Post-Selection Inference with HSIC-Lasso

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

Abstract

Detecting influential features in non-linear and/or high-dimensional data is a challenging and increasingly important task in machine learning. Variable selection methods have thus been gaining much attention as well as post-selection inference. Indeed, the selected features can be significantly flawed when the selection procedure is not accounted for. We propose a selective inference procedure using the so-called model-free "HSIC-Lasso" based on the framework of truncated Gaussians combined with the polyhedral lemma. We then develop an algorithm, which allows for low computational costs and provides a selection of the regularisation parameter. The performance of our method is illustrated by both artificial and real-world data based experiments, which emphasise a tight control of the type-I error, even for small sample sizes.

1. Introduction

The choice of a relevant statistical model in light of the observations prior to any statistical inference is ubiquitous in statistics. This reduces computational costs and fosters parsimonious models. For example, in linear regression analysis, allowing for a subset of features to enter the model, i.e. the sparsity assumption, is particularly suited to high-dimensional data, which potentially provides more robust predictions. Yet, should one follow the standard procedure, which assumes a model specified a priori, the data-driven nature of the selection procedure would be tacitly overlooked. Therefore, applying classical inference methods may entail seriously flawed results, as it was highlighted by Leeb & Pötscher (2005; 2006), among others.

[1]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, United Kingdom [2]Graduate School of Economics, Osaka University, Osaka, Japan [3]Center for Advanced Intelligence Project (AIP), RIKEN, Kyoto, Japan [4]Graduate School of Informatics, Kyoto University, Kyoto, Japan. Correspondence to: Tobias Freidling <taf40@cam.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

The most straightforward remedy is sample splitting, which uses one part of the data for model selection and the other part for inference, cf. (Cox, 1975). However, this approach leaves space for improvement as the selection-data is outright disregarded at the inference step. In particular, two major paradigms regarding the treatment of model selection have evolved. Berk et al. (2013) developed a post-selection inference (PSI) method, which enables to circumvent distorting effects inherent to any selection procedure. Yet, in practice this method often leads to overly conservative confidence intervals and entails high computational costs.

In contrast to this work, Lee et al. (2016) and Tibshirani et al. (2016) only accounted for the actual selection outcome by conditioning on the latter at the inference step. This requires control over or at least insight into the selection procedure in order to characterise the selection event. Under the Gaussian assumption, it is possible to express this event as a restriction on the distribution of the inference target, using the so-called polyhedral lemma, and derive a pivotal quantity. Due to its comparatively low computational cost and general set-up, this method was applied with great success to a variety of model selection approaches, e.g. (Hyun et al., 2018) and (Taylor & Tibshirani, 2018).

Variable selection procedure methods typically include step-wise procedures - see (Hocking, 1976) -, information criteria, with the AIC and its extensions - see (Akaike, 1974) -, and regularisation methods, pioneered by Tibshirani's work on the Lasso (1996). Their limitations, especially regarding assumptions such as linearity or certain probability distributions, fostered a flourishing research on model-free feature selection. The kernel-based Hilbert-Schmidt independence criterion (HSIC) (2005), proposed by Gretton et al., emerged as a key ingredient as it enables to quantify the dependence between two random variables. This allows for simple feature selection: For instance, one can select a fixed number of covariates that exhibit the highest HSIC-estimates with the response, which is referred to as HSIC-ordering in the rest of the paper. HSIC-Lasso (2014), proposed by Yamada et al., additionally accounts for the dependence structure among the covariates and selects features with an L^1-penalty. Yamada et al. (2018) developed a selective inference approach for the former selection procedure; the analogous results for HSIC-Lasso, however, are still an important blind spot in PSI.

In this study, we are interested in the following problem: Given n observations of a response variable and potential features, select the relevant covariates using the HSIC-Lasso screening procedure; then, based on the polyhedral lemma, develop asymptotic inference *given* the Lasso selection procedure in the same spirit as Lee et al. (2016). Our contributions are as follows: First, we derive a tailored, novel asymptotic pivot and characterise the selection event of HSIC-Lasso in an affine linear fashion; then we propose an algorithm that solves cumbersome issues arising in applications, such as high computational costs and hyper-parameter choice; finally, to illustrate our theoretical results and the relevance of the proposed method, we conduct an empirical analysis using artificial and real-world data. To the best of our knowledge, this work is the first approach to tackle the question of PSI with HSIC-Lasso.

