Model uncertainty in statistical inference

Tobias Freidling

Department of Mathematics, Technical University of Munich

September 9, 2020

Outline

Causal effect inference
- Linear Structural Equation Models
- Resampling
- Inverting tests
- Experiments

Post-selection inference with HSIC-Lasso
- Post-selection inference (PSI)
- Hilbert-Schmidt Independence Criterion (HSIC)
- PSI with HSIC-Lasso
- Experiments

Linear Structural Equation Models

Let $\mathcal{G}$ be a directed acyclic graph (DAG) with d nodes and denote the parents of node j as $\mathrm{pa}(j)$.

Definition

Let $X = (X_1, \ldots, X_d)$ be a centred random vector with a causal structure that is linked to a DAG $\mathcal{G}$. A *Linear Structural Equation Model* (LSEM) is given by

$$X_j := \sum_{k \in \mathrm{pa}(j)} \beta_{jk} X_k + \varepsilon_j \qquad \forall j \in \{1, \ldots, d\},$$

where $\{\varepsilon_j\}_{j \in \{1,\ldots,d\}}$ are independent, centred random variables.

Linear Structural Equation Models

Let $\mathcal{G}$ be a directed acyclic graph (DAG) with d nodes and denote the parents of node j as $\mathrm{pa}(j)$.

Definition

Let $X = (X_1, \ldots, X_d)$ be a centred random vector with a causal structure that is linked to a DAG $\mathcal{G}$. A *Linear Structural Equation Model* (LSEM) is given by

$$X_j := \sum_{k \in \mathrm{pa}(j)} \beta_{jk} X_k + \varepsilon_j \qquad \forall j \in \{1, \ldots, d\},$$

where $\{\varepsilon_j\}_{j \in \{1,\ldots,d\}}$ are independent, centred random variables.

Possible two-variable LSEM's:

$$\begin{aligned}
&(\mathrm{M1}) \quad X_1 = \beta_{12} X_2 + \varepsilon_1, \quad && X_2 = \varepsilon_2,\\
&(\mathrm{M2}) \quad X_1 = \varepsilon_1, && X_2 = \beta_{21} X_1 + \varepsilon_2,\\
&(\mathrm{M3}) \quad X_1 = \varepsilon_1, && X_2 = \varepsilon_2.
\end{aligned}$$

Causal effect (e.g. Pearl 2009)

Causal effect of X_2 on X_1: $\beta_{12} 1_{\{X_1 \leftarrow X_2\}}$

Assume equal variance, $\operatorname{var}(\varepsilon_1) = \operatorname{var}(\varepsilon_2) = \sigma^2$, and $X_1 \leftarrow X_2$:

$$\Sigma = \mathrm{E}\left[XX^T\right] = \sigma^2 \begin{pmatrix} 1 + \beta_{12}^2 & \beta_{12} \\ \beta_{12} & 1 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

Causal effect (e.g. Pearl 2009)

Causal effect of X_2 on X_1: $\beta_{12} 1_{\{X_1 \leftarrow X_2\}}$

Assume equal variance, $\mathrm{var}(\varepsilon_1) = \mathrm{var}(\varepsilon_2) = \sigma^2$, and $X_1 \leftarrow X_2$:

$$\Sigma = \mathrm{E}\left[XX^T\right] = \sigma^2 \begin{pmatrix} 1 + \beta_{12}^2 & \beta_{12} \\ \beta_{12} & 1 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

Consequences:

- The parent variable has smaller variance.
- Causal effect of X_2 on X_1:

$$t(\Sigma) = \frac{\sigma_{12}}{\sigma_{22}} 1_{\{\sigma_{11} > \sigma_{22}\}}$$

For point estimate we can plug-in the covariance-estimator $\widehat{\Sigma}$, but what about confidence intervals?

Bootstrapping and subsampling

Since the causal effect is only defined for LSEM's, we need a continuous extension $T(\Sigma)$ accepting general covariance matrices.

We have to choose τ_n such that the root for the construction of confidence intervals

$$\tau_n\big(T(\widehat{\Sigma}) - T(\Sigma)\big)$$

converges to a non-degenerate distribution.

Bootstrapping and subsampling

Since the causal effect is only defined for LSEM's, we need a continuous extension $T(\Sigma)$ accepting general covariance matrices.

