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

Introduction

- Response variable is an unknown function of different features.
- Feature-selection is important to understand the data-generating process and build parsimonious models, esp. for 'small n large p' data.
- Few assumptions are desirable. $\rightarrow$ model-free selection methods
- Inference on the selected features is only correct when the selection is accounted for.

HSIC-Lasso

Hilbert-Schmidt Independence Criterion (HSIC)

HSIC[1] measures the dependence between two random variables X and Y:

$$\mathrm{HSIC}(X,Y) = \mathrm{E}_{X,X',Y,Y'}[k(X,X')\, l(Y,Y')] + \mathrm{E}_{X,X'}[k(X,X')]\, \mathrm{E}_{Y,Y'}[l(Y,Y')] - 2\,\mathrm{E}_{X,Y}[\mathrm{E}_{X'}[k(X,X')]\, \mathrm{E}_{Y'}[l(Y,Y')]],$$

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[2]: Select k features for which $\widehat{\mathrm{HSIC}}(Y, X_j)$ is largest
- **HSIC-Lasso**[3]: Select j-th feature if $\hat{\beta}_j$ is positive, where

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

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[4]

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

Type-I Error and Power

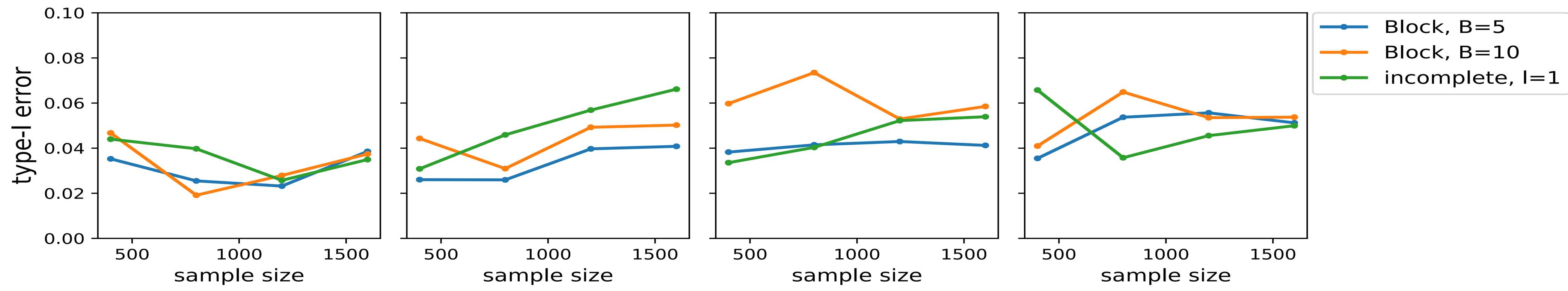

Figure 1. Empirical type-I error for HSIC-target and envisaged level 0.05. Asymptotically normal block and incomplete U-statistics estimator with varying sizes. Toy models with continuous (1st & 2nd panel) and categorical (3rd & 4th panel) response; with and without correlation in features.

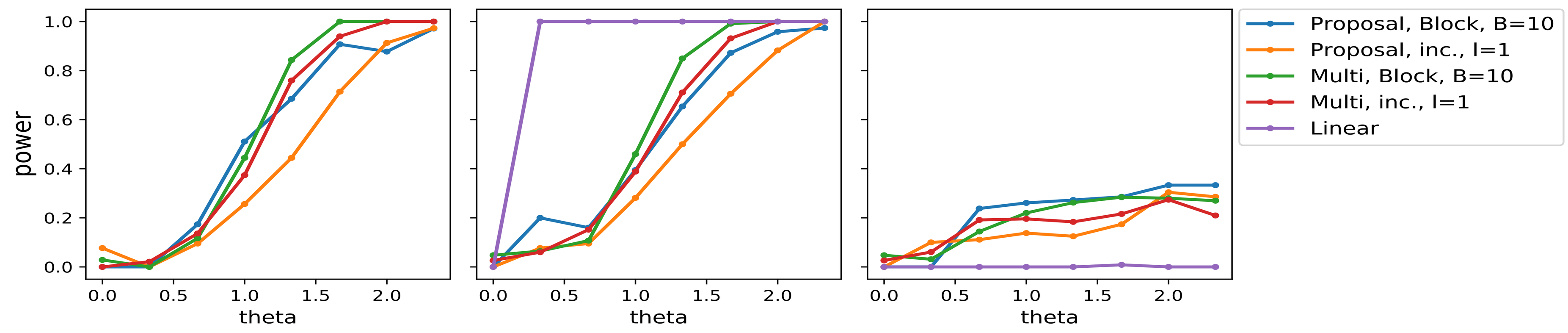

Figure 2. Empirical power for detecting the feature θX_1. Proposed method, multiscale bootstrapping[5] and linear PSI-model. Toy models with discrete (1st panel), linear (2nd panel) and non-linear (3rd panel) data-generating process.

PSI with HSIC-Lasso

- **Normal HSIC-Lasso**: $\hat{\beta} = \mathrm{argmin}_{\beta \in \mathbb{R}^p_+} \; -\beta^T H + \frac{1}{2}\beta^T M \beta + \lambda\, \beta^T w$, 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 $\hat{S} = \{j \colon \hat{\beta}_j > 0\}$
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 (similar to regression coefficient) $\hat{\beta}^{\mathrm{par}}_{j,S} = M_{SS}^{-1} H_S \;\Rightarrow\; \eta = (M_{SS}^{-1} H \,|\, 0)^T \mathrm{e}_j$.
- Polyhedral Lemma for **asymptotically normal** random variables

Application in Practice

- Challenges: (1) Positive definiteness of M, (2) Computational costs of HSIC-estimation, (3) choice of λ
- Solution: (1) Positive definite approximation, (2) & (3) Set data aside to screen for relevant features and estimate λ
- Flexibility: 2 asymptotically normal HSIC-estimators[5] with adjustable size; Adaptive- and non-adaptive Lasso penalty; Hyper-parameter choice via cross-validation or AIC

Real-world Data

- RNAseq data from the Broad Institute's Single Cell Portal
- Response: type of blood cell (10-level categorical), Features: 26 593 genes, Sample size: 1 078
- Half of the data used for screening 1 000 features and choice of λ with cross-validation; Incomplete U-statistics estimator of size 20 and partial target
- HSIC-Lasso selects 13 features; 9 of them are significant
- Found potentially new molecular signatures; Confidence statement on selected features

Gene	p-value
ACTB	0.961
IGJ	0.001
CD14	0.026
LYZ	0.001
FCER1A	0.001
MTRNR2L2	0.420
FCGR3A	0.001
RPS3A	0.001
FTL	0.968
TMSB4X	0.012
HLA-DPA1	0.001
TVAS5	0.553
IFI30	0.002

References

[1] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory. 16th international conference, ALT 2005*. Berlin: Springer, 2005.

[2] Makoto Yamada, Yuta Umezu, Kenji Fukumizu, and Ichiro Takeuchi. Post Selection Inference with Kernels. Proceedings of Machine Learning Research, 2018.

[3] Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric P. Xing, and Masashi Sugiyama. High-dimensional feature selection by feature-wise kernelized Lasso. *Neural Computation*, 2014.

[4] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the Lasso. *The Annals of Statistics*, 2016.

[5] Jen Ning Lim, Makoto Yamada, Wittawat Jitkrittum, Yoshikazu Terada, Shigeyuki Matsui, and Hidetoshi Shimodaira. More Powerful Selective Kernel Tests for Feature Selection. Proceedings of Machine Learning Research, 2020.