2. Background

In this section the two theoretical cornerstones which our work is founded on - namely PSI based on truncated Gaussians and the Hilbert-Schmidt independence criterion - are reviewed.

2.1. PSI with Truncated Gaussians

We first review the PSI-approach (2016), which was pioneered by Lee et al. and which considers the distribution of the quantity of interest, alias target, conditionally on a selection event at the inference step. We denote the set of potential models $\mathcal{S}$, the model estimator $\hat{S}$ and suppose that the response Y follows the distribution $\mathcal{N}(\mu, \Sigma)$, where μ is unknown and Σ is given. It is assumed that the inference target can be expressed as $\eta_S^T \mu$ for a vector η_S that can depend on the previously chosen model S. Consequently, the distribution of interest is $\eta_S^T Y | \{\hat{S} = S\}$. If the selection event allows for an affine linear representation, i.e. $\{\hat{S} = S\} = \{AY \leq b\}$, with A and b only depending on S and the covariates, a specific model choice can be seen as a restriction on the distribution of the inference target. This is the object of the following result.

Lemma 1 (Polyhedral lemma)**.** *Let $Y \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$, $\eta \in \mathbb{R}^n$, $A \in \mathbb{R}^{k \times n}$ and $b \in \mathbb{R}^k$. Defining $Z := \left(\mathrm{Id} - C\eta^T\right) Y$ and $C := (\eta^T \Sigma \eta)^{-1} \Sigma \eta$, then*

$$\{AY \leq b\} = \left\{\mathcal{V}^-(Z) \leq \eta^T Y \leq \mathcal{V}^+(Z)\right\},$$

holds almost surely with

$$\mathcal{V}^-(Z) := \max_{j:(AC)_j < 0} \frac{b_j - (AZ)_j}{(AC)_j}, \tag{1}$$

$$\mathcal{V}^+(Z) := \min_{j:(AC)_j > 0} \frac{b_j - (AZ)_j}{(AC)_j}. \tag{2}$$

Hence, selection restricts the values the inference target can take, which gives rise to the definition of truncated Gaussians.

Definition 2. Let $\mu \in \mathbb{R}$, $\sigma^2 > 0$ and $a, b \in \mathbb{R}$ such that $a < b$. Then the cumulative distribution function of a *Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ truncated to the interval* $[a, b]$ is given by

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

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

Concluding the line of thought, we are now able to state a pivotal quantity that can be used for inference.

Theorem 3. *Under the assumptions of Lemma 1 it holds that*

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

where $\mathcal{V}^-$ and $\mathcal{V}^+$ are given by (1) and (2), respectively.

This result corresponds to Theorem 5.2 of (Lee et al., 2016), which characterises the distribution of $\eta^T Y$ conditionally on Y belonging to the polyhedron $\{AY \leq b\}$.

2.2. Hilbert-Schmidt Independence Criterion

Gretton et al. proposed the Hilbert-Schmidt independence criterion (2005) as a model-free measure for the dependence between two random variables. It relies on embedding probability measures $\mathbb{P}$ into a reproducing kernel Hilbert space $\mathcal{H}_k$ with associated kernel function k. If k is universal, i.e. $\mathcal{H}_k$ is dense in the space of continuous functions, then the embedding $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ is injective. Hence, the squared distance between $\mu_k(\mathbb{P})$ and $\mu_k(\mathbb{Q})$ (maximum mean discrepancy) forms a metric so that $\mathbb{P} = \mathbb{Q} \Leftrightarrow \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}^2 = 0$ for any probability measures $\mathbb{P}$ and $\mathbb{Q}$; see (Smola et al., 2007) for further details on RKHSs. The Gaussian kernel, which is universal, is probably the most commonly used kernel, and is defined as

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \quad \sigma^2 > 0.$$

A way to detect the dependence between two random variables X and Y is the comparison between the joint distribution $\mathbb{P}_{X,Y}$ and the product of the marginals $\mathbb{P}_X \mathbb{P}_Y$ using the maximum mean discrepancy. This metric is precisely the Hilbert-Schmidt independence criterion, which can also be defined in terms of the involved kernels as follows.