We have to choose τ_n such that the root for the construction of confidence intervals

$$\tau_n\big(T(\widehat{\Sigma}) - T(\Sigma)\big)$$

converges to a non-degenerate distribution.

Two continuous extensions were investigated. For both, we find that the asymptotic order of τ_n depends on the unknown causal effect.
E.g. we get $\tau_n = \mathcal{O}(n^{1/2})$ for (M1), $\tau_n = \mathcal{O}(n^{1/8})$ for (M2) and $\tau_n = \mathcal{O}(n^{3/8})$ for (M3).

Hence we cannot use bootstrapping or subsampling. ☹

Inverting tests and constrained statistical inference

Suppose we can test the statistic c for different values c_0 at level α. Then

$$\{c_0 \colon c_0 \text{ is accepted}\}$$

is a $(1-\alpha)$-confidence interval for c.

Inverting tests and constrained statistical inference

Suppose we can test the statistic c for different values c_0 at level α. Then

$$\{c_0 \colon c_0 \text{ is accepted}\}$$

is a $(1-\alpha)$-confidence interval for c.

Constrained statistical inference (cf. Silvapulle and Sen 2005) is a generalisation of likelihood ratio tests. For the nested models $\Theta_0 \subseteq \Theta_1 \subseteq \Theta$ we test

$$\mathrm{H}_0 : \theta \in \Theta_0 \quad \text{against} \quad \mathrm{H}_1 : \theta \in \Theta_1$$

Definition

The *likelihood ratio statistic* λ_n for a sample of size n is defined as

$$\lambda_n = 2\left(\sup_{\theta\in\Theta_1} \ell_n(\theta) - \sup_{\theta\in\Theta_0} \ell_n(\theta)\right).$$

Constrained statistical inference II

General idea: If Θ_0 can be approximated with linear equations and inequalities at the true value θ_0, the asymptotic distribution of λ_n is a mixture of χ^2-distributions.

Constrained statistical inference II

General idea: If Θ_0 can be approximated with linear equations and inequalities at the true value θ_0, the asymptotic distribution of λ_n is a mixture of χ^2-distributions.

Theorem

Let $h^{(1)}(\theta)$ and $h^{(2)}(\theta)$ vector-valued functions. Consider the testing problem

$$H_0 : \theta \in \Theta_0 = \{\theta \colon h^{(1)}(\theta) \geq 0,\ h^{(2)}(\theta) = 0\} \quad vs \quad H_1 : \theta \in \Theta \subseteq \mathbb{R}^p.$$

Under several assumptions, it holds that

$$\mathbb{P}(\lambda_n \geq \cdot \,|\, \theta = \theta_0) \to \sum_{i=0}^{m} w_{m-i}(m, V(\theta_0))\, \mathbb{P}(\chi^2_{r+i} \geq \cdot), \quad as\ n \to \infty,$$

where $\{w_k\}_{k \in \{0,\dots,m\}}$ are positive weights such that $\sum_{k=0}^{m} w_k = 1$.

Constrained statistical inference for causal effects

Goal: developing tests for different values c_0 of the causal effect

- determine log-likelihood, e.g. Gaussian
- translate a causal effect of c_0 into constraint on covariance
- asymptotic distribution of λ_n for different c_0-values

Constrained statistical inference for causal effects

Goal: developing tests for different values c_0 of the causal effect

- determine log-likelihood, e.g. Gaussian
- translate a causal effect of c_0 into constraint on covariance
- asymptotic distribution of λ_n for different c_0-values

Example: $0 < |c_0| < 1$ leads to the constraints $\sigma_{11} > \sigma_{22}$ and $\sigma_{12} = c_0\,\sigma_{22}$. Hence, we test

$$\mathrm{H}_0^{c_0} : \sigma_{12} = c_0\,\sigma_{22},\ \sigma_{11} \geq \sigma_{22} \quad \text{vs} \quad \mathrm{H}_1 : \Sigma \in C$$

and get as (least favourable) asymptotic distribution

$$\sup_{\Sigma \in \mathrm{H}_0^{c_0}} \mathbb{P}\left(\lambda_n^{c_0} \leq \cdot\right) \to \frac{1}{2}\mathbb{P}\left(\chi_1^2 \leq \cdot\right) + \frac{1}{2}\mathbb{P}\left(\chi_2^2 \leq \cdot\right), \quad \text{as } n \to \infty.$$

