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

Outline

Post-selection Inference - Toy Example

Linear regression model with 50 features and sample size 300.

$$Y_i = \sum_{j=1}^{50} X_{ij}\beta_j + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

Task: Select the 5 most influential features and construct 90% - confidence intervals for them.

Data generation: Draw standardnormal random numbers for X and ε, and set $\beta_j = 0$ for all $j \in \{1, \ldots, 50\}$.

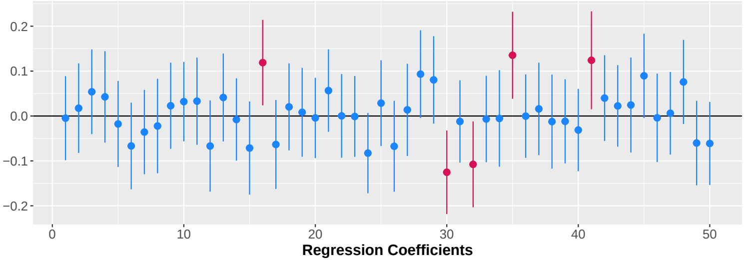
Post-selection Inference - Toy Example

Linear regression model with 50 features and sample size 300.

$$Y_i = \sum_{j=1}^{50} X_{ij}\beta_j + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

Task: Select the 5 most influential features and construct 90% - confidence intervals for them.

Data generation: Draw standardnormal random numbers for X and ε, and set $\beta_j = 0$ for all $j \in \{1, \ldots, 50\}$.



Post-selection Inference

Problem: Hypothesis tests and confidence interval can be seriously flawed when the information encapsulated in the selection is not accounted for.

Post-selection Inference

Problem: Hypothesis tests and confidence interval can be seriously flawed when the information encapsulated in the selection is not accounted for.
Solution: Inference conditional on selection information.

Post-selection Inference

Problem: Hypothesis tests and confidence interval can be seriously flawed when the information encapsulated in the selection is not accounted for.
Solution: Inference conditional on selection information.

In the example: $S = \{16, 30, 32, 35, 41\}, S^c = \{1, \dots, 50\} \setminus S$

$$\mathbb{P}\Big(\beta_{16} \in C \Big| |\hat{\beta}_{16}| \geq |\hat{\beta}_j| \ \forall j \in S^c\Big) \geq 0.9.$$

Post-selection Inference

Problem: Hypothesis tests and confidence interval can be seriously flawed when the information encapsulated in the selection is not accounted for.
Solution: Inference conditional on selection information.

In the example: $S = \{16, 30, 32, 35, 41\}, S^c = \{1, \ldots, 50\} \setminus S$

$$\mathbb{P}\left(\beta_{16} \in C \middle| |\hat{\beta}_{16}| \geq |\hat{\beta}_j| \ \forall\, j \in S^c\right) \geq 0.9.$$

More generally, we are interested in the distribution of $\eta^T Y | \{AY \leq b\}$ for $\eta \in \mathbb{R}^n, A \in \mathbb{R}^{q \times n}, b \in \mathbb{R}^q$.

Post-selection Inference with Polyhedral Lemma

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 $\mathcal{N}(0, 1)$.

Post-selection Inference with Polyhedral Lemma

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 (Polyhedral Lemma, Lee et al. 2016)

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

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

where $z = \left(\mathrm{Id} - (\eta^T\Sigma\eta)^{-1}\Sigma\eta\eta^T\right)Y$ and $\mathcal{V}^-$ and $\mathcal{V}^+$ are known.

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

Hilbert-Schmidt Independence Criterion (HSIC)

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 MMD-distance in RKHS

Definition (HSIC, Gretton et al. 2005)

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

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

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

Hilbert-Schmidt Independence Criterion (HSIC)

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 MMD-distance in RKHS

Definition (HSIC, Gretton et al. 2005)

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!
- $\mathrm{HSIC}(X,Y) \geq 0, \quad \mathrm{HSIC}(X,Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y.$
- Classification and regression settings with suitable kernels possible.