Definition 4. Let X and Y be random variables, X' and Y' independent copies, and k and l be bounded kernels. The *Hilbert-Schmidt independence criterion* $\mathrm{HSIC}_{k,l}$ is given

by

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

Several approaches for estimating the Hilbert-Schmidt independence criterion from a data sample $\{(x_j, y_j)\}_{j=1}^n$ have been put forward. Gretton et al. (2005) proposed a V-statistic based estimator $\widehat{\mathrm{HSIC}}_{\mathrm{b}}(X,Y)$, which is biased, whereas Song et al. (2012) provided an unbiased U-statistic version $\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X,Y)$. Regarding the asymptotic distribution, Zhang et al. (2018) established that both these estimators scaled by $n^{1/2}$ converge to a Gaussian distribution if X and Y are dependent. However for $X \perp\!\!\!\perp Y$, the asymptotic distribution is not normal. Since the true dependence between X and Y is unknown, we focus on estimators that are asymptotically normal in either case.

In this respect, it is helpful to express the unbiased version of the HSIC-estimator as a U-statistic of degree 4 with kernel function h provided, e.g. in Theorem 3 in Song et al. (2012). Zhang et al. (2018) and Lim et al. (2020) suggested the following estimators.

Definition 5. Let $B \in \mathbb{N}$ and subdivide the data into folds of size B, $\{\{(x_i^b, y_i^b)\}_{i=1}^B\}_{b=1}^{n/B}$. The *block estimator* $\widehat{\mathrm{HSIC}}_{\mathrm{block}}$ with block size B is given by

$$\widehat{\mathrm{HSIC}}_{\mathrm{block}}(X,Y) = \frac{1}{n/B}\sum_{b=1}^{n/B} \widehat{\mathrm{HSIC}}_{\mathrm{u}}(X^b, Y^b), \quad (3)$$

where $\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X^b, Y^b)$ denotes the unbiased estimate on the data $\{(x_i^b, y_i^b)\}_{i=1}^B$. Let $\mathcal{S}_{n,4}$ be the set of all 4-subsets of $\{1,\dots,n\}$ and let $\mathcal{D}$ be a multiset containing m elements of $\mathcal{S}_{n,4}$ randomly chosen with replacement. Further, suppose $m = \mathcal{O}(n)$ and define $l := \lim_{n,m\to\infty} m/n$. The *incomplete U-statistic estimator* $\widehat{\mathrm{HSIC}}_{\mathrm{inc}}$ of size l is defined by

$$\widehat{\mathrm{HSIC}}_{\mathrm{inc}}(X,Y) = \frac{1}{m}\sum_{(i,j,q,r)\in\mathcal{D}} h(i,j,q,r). \quad (4)$$

Both estimators are asymptotically normal.

Theorem 6. *Let $\{(x_j^{(1)},\dots,x_j^{(p)}, y_j)\}_{j=1}^n$ be an i.i.d. data sample and define $H_0 = (\mathrm{HSIC}(X^{(1)},Y),\dots$ $\dots,\mathrm{HSIC}(X^{(p)},Y))^T$, and H_{block} and H_{inc} accordingly. Assume that B and l are the same for all entries of H_{block} and H_{inc}, respectively, let $n/B \to \infty$ and choose $\mathcal{D}$ for all elements of H_{inc} independently. Then*

$$\begin{aligned}\sqrt{n/B}\big(H_{\mathrm{block}} - H_0\big) &\xrightarrow{D} \mathcal{N}(0, \Sigma_{\mathrm{block}}), \\ \sqrt{m}\big(H_{\mathrm{inc}} - H_0\big) &\xrightarrow{D} \mathcal{N}(0, \Sigma_{\mathrm{inc}}),\end{aligned}$$

with positive definite matrices Σ_{block} and Σ_{inc}.

The first statement is a direct consequence of the multidimensional central limit theorem; to prove the second asymptotic result, we use the framework of U-statistics. The details are deferred to the supplementary material.

3. PSI for HSIC-Lasso

This section contains our main theoretical results for PSI with HSIC-Lasso and addresses the difficulties arising in practical applications.

3.1. Feature Selection with HSIC-Lasso

Yamada et al. introduced the model-free HSIC-Lasso (2014) feature selection method as follows.

Definition 7. Let $\{(x_j^{(1)},\dots,x_j^{(p)}, y_j)\}_{j=1}^n$ be an i.i.d. data sample, k and l be kernels, let K and L be given by $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$ for $i,j \in \{1,\dots,n\}$ and set $\Gamma = \mathrm{Id} - \frac{1}{n}11^T$. Denoting $\bar{L} = \Gamma L \Gamma$ and $\bar{K}^{(s)} = \Gamma K^{(s)}\Gamma$, $\hat{\beta}$ is given by

$$\hat{\beta} = \underset{\beta\in\mathbb{R}_+^p}{\mathrm{argmin}}\ \frac{1}{2}\|\bar{L} - \sum_{s=1}^p \beta_s \bar{K}^{(s)}\|_{\mathrm{Frob}}^2 + \lambda\|\beta\|_1,$$

where $\lambda > 0$ and $\mathbb{R}_+ = [0,\infty)$. The *HSIC-Lasso selection procedure* is defined by $\hat{S} = \{s\colon \hat{\beta}_s > 0\}$.