Simulated data

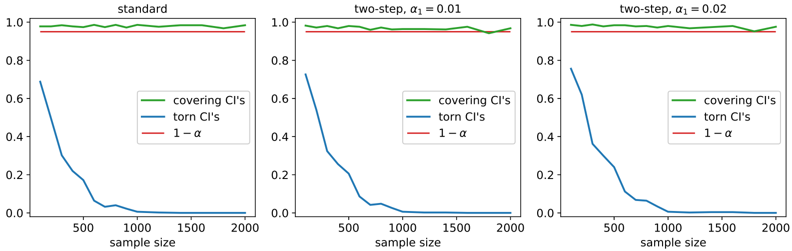


Figure: Share of covering and torn confidence intervals for $\beta = 0.5$ and $X_1 \leftarrow X_2$.

Benchmark data

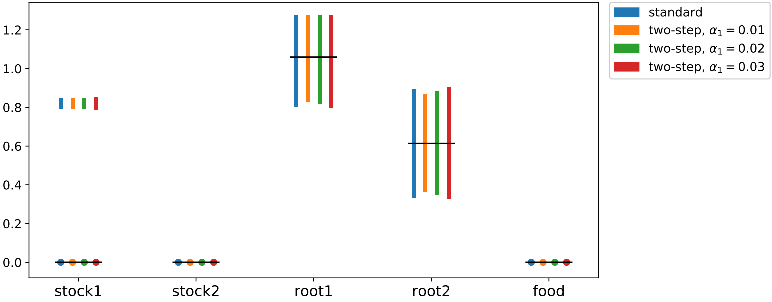


Figure: 95% confidence intervals for the causal effect of X_2 on X_1.

Post-selection inference (PSI)

Idea: When we select a model based on data, we have to account for the selection in the inference results.

Example: We select a subset of covariates with Lasso in a linear regression setting. When testing hypothesis in the selected model, we have to take into account that the selected covariates are overly significant.

Post-selection inference (PSI)

Idea: When we select a model based on data, we have to account for the selection in the inference results.

Example: We select a subset of covariates with Lasso in a linear regression setting. When testing hypothesis in the selected model, we have to take into account that the selected covariates are overly significant.

Polyhedral Lemma setting (cf. Lee et al. 2016):
$Y \sim \mathcal{N}(\mu, \Sigma) \rightarrow$ selection procedure $\rightarrow$ inference for $\eta^T Y$

- selection procedure restricts Y according to $\{AY \leq b\}$
- inference for $\eta^T Y | \{AY \leq b\}$ is valid
- $\{AY \leq b\} = \{\mathcal{V}^- \leq \eta^T Y \leq \mathcal{V}^+\}$ (Polyhedral Lemma)
- $\eta^T Y | \{AY \leq b\} = \eta^T Y | \{\mathcal{V}^- \leq \eta^T Y \leq \mathcal{V}^+\}$, i.e. a truncated normal distribution

Truncated Gaussians

Let $F^{[a,b]}_{\mu,\sigma^2}$ denote the cdf of a $\mathcal{N}(\mu,\sigma^2)$ truncated to the interval $[a,b]$, that is

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi(\frac{x-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})},$$

where Φ is the cdf of $\mathcal{N}(0,1)$.

Truncated Gaussians

Let $F_{\mu,\sigma^2}^{[a,b]}$ denote the cdf of a $\mathcal{N}(\mu,\sigma^2)$ truncated to the interval $[a,b]$, that is

$$F_{\mu,\sigma^2}^{[a,b]}(x) = \frac{\Phi(\frac{x-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})},$$

where Φ is the cdf of $\mathcal{N}(0,1)$.

Theorem

Let $Y \sim \mathcal{N}(\mu,\Sigma)$, then

$$F_{\eta^T\mu,\eta^T\Sigma\eta}^{[\mathcal{V}^-(z),\mathcal{V}^+(z)]}(\eta^T Y) | \{AY \leq b\} \sim Unif(0,1),$$

where

$$z = \left(\mathrm{Id} - (\eta^T\Sigma\eta)^{-1}\Sigma\eta\eta^T\right)Y.$$

Note:
If X is a random variable and F is its cdf, then $F(X) \sim \mathrm{Unif}(0,1)$.