HSIC estimators I

Suppose that we are given an i.i.d. sample $\{y_i, x_i\}_{i=1}^n$ and define K and L by $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$ for $i, j \in \{1, \ldots, n\}$. $\tilde{K} = K -$ diag(K), $\tilde{L} = L -$ diag(L) and $\Gamma = \mathrm{Id} - \frac{1}{n} 1 1^T$.

Biased estimator (Gretton et al. 2005):

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

Unbiased estimator (Song et al. 2012):

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

HSIC estimators I

Suppose that we are given an i.i.d. sample $\{y_i, x_i\}_{i=1}^n$ and define K and L by $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$ for $i, j \in \{1, \ldots, n\}$. $\tilde{K} = K - \text{diag}(K)$, $\tilde{L} = L - \text{diag}(L)$ and $\Gamma = \text{Id} - \frac{1}{n} 1 1^T$.

Biased estimator (Gretton et al. 2005):

$$\widehat{\text{HSIC}}_{\text{b}}(X, Y) = (n-1)^{-2} \operatorname{tr}(K \Gamma L \Gamma)$$

Unbiased estimator (Song et al. 2012):

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

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

HSIC estimators II

Block estimator (Zhang et al. 2018):
Divide sample into blocks of size B, $\{\{y_i^b, x_i^b\}_{i=1}^{B}\}_{b=1}^{n/B}$.

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

HSIC estimators II

Block estimator (Zhang et al. 2018):
Divide sample into blocks of size B, $\{\{y_i^b, x_i^b\}_{i=1}^B\}_{b=1}^{n/B}$.

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

Incomplete U-statistics estimator (Lim et al. 2020):
HSIC is a U-statistic of degree 4, i.e. there exists h such that $\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X, Y) = \binom{n}{4}^{-1} \sum_{(i,j,q,r) \in \mathcal{S}_{n,4}} h(i, j, q, r)$, 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. 2018):
Divide sample into blocks of size B, $\{\{y_i^b, x_i^b\}_{i=1}^B\}_{b=1}^{n/B}$.

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

Incomplete U-statistics estimator (Lim et al. 2020):
HSIC is a U-statistic of degree 4, i.e. there exists h such that $\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X,Y) = \binom{n}{4}^{-1}\sum_{(i,j,q,r)\in\mathcal{S}_{n,4}} h(i,j,q,r)$, 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

Goal: Use HSIC to select *non-redundant* features.

Let $\bar{L} = \Gamma L \Gamma$ and $\bar{K}^{(j)} = \Gamma K^{(j)} \Gamma, j \in \{1, \ldots, p\}$. The HSIC-Lasso (Yamada et al. 2014) solution is given by

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

$$= \underset{\beta \geq 0}{\operatorname{argmin}} - \sum_{j=1}^{p} \beta_j \widehat{\mathrm{HSIC}}_{\mathrm{b}}(X^{(j)}, Y) + \frac{1}{2} \sum_{i,j=1}^{p} \beta_i \beta_j \widehat{\mathrm{HSIC}}_{\mathrm{b}}(X^{(i)}, X^{(j)}) + \lambda \|\beta\|_1$$

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

Post-selection Inference with HSIC-Lasso

Goal: Create PSI-procedure for HSIC-Lasso

- Version of Polyhedral Lemma for asymptotically normal random variables
- Asymptotically normal HSIC-Lasso
- Expression for inference targets
- Characterisation of selection in affine linear way

Normal HSIC-Lasso and Inference Targets

We replace the biased estimator with the block or the incomplete U-statistics estimator, for example

$$\hat{\beta} = \operatorname*{argmin}_{\beta \geq 0} - \sum_{j=1}^{p} \beta_j \widehat{\mathrm{HSIC}}_{\mathrm{block}}(X^{(j)}, Y) + \frac{1}{2} \sum_{i,j=1}^{p} \beta_i \beta_j \widehat{\mathrm{HSIC}}(X^{(i)}, X^{(j)}) + \lambda \|\beta\|_1$$

