Post-Selection Inference with HSIC-Lasso

Tobias Freidling[1], Benjamin Poignard[2,3], Héctor Climente-González[3], Makoto Yamada[3,4]

[1]DPMMS, University of Cambridge [2]Graduate School of Economics, Osaka University [3]Center for Advanced Intelligence Project (AIP), RIKEN, Kyoto [4]Graduate School of Informatics, Kyoto University

ICML 2021

Hilbert-Schmidt Independence Criterion

The Hilbert-Schmidt Independence Criterion (HSIC) measures the dependence between two random variables X and Y:

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

where k and l are kernel functions and X' and Y' are i.i.d. copies.

- $\mathrm{HSIC}(X, Y) \geq 0, \quad \mathrm{HSIC}(X, Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$
- **Modelfree**, i.e. no assumptions on distribution of X and Y required

Feature Selection

Goal: Selection of (non-redundant) subset of features $X_1, \ldots, X_p$ that are strongly associated with response Y.

- HSIC-ordering: Select k features for which $\widehat{\mathrm{HSIC}}(Y, X_j)$ is largest
- **HSIC-Lasso**: Select j-th feature if $\hat{\beta}_j$ is positive, where

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p_+}{\operatorname{argmin}} \; -\sum_{j=1}^{p} \beta_j \widehat{\mathrm{HSIC}}(X_j, Y) + \frac{1}{2} \sum_{i,j=1}^{p} \beta_i \beta_j \widehat{\mathrm{HSIC}}(X_i, X_j) + \lambda \|\beta\|_1.$$

Post-selection Inference (PSI)

To guarantee correct inference on the selected features, account for/condition on the information encapsulated in the selection. For (affine) linear inference target $\eta^T \mu$ and selection procedure $\{AY \leq b\}$:

Polyhedral Lemma

Let $Y \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^q$ and $\Sigma \in \mathbb{R}^{q \times q}$, $\eta \in \mathbb{R}^q$, $A \in \mathbb{R}^{m \times q}$ and $b \in \mathbb{R}^m$. Then,

$$\eta^T Y | \{AY \leq b\} \sim \text{TruncatedNormal}(\eta^T \mu,\ \eta^T \Sigma \eta,\ \mathcal{V}^-, \mathcal{V}^+),$$

where $\mathcal{V}^-$ and $\mathcal{V}^+$ are the lower and upper truncation point.

PSI with HSIC-Lasso

- **Normal HSIC-Lasso**:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p_+}{\operatorname{argmin}} \; -\beta^T H + \frac{1}{2}\beta^T M \beta + \lambda \beta^T w,$$

  where $M_{ij} = \widehat{\mathrm{HSIC}}(X_i, X_j)$, asymptotically normal $H_j = \widehat{\mathrm{HSIC}}_N(X_j, Y)$ and weight vector w.

- **Affine linear selection**: Selection procedure $\hat{S} = \{j : \hat{\beta}_j > 0\}$
  For positive definite M, $\{\hat{S} = S\} = \{A(H_S, H_{S^c})^T \leq b\}$ with

$$A = -\frac{1}{\lambda}\begin{pmatrix} M_{SS}^{-1} & | & 0 \\ M_{S^cS}M_{SS}^{-1} & | & \mathrm{Id} \end{pmatrix}, \quad b = \begin{pmatrix} -M_{SS}^{-1} w_S \\ w_{S^c} - M_{S^cS}M_{SS}^{-1} w_S \end{pmatrix}.$$

- **Inference targets**: HSIC-target $H_j = \mathrm{e}_j^T H \;\Rightarrow\; \eta = \mathrm{e}_j$; partial target $\hat{\beta}^{\mathrm{par}}_{j,S} = M_{SS}^{-1} H_S \;\Rightarrow\; \eta = (M_{SS}^{-1} H \,|\, 0)^T \mathrm{e}_j$.
- Polyhedral Lemma for **asymptotically normal** random variables

Application in Practice

Challenges

- Positive definiteness of M: positive definite approximation
- Computational costs of HSIC-estimation and choice of λ: set data aside to screen for relevant features and estimate λ

Flexibility

- 2 asymptotically normal HSIC-estimators (block and incomplete U-statistics)
- Adaptive and non-adaptive Lasso-penalty
- Hyper-parameter choice via cross-validation or AIC