HSIC (Gretton et al. 2005)

Idea: embed probability measures $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$ in Reproducing Kernel Hilbert Space (RKHS) and compare them through the distance in RKHS

Definition

Let X and Y be random variables and $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ kernel functions. The *Hilbert-Schmidt independence criterion* is given by

$$\begin{aligned}\mathrm{HSIC}(X,Y) =& \mathrm{E}_{x,x',y,y'}[k(x,x')l(y,y')] + \mathrm{E}_{x,x'}[k(x,x')]\,\mathrm{E}_{y,y'}[l(y,y')] \\ &- 2\mathrm{E}_{x,y}\left[\mathrm{E}_{x'}[k(x,x')]\mathrm{E}_{y}[l(y,y')]\right],\end{aligned}$$

where $\mathrm{E}_{x,x',y,y'}$ denotes the expectation over independent pairs (x,y) and (x',y').

HSIC (Gretton et al. 2005)

Idea: embed probability measures $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$ in Reproducing Kernel Hilbert Space (RKHS) and compare them through the distance in RKHS

Definition

Let X and Y be random variables and $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ kernel functions. The *Hilbert-Schmidt independence criterion* is given by

$$\begin{aligned}\mathrm{HSIC}(X,Y) =& \mathrm{E}_{x,x',y,y'}[k(x,x')l(y,y')] + \mathrm{E}_{x,x'}[k(x,x')]\,\mathrm{E}_{y,y'}[l(y,y')] \\ & - 2\mathrm{E}_{x,y}\left[\mathrm{E}_{x'}[k(x,x')]\mathrm{E}_{y}[l(y,y')]\right],\end{aligned}$$

where $\mathrm{E}_{x,x',y,y'}$ denotes the expectation over independent pairs (x,y) and (x',y').

Properties:

- No assumptions on X, Y and their relationship. Modelfree!
- $\mathrm{HSIC}(X,Y) \geq 0, \quad \mathrm{HSIC}(X,Y) = 0 \Leftrightarrow X \perp Y$
- Classification and regression settings with suitable kernels possible!

HSIC estimators I

We are given a sample $\{y_i, x_i\}_{i=1}^n$ and define K and L by $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(x_i, x_j)$ for $1 \leq i, j \leq n$. $\tilde{K} = K - \mathrm{diag}(K)$, $\tilde{L} = L - \mathrm{diag}(L)$ and $\Gamma = \mathrm{Id} - \frac{1}{n}11^T$.

Biased estimator (Gretton et al. 2005):

$$\widehat{\mathrm{HSIC}}_b(X, Y) = (n-1)^{-2}\,\mathrm{tr}(K\Gamma L\Gamma)$$

Unbiased estimator (Song et al. 2012):

$$\widehat{\mathrm{HSIC}}_u(X, Y) = \frac{1}{n(n-3)}\left(\mathrm{tr}(\tilde{K}\tilde{L}) + \frac{1^T\tilde{K}1\,1^T\tilde{L}1}{(n-1)(n-2)} - \frac{2}{n-2}1^T\tilde{K}\tilde{L}1\right)$$

HSIC estimators I

We are given a sample $\{y_i, x_i\}_{i=1}^n$ and define K and L by $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(x_i, x_j)$ for $1 \leq i, j \leq n$. $\tilde{K} = K - \mathrm{diag}(K)$, $\tilde{L} = L - \mathrm{diag}(L)$ and $\Gamma = \mathrm{Id} - \frac{1}{n} 1 1^T$.

