & 0.332 & 0.322 \\ 0.516 & 1.000 & 0.438 & 0.417 & 0.405 \\ 0.453 & 0.438 & 1.000 & 0.538 & 0.596 \\ 0.332 & 0.417 & 0.538 & 1.000 & 0.541 \\ 0.322 & 0.405 & 0.596 & 0.541 & 1.000 \end{pmatrix}.$$

Limit experiment

Consider an "experiment" $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_h : h \in H)$ in the sense of van der Vaart. h is typically a local parameter.

Fix a "base" $h_0 \in H$. The likelihood ratio process is

$$\left(\frac{\mathrm{d}P_h}{\mathrm{d}P_{h_0}}(X)\right)_{h \in H}, \quad X \sim P_{h_0}.$$

A sequence of experiments $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{h,n} : h \in H)$ converges a limit experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_h : h \in H)$ if the likelihood ratio process weakly converges (marginally). That is, for any finite subset $I \subset H$ and any $h_0 \in H$,

$$\left(\frac{\mathrm{d}P_{h,n}}{\mathrm{d}P_{h_0,n}}(X_n)\right)_{h \in I} \overset{h_0}{\rightsquigarrow} \left(\frac{\mathrm{d}P_h}{\mathrm{d}P_{h_0}}(X)\right)_{h \in I}.$$

Limit experiment

If $(P_{n,\theta} : \theta \in \Theta)$ is locally asymptotic normal (LAN) with norming sequence $n^{-1/2}$ and non-singular I_θ, then the sequence of experiments $(P_{\theta+n^{-1/2},n} : h \in \mathbb{R}^d)$ converges to the limit experiment $(\mathcal{N}(h, I_\theta^{-1}) : h \in \mathbb{R}^d)$.

References

- Blau, Peter M and Otis Dudley Duncan (1967). *The American Occupational Structure*. Wiley New York.
- Drton, Mathias (2006). "Algebraic techniques for Gaussian models". In: *Prague Stochastics*. Ed. by M. Hušková and M. Janžura. Matfyzpress, Charles Univ.
- – (2009). "Likelihood ratio tests and singularities". In: *The Annals of Statistics* 37.2, pp. 979–1012.
- Drton, Mathias and Marloes H Maathuis (2017). "Structure learning in graphical modeling". In: *Annual Review of Statistics and Its Application* 4, pp. 365–393.

References ii

Drton, Mathias and Michael D Perlman (2004). "Model selection for Gaussian concentration graphs". In: *Biometrika* 91.3, pp. 591–602.

Drton, Mathias and Thomas S Richardson (2008). "Graphical methods for efficient likelihood inference in Gaussian covariance models". In: *Journal of Machine Learning Research* 9.May, pp. 893–914.

Drton, Mathias and Seth Sullivant (2007). "Algebraic statistical models". In: *Statistica Sinica*, pp. 1273–1297.

Evans, Robin J (2018). "Model selection and local geometry". In: *arXiv preprint arXiv:1801.08364v3*.