For our purposes, however, we consider the following alternative representation of $\hat{\beta}$:

$$\begin{aligned}\hat{\beta} = \underset{\beta\in\mathbb{R}_+^p}{\mathrm{argmin}}\ & -\sum_{s=1}^p \beta_s\, \widehat{\mathrm{HSIC}}_{\mathrm{b}}(X^{(s)}, Y) \\ & + \frac{1}{2}\sum_{s,r=1}^p \beta_s\beta_r\, \widehat{\mathrm{HSIC}}_{\mathrm{b}}(X^{(s)}, X^{(r)}) + \lambda\|\beta\|_1. \quad (5)\end{aligned}$$

It becomes apparent that feature selection is driven by three competing components. Considering the first and third term together, we notice that covariates with high dependence to the response achieve positive $\hat{\beta}$-values, whereas the $\hat{\beta}$-entries of non-influential features are forced to zero by the regularisation term. Moreover, the second term penalises the selection of covariates showing high dependence on other features.

Taking (5) as a starting point, we replace the biased V-statistic based estimators with asymptotically normal ones and allow for a more general weighted Lasso-penalty.

Definition 8. Let H be an asymptotically Gaussian and $\tilde{\mathrm{H}}$ be any HSIC-estimator and define H and M by $H_s = \mathrm{H}(X^{(s)}, Y)$, $M_{sr} = \tilde{\mathrm{H}}(X^{(s)}, X^{(r)})$ for $s,r \in \{1,\dots,p\}$. The *normal weighted HSIC-Lasso selection procedure* is given by $\hat{S} = \{s\colon \hat{\beta}_s > 0\}$ with

$$\hat{\beta} = \underset{\beta\in\mathbb{R}_+^p}{\mathrm{argmin}}\ -\beta^T H + \frac{1}{2}\beta^T M\beta + \lambda\,\beta^T w, \quad (6)$$

where $w \in \mathbb{R}^p_+$ is a fixed weight vector.

Using the asymptotically normal response H, the framework of the polyhedral lemma can be applied; however, we need to provide an asymptotic pivot.

Theorem 9. *Let $(H_n)_{n\in\mathbb{N}}$, $(M_n)_{n\in\mathbb{N}}$ and $(\widehat{\Sigma}_n)_{n\in\mathbb{N}}$ be sequences of random vectors and matrices, respectively, such that $H_n \to \mathcal{N}(\mu, \Sigma)$ in distribution, $M_n \to M$, $\widehat{\Sigma}_n \to \Sigma$ almost surely and $\widehat{\Sigma}_n \perp\!\!\!\perp (H_n, M_n)$. For a selected model S and $\tilde{S} \subseteq S$, let $\eta_{\tilde{S}}$, $A_{\tilde{S}}$ and $b_{\tilde{S}}$ be a.s. continuous functions of M and assume that the selection events are given by $\{A_{\tilde{S}}(M_n)H_n \leq b_{\tilde{S}}(M_n)\}$ and that $\operatorname{int}(\{A_{\tilde{S}}(M)H \leq b_{\tilde{S}}(M)\}) \neq \emptyset$. Then*

$$F^{[\mathcal{V}^-(Z_n),\mathcal{V}^+(Z_n)]}_{\eta_n^T\mu,\eta_n^T\widehat{\Sigma}_n\eta_n}(\eta_n^T H_n) | \{A_n H_n \leq b_n\} \xrightarrow{D} \operatorname{Unif}(0,1), \tag{7}$$

as $n \to \infty$ where $\eta_n = \eta_{\tilde{S}}(M_n)$, $A_n = A_{\tilde{S}}(M_n)$ and $b_n = b_{\tilde{S}}(M_n)$.

This statement is tailored to selection with normal weighted HSIC-Lasso and generalises Theorem 3 as it relaxes the requirements of a normal response and known covariance Σ: Under the assumption of an asymptotically Gaussian response and a consistent estimator for Σ, we obtain an asymptotic pivot.
To prove this result, we show that $(H_n, M_n, \widehat{\Sigma}_n) | \{A_n H_n \leq b_n\} \to (H, M, \Sigma) | \{AH \leq b\}$ in distribution and then apply the continuous mapping theorem to the a.s. continuous function F. The details of the proof can be found in the supplement.