Biased estimator (Gretton et al. 2005):

$$\widehat{\mathrm{HSIC}}_b(X, Y) = (n-1)^{-2}\,\mathrm{tr}(K \Gamma L \Gamma)$$

Unbiased estimator (Song et al. 2012):

$$\widehat{\mathrm{HSIC}}_u(X, Y) = \frac{1}{n(n-3)} \left(\mathrm{tr}(\tilde{K}\tilde{L}) + \frac{1^T \tilde{K} 1\, 1^T \tilde{L} 1}{(n-1)(n-2)} - \frac{2}{n-2} 1^T \tilde{K} \tilde{L} 1 \right)$$

If X and Y are independent, for both estimators $n\,\widehat{\mathrm{HSIC}}(X, Y)$ does not converge to a Gaussian random variable. ☹

HSIC estimators II

Block estimator (Zhang et al. 2017):
Divide sample into blocks of size B, $\{\{y_i^b, x_i^b\}_{i=1}^B\}_{b=1}^{n/B}$.

$$\widehat{\text{HSIC}}_{\text{Block}}(X, Y) = \frac{1}{n/B} \sum_{b=1}^{n/B} \widehat{\text{HSIC}}_u(X^b, Y^b)$$

HSIC estimators II

Block estimator (Zhang et al. 2017):
Divide sample into blocks of size B, $\{\{y_i^b, x_i^b\}_{i=1}^{B}\}_{b=1}^{n/B}$.

$$\widehat{\mathrm{HSIC}}_{\mathrm{Block}}(X, Y) = \frac{1}{n/B} \sum_{b=1}^{n/B} \widehat{\mathrm{HSIC}}_u(X^b, Y^b)$$

Incomplete U-statistics estimator (Lim et al. 2020):
HSIC is a U-statistic of degree 4, i.e. there exists h such that $\widehat{\mathrm{HSIC}}_u(X, Y) = \binom{n}{4}^{-1} \sum_{(i,j,q,r) \in \mathcal{S}_{n,4}} h(i, j, q, r)$, where $\mathcal{S}_{n,4}$ is the set of all 4-subsets of $\{1, \ldots, n\}$. Let $\mathcal{D} \subset \mathcal{S}_{n,4}$ and $|\mathcal{D}| = m = \mathcal{O}(n)$, then

$$\widehat{\mathrm{HSIC}}_{\mathrm{inc}}(X, Y) = m^{-1} \sum_{(i,j,q,r) \in \mathcal{D}} h(i, j, q, r).$$

HSIC estimators II

Block estimator (Zhang et al. 2017):
Divide sample into blocks of size B, $\{\{y_i^b, x_i^b\}_{i=1}^{B}\}_{b=1}^{n/B}$.

$$\widehat{\mathrm{HSIC}}_{\mathrm{Block}}(X, Y) = \frac{1}{n/B} \sum_{b=1}^{n/B} \widehat{\mathrm{HSIC}}_u(X^b, Y^b)$$

Incomplete U-statistics estimator (Lim et al. 2020):
HSIC is a U-statistic of degree 4, i.e. there exists h such that $\widehat{\mathrm{HSIC}}_u(X, Y) = \binom{n}{4}^{-1} \sum_{(i,j,q,r)\in\mathcal{S}_{n,4}} h(i,j,q,r)$, where $\mathcal{S}_{n,4}$ is the set of all 4-subsets of $\{1, \ldots, n\}$. Let $\mathcal{D} \subset \mathcal{S}_{n,4}$ and $|\mathcal{D}| = m = \mathcal{O}(n)$, then

$$\widehat{\mathrm{HSIC}}_{\mathrm{inc}}(X, Y) = m^{-1} \sum_{(i,j,q,r)\in\mathcal{D}} h(i,j,q,r).$$

Both $\sqrt{n/B}\,\widehat{\mathrm{HSIC}}_{\mathrm{Block}}(X, Y)$ and $\sqrt{m}\,\widehat{\mathrm{HSIC}}_{\mathrm{inc}}(X, Y)$ are asymptotically normal. ☺

HSIC-Lasso (Yamada 2014)

Let $\bar{L} = \Gamma L \Gamma$ and $\bar{K}^{(k)} = \Gamma K^{(k)} \Gamma$. The HSIC-Lasso solution is given by

$$\begin{aligned}
\hat{\beta} &= \operatorname*{argmin}_{\beta \geq 0} \frac{1}{2} \| \bar{L} - \sum_{k=1}^{p} \beta_k \bar{K}^{(k)} \|_{\text{Frob}}^2 + \lambda \|\beta\|_1 \\
&= \operatorname*{argmin}_{\beta \geq 0} - \sum_{k=1}^{p} \beta_k \widehat{\text{HSIC}}_b(X^{(k)}, Y) + \frac{1}{2} \sum_{k,l=1}^{p} \beta_k \beta_l \widehat{\text{HSIC}}_b(X^{(k)}, X^{(l)}) + \lambda \|\beta\|_1
\end{aligned}$$