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

where $H_j = \widehat{\mathrm{HSIC}}_{\mathrm{block}}(X^{(j)}, Y)$ and $M_{ij} = \widehat{\mathrm{HSIC}}(X^{(i)}, X^{(j)})$. We define the selection procedure as $\hat{S} := \{j : \hat{\beta}_j > 0\}$, denote its value by S and set $S^c = \{1, \ldots, p\} \setminus S$. Moreover, we assume that M is positive definite.

Normal HSIC-Lasso and Inference Targets

We replace the biased estimator with the block or the incomplete U-statistics estimator, for example

$$\hat{\beta} = \underset{\beta\geq 0}{\operatorname{argmin}} - \sum_{j=1}^{p} \beta_j \widehat{\mathrm{HSIC}}_{\mathrm{block}}(X^{(j)}, Y) + \frac{1}{2}\sum_{i,j=1}^{p} \beta_i \beta_j \widehat{\mathrm{HSIC}}(X^{(i)}, X^{(j)}) + \lambda \|\beta\|_1$$

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

where $H_j = \widehat{\mathrm{HSIC}}_{\mathrm{block}}(X^{(j)}, Y)$ and $M_{ij} = \widehat{\mathrm{HSIC}}(X^{(i)}, X^{(j)})$. We define the selection procedure as $\hat{S} := \{j : \hat{\beta}_j > 0\}$, denote its value by S and set $S^c = \{1, \ldots, p\} \setminus S$. Moreover, we assume that M is positive definite.

Partial target: 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-target: $H_j = \mathrm{e}_j^T H =: \eta^T H$.

Affine Linear Selection

Partial target:
Similarly to linear regression with Lasso-regularisation, the selection event can be characterised using the Karush-Kuhn-Tucker (KKT) 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-target:
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 $\mathcal{V}^+$:

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

Testing

For all $j \in S$, we conduct the tests

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

The p-value is given by $p = 1 - F_{0,\eta^T \Sigma \eta}^{[\mathcal{V}^-,\mathcal{V}^+]}(\eta^T H)$, where η is set according to the target.

Practical Application

Challenges

- Positive definiteness of M: positive definite approximation
- Computational costs of HSIC-estimation: screen for relevant features entering HSIC-Lasso
- Choice of hyper-parameter: set data aside to estimate λ via cross-validation or AIC

Practical Application

Challenges

- Positive definiteness of M: positive definite approximation
- Computational costs of HSIC-estimation: screen for relevant features entering HSIC-Lasso
- Choice of hyper-parameter: set data aside to estimate λ via cross-validation or AIC

Outline of algorithm

- Split data into two folds
- 1st fold:
  - Screening of relevant features
  - Estimation of λ
- 2nd fold:
  - Computing H and M
  - HSIC-Lasso estimate $\hat{\beta}$ and obtaining selected indices S
  - Post-selection inference for targets

Toy Models

Type-I error:

$$\text{(M1)} \quad \begin{gathered} Y \sim \mathsf{Ber}\Big(g\big(\textstyle\sum_{i=1}^{10} X_i\big)\Big), \quad X \sim \mathcal{N}(0_{50}, \Xi), \\ g(x) = \mathrm{e}^x/(1+\mathrm{e}^x), \end{gathered}$$

$$\text{(M2)} \quad \begin{gathered} Y = \textstyle\sum_{i=1}^{5} X_i X_{i+5} + \varepsilon, \quad X \sim \mathcal{N}(0_{50}, \Xi), \\ \varepsilon \sim \mathcal{N}(0, \sigma^2), \end{gathered}$$

where Ξ is either set to Id or $\Xi_{ij} = 0.5^{|i-j|}$, and σ^2 is chosen to be a fifth of the variance in the X-terms.