Type-I Error and Power

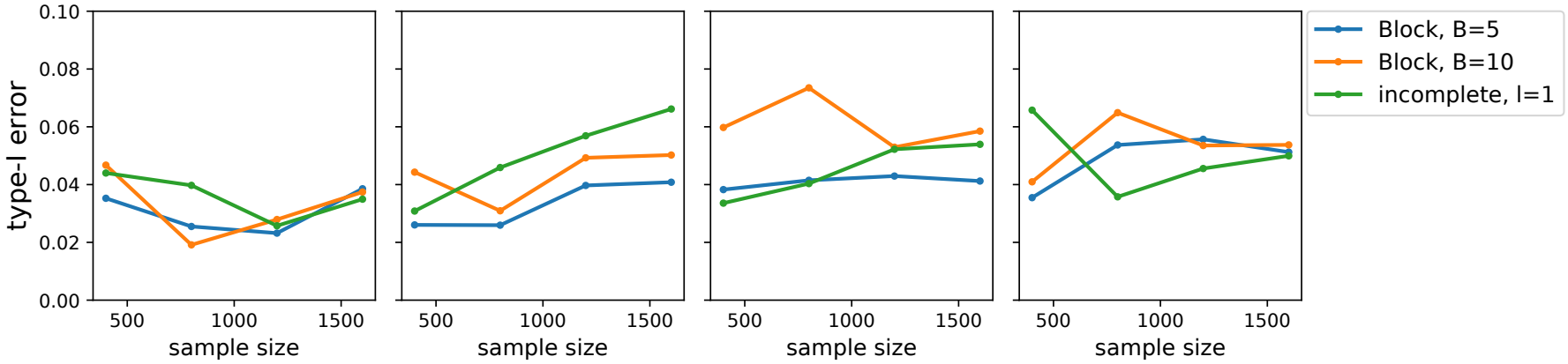

Figure: Empirical type-I error for different toy models

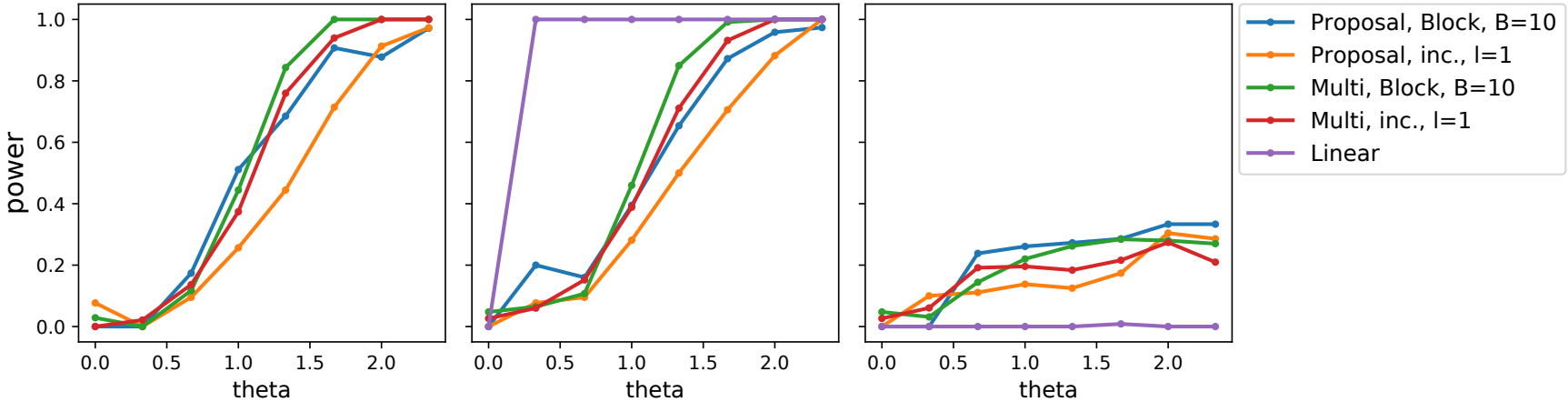

Figure: Empirical power for discrete, linear and non-linear toy model

Real-World Data

Evaluation of our proposal on three datasets and comparison with post-selection inference with HSIC-ordering:

- Parsimonious model
- Discovering potentially novel dependencies
- Effective reduction of selecting strongly correlated features
- Utility of partial target fostering a more interpretable view on the data

Slides of the 5-minute presentation following.

Feature Selection with HSIC-Lasso

Hilbert-Schmidt Independence Criterion (HSIC)

- HSIC(X, Y): measure of dependence between X and Y
- HSIC$(X, Y) \geq 0, \quad$ HSIC$(X, Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$
- Model-free: no assumptions on distribution of X and Y

Feature Selection

Data sample from response Y and features $X_1, \ldots, X_p$

- HSIC-ordering: Select k features for which $\widehat{\text{HSIC}}(Y, X_j)$ is largest
- HSIC-Lasso: Select j-th feature if $\hat{\beta}_j$ is positive, where

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p_+}{\text{argmin}} \; -\sum_{j=1}^{p} \beta_j \widehat{\text{HSIC}}(X_j, Y) + \frac{1}{2} \sum_{i,j=1}^{p} \beta_i \beta_j \widehat{\text{HSIC}}(X_i, X_j) + \lambda \|\beta\|_1.$$

Post-Selection Inference with Polyhedral Lemma

For correct inference on HSIC(Y, X_j) (HSIC-target) we condition on the selection procedure.