- 1^{st} term selects influential covariates
- 2^{nd} term punishes selection of dependent variables
- 3^{rd} term enforces sparsity

HSIC-Lasso (Yamada 2014)

Let $\bar{L} = \Gamma L \Gamma$ and $\bar{K}^{(k)} = \Gamma K^{(k)} \Gamma$. The HSIC-Lasso solution is given by

$$\hat{\beta} = \underset{\beta \geq 0}{\operatorname{argmin}} \frac{1}{2} \|\bar{L} - \sum_{k=1}^{p} \beta_k \bar{K}^{(k)}\|_{\text{Frob}}^2 + \lambda \|\beta\|_1$$

$$= \underset{\beta \geq 0}{\operatorname{argmin}} - \sum_{k=1}^{p} \beta_k \widehat{\text{HSIC}}_b(X^{(k)}, Y) + \frac{1}{2} \sum_{k,l=1}^{p} \beta_k \beta_l \widehat{\text{HSIC}}_b(X^{(k)}, X^{(l)}) + \lambda \|\beta\|_1$$

- 1st term selects influential covariates
- 2nd term punishes selection of dependent variables
- 3rd term enforces sparsity

How to do post-selection inference with the Polyhedral Lemma? We need a multivariate Gaussian random variable, a quantity for inference and a characterisation of the truncation points.

Multivariate Gaussian & inference targets

We replace the biased estimator with the Block or the incomplete U-statistics estimator which are asymptotically normal, for example:

$$\hat{\beta} = \underset{\beta \geq 0}{\operatorname{argmin}} - \sum_{k=1}^{p} \beta_k \widehat{\mathrm{HSIC}}_{\mathrm{Block}}(X^{(k)}, Y) + \frac{1}{2} \sum_{k,l=1}^{p} \beta_k \beta_l \widehat{\mathrm{HSIC}}(X^{(k)}, X^{(l)}) + \lambda \|\beta\|_1$$

$$=: \underset{\beta \geq 0}{\operatorname{argmin}} -\beta^T H + \frac{1}{2} \beta^T M \beta + \lambda \|\beta\|_1,$$

where $H_k = \widehat{\mathrm{HSIC}}_{\mathrm{Block}}(X^{(k)}, Y)$ and $M_{kl} = \widehat{\mathrm{HSIC}}(X^{(k)}, X^{(l)})$. We define $\hat{S} := \{j : \hat{\beta}_j > 0\}$ and assume it takes the value S.

Multivariate Gaussian & inference targets

We replace the biased estimator with the Block or the incomplete U-statistics estimator which are asymptotically normal, for example:

$$\begin{aligned}\hat{\beta} &= \underset{\beta \geq 0}{\operatorname{argmin}} - \sum_{k=1}^{p} \beta_k \widehat{\mathrm{HSIC}}_{\mathrm{Block}}(X^{(k)}, Y) + \frac{1}{2} \sum_{k,l=1}^{p} \beta_k \beta_l \widehat{\mathrm{HSIC}}(X^{(k)}, X^{(l)}) + \lambda \|\beta\|_1 \\ &=: \underset{\beta \geq 0}{\operatorname{argmin}} - \beta^T H + \frac{1}{2} \beta^T M \beta + \lambda \|\beta\|_1,\end{aligned}$$

where $H_k = \widehat{\mathrm{HSIC}}_{\mathrm{Block}}(X^{(k)}, Y)$ and $M_{kl} = \widehat{\mathrm{HSIC}}(X^{(k)}, X^{(l)})$. We define $\hat{S} := \{j : \hat{\beta}_j > 0\}$ and assume it takes the value S.

Partial regression coefficient:
In analogy with linear regression, we look at partial regression coefficients

$$\hat{\beta}_j^{\mathrm{par}} = \mathrm{e}_j^T M_{SS}^{-1} H_S = \mathrm{e}_j^T (M_{SS}^{-1} | 0) H =: \eta^T H.$$

HSIC estimate:

$$H_j = \mathrm{e}_j^T H =: \eta^T H$$

Truncation points

We denote $S^c := \{1, \dots, p\} \setminus S$.