Power: We replace X_1 by θX_1 in model (M1) and denote it (M1') and introduce

$$\text{(M3)} \quad \begin{gathered} Y = \theta X_1 + \textstyle\sum_{i=2}^{10} X_i + \varepsilon, \quad X \sim \mathcal{N}(0_{50}, \mathrm{Id}), \\ \varepsilon \sim \mathcal{N}(0, \sigma^2), \end{gathered}$$

$$\text{(M4)} \quad \begin{gathered} Y = \theta\, h(X_1) + \textstyle\sum_{i=2}^{10} X_i + \varepsilon, \quad X \sim \mathcal{N}(0_{50}, \mathrm{Id}), \\ h(x) = x - x^3, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2). \end{gathered}$$

Type-I Error and Power

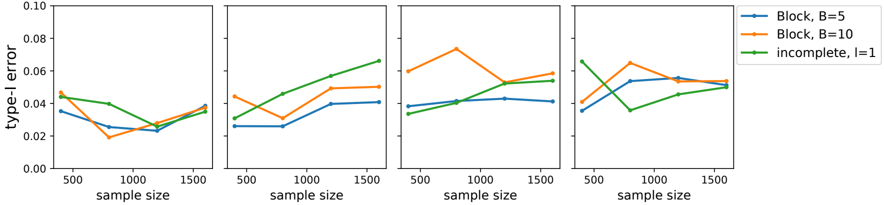


Figure: Type-I error for the HSIC-target: (M1) (1st and 2nd), (M2) (3rd and 4th)

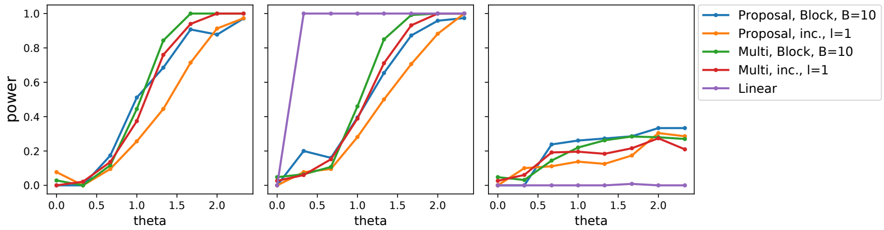


Figure: Power of detecting X_1: (M1'), (M3), (M4)

Performance on 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

Potential Future Work

- Wider investigation of method, e.g. split ratio, size of estimators, estimation of λ, behaviour for correlated features
- Development of/ Integration into a Python-package
- Application to more datasets (analysis of Turkish Student and Communities & Crimes data in the paper and supplement)
- Integration of screening and hyper-parameter estimation in PSI-procedure
- Improvement through novel ideas in PSI

Thank you for your attention!

Paper: Proceedings of ICML 2021 and on arXiv (2010.15659)

Code: Github `tobias-freidling/hsic-lasso-psi`

Slides: Website `tobias-freidling.onrender.com`

References

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

Lee, Jason D., Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor (2016). "Exact post-selection inference, with application to the Lasso.". In: *The Annals of Statistics*.

Lim, Jen Ning, Makoto Yamada, Wittawat Jitkrittum, Yoshikazu Terada, Shigeyuki Matsui, and Hidetoshi Shimodaira (2020). "More Powerful Selective Kernel Tests for Feature Selection". In: Proceedings of Machine Learning Research.

Song, Le, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt (2012). "Feature selection via dependence maximization.". In: *Journal of Machine Learning Research (JMLR)*.

Yamada, Makoto, Wittawat Jitkrittum, Leonid Sigal, Eric P. Xing, and Masashi Sugiyama (2014). "High-dimensional feature selection by feature-wise kernelized Lasso.". In: *Neural Computation*.

Zhang, Qinyi, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic (2018). "Large-scale kernel methods for independence testing.". In: *Statistics and Computing*.