Polyhedral Lemma

Let $Y \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$, $\eta \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then

$$\eta^T Y | \{AY \leq b\} \sim \mathrm{TN}(\eta^T \mu, \eta^T \Sigma \eta, \mathcal{V}^-, \mathcal{V}^+).$$

- Prove asymptotic version of Polyhedral Lemma
- Use asymptotically normal HSIC-estimators
- Show that HSIC-Lasso selection is affine linear
- Express HSIC-target in the form $\eta^T Y$ and introduce the novel partial target

Practical Application

Computational costs:

- High-dimensional data: upstream screening stage to reduce features considered by HSIC-Lasso
- Less expensive than multiscale bootstrap

Choice of the hyper-parameter λ:

- Set data aside to estimate λ
- Cross validation and AIC; adaptive and non-adaptive Lasso penalty

We give some intuition for different set-ups.

Type-I Error and Power

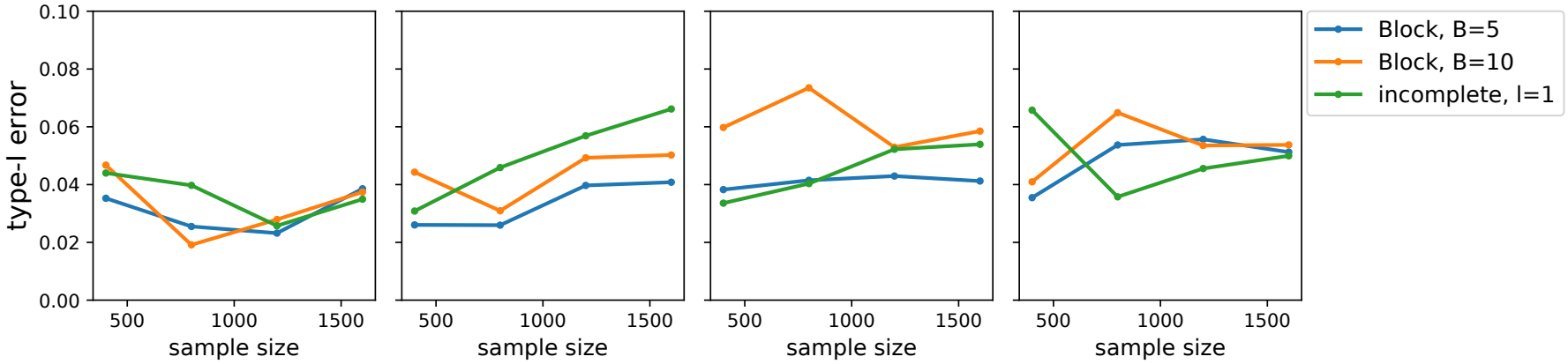

Figure: Empirical type-I error for different toy models

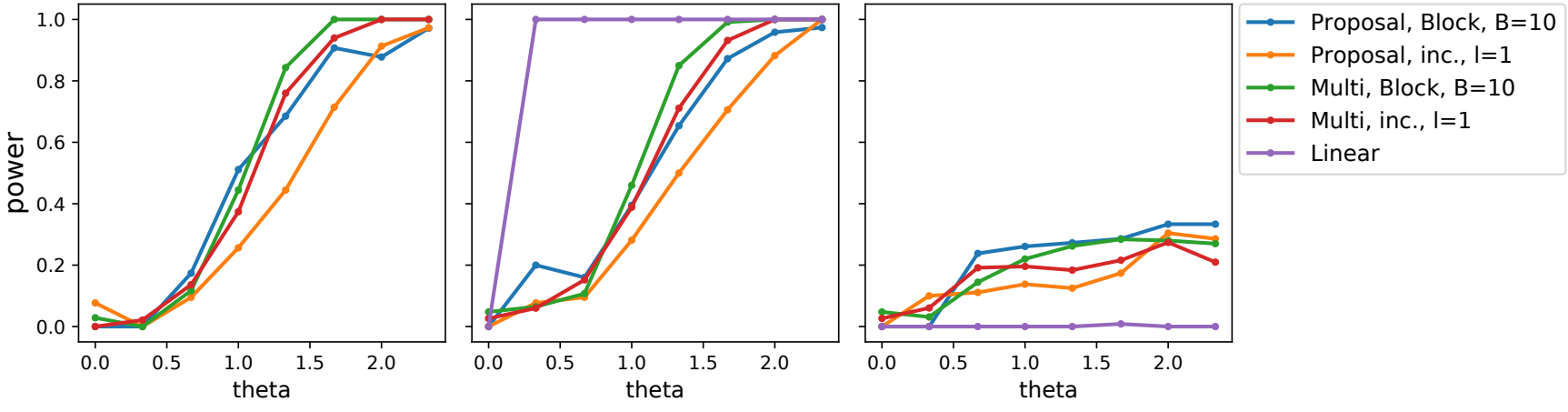

Figure: Empirical power for discrete, linear and non-linear toy model

Real-World Data

Evaluation of our proposal on three datasets and comparison with post-selection inference with HSIC-ordering:

- Parsimonious model
- Discovering potentially novel dependencies
- Effective reduction of selecting strongly correlated features
- Utility of partial target fostering a more interpretable view on the data