Partial regression coefficients:
Similarly to the Lasso-example, the selection event can be characterised using the Karush-Kuhn-Tucker conditions. We get

$$\frac{1}{\lambda} \begin{pmatrix} -M_{SS}^{-1} & \mid & 0 \\ -M_{S^cS} M_{SS}^{-1} & \mid & \mathrm{Id} \end{pmatrix} H \leq \begin{pmatrix} -M_{SS}^{-1} 1 \\ 1 - M_{S^cS} M_{SS}^{-1} 1 \end{pmatrix}.$$

The truncation points $\mathcal{V}^-$ and $\mathcal{V}^+$ are given by the Polyhedral Lemma.

HSIC estimate:
We define $\hat{\beta}_{-j}$ as $\hat{\beta}$ with 0 at the j-th position and can directly derive the truncation points $\mathcal{V}^-$ and $\mathcal{V}^+$

$$\mathcal{V}^- = \lambda + (M\hat{\beta}_{-j})_j$$
$$\mathcal{V}^+ = \infty$$

Testing

We conduct the tests

$$\mathrm{H}_0 : \hat{\beta}_j^{\mathrm{par}} = 0 \quad \text{against} \quad \mathrm{H}_1 : \hat{\beta}_j^{\mathrm{par}} > 0 \quad \text{and}$$
$$\mathrm{H}_0 : H_j = 0 \quad \text{against} \quad \mathrm{H}_1 : H_j > 0.$$

The p-value is given by

$$p = 1 - F_{0,\eta^T \Sigma \eta}^{[\mathcal{V}^-, \mathcal{V}^+]}(\eta^T H).$$

Issues in practical application

- M must be positive definite to apply KKT conditions
  solution: positive definite approximation $\tilde{M}$
- high computational costs
  solution: screening step for potentially influential covariates
- hyperparameter choice
  solution: data splitting into two folds; hyperparameter estimation on first fold, HSIC-Lasso selection on second fold

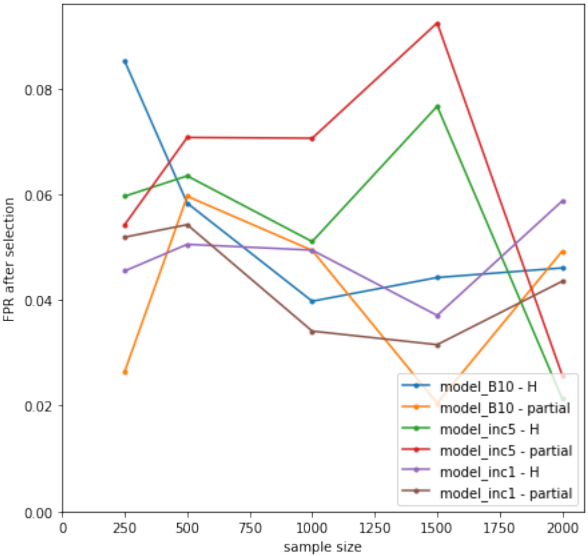
Simulation Example

We generate $n \in \{250, 500, 1000, 1500, 2000\}$ samples from

$$
\begin{aligned}
& X \sim \mathcal{N}(0_{500}, \Sigma), \quad \Sigma_{ij} = 0.25^{|i-j|}, \quad 1 \leq i, j \leq 500, \\
& E \sim \mathcal{N}(0, 0.6), \\
& Y = (X_1 - 1)\tanh(X_2 + X_3 + 1) + \operatorname{sign}(X_4) + E,
\end{aligned}
$$

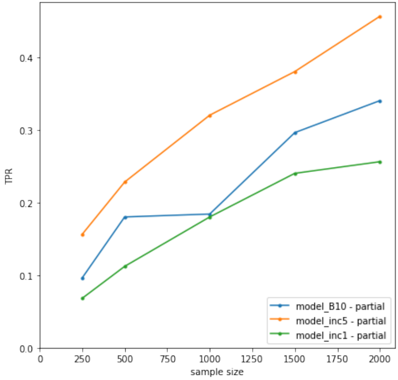
and set the confidence level $\alpha = 0.05$.

False Positive Rate (FPR)



True Positive Rate (TPR)

Partial target



HSIC target