3.2. Inference Targets and Selection Event

We now define the inference targets and characterise the selection events of the normal weighted HSIC-Lasso. To do so, we first introduce the following notations. For any matrix $B \in \mathbb{R}^{q\times q}$, $v \in \mathbb{R}^q$ and index sets $I, J \subset \{1, \dots, q\}$, we define $I^c := \{1, \dots, q\} \setminus I$. Moreover, v_I contains all entries at positions in I and $B_{IJ} \in \mathbb{R}^{|I|\times|J|}$ is given by the rows and columns of B whose indices are in I and J, respectively.
For a selected model S and $j \in S$, we consider the HSIC-target $H_j := \mathrm{e}_j^T H$ and the partial target $\hat{\beta}^{\text{par}}_{j,S} := \mathrm{e}_j^T M_{SS}^{-1} H_S$, where e_j denotes the j-th unit vector. The former target describes the dependence between response Y and feature $X^{(j)}$. In the same spirit of a partial regression coefficient, the latter can be interpreted as the degree of influence of $X^{(j)}$ on Y adjusted to the dependence structure among the covariates. Expressing both targets in the form of $\eta_S^T H$, the respective η-vectors for the HSIC- and partial target are e_j and $(M_{SS}^{-1} \,|\, 0)^T \mathrm{e}_j$.
We notice that the HSIC-target is influenced by the selection information $\{j \in \hat{S}\}$ only, whereas the partial target is, by definition, affected by the entire chosen set of covariates $\{\hat{S} = S\}$. We characterise these selection events as follows.

Theorem 10. *Assume the same framework as in Definition 8, suppose that M is positive definite and let $\eta \in \mathbb{R}^p$. Then $\{\hat{S} = S\} = \{A\,(H_S, H_{S^c})^T \leq b\}$ holds, where*

$$A = -\frac{1}{\lambda}\begin{pmatrix} M_{SS}^{-1} & | & 0 \\ M_{S^cS}M_{SS}^{-1} & | & \mathrm{Id} \end{pmatrix}, \quad b = \begin{pmatrix} -M_{SS}^{-1} w_S \\ w_{S^c} - M_{S^cS}M_{SS}^{-1} w_S \end{pmatrix}, \tag{8}$$

and 0 denotes a matrix of size $|S| \times |S^c|$ filled with zeros. Moreover, $\{j \in \hat{S}\} = \{AH \leq b\}$ holds for

$$A = -\mathrm{e}_j^T, \quad b = -\mathrm{e}_j^T M \hat{\beta}_{-j} - \lambda w_j, \tag{9}$$

where $\hat{\beta}_{-j}$ denotes $\hat{\beta}$ with the j-th entry set to zero.

To prove these statements, we use the Karush-Kuhn-Tucker (KKT) conditions, which characterise the solution of (6) by a set of inequalities. Manipulating these, we obtain the affine linear representation of $\{\hat{S} = S\}$ and $\{j \in \hat{S}\}$. The details can be found in the supplementary material.
Theorem 10 is the key result that allows us to carry out post-selection inference with HSIC-Lasso. In summary, we have to consider the distributions $\mathrm{e}_j^T H \,|\, \{j \in \hat{S}\}$ and $\mathrm{e}_j^T (M_{SS}^{-1} H \,|\, 0) \,|\, \{\hat{S} = S\}$ for the HSIC- and partial target, respectively, in order to account for the selection. With the affine linear representations (8) and (9), we can apply Theorem 9 and get an asymptotic pivot for inference.

3.3. Practical Applications

Equipped with these theoretical results, we now propose an algorithm that handles difficulties arising in practical applications.

3.3.1. Positive Definiteness

Theorem 10 requires M to be positive definite. For the original version of HSIC-Lasso (5), this condition is always fulfilled by the structure of the biased HSIC-estimates. However, there is no guarantee for other estimation procedures. For this reason, we project M onto the space of positive definite matrices, as proposed by Higham (1988): The spectral decomposition of M is computed and all negative eigenvalues are replaced with a small positive value $\varepsilon > 0$.

3.3.2. Computational Costs

HSIC-Lasso is frequently applied to high-dimensional data where the number of covariates p exceeds the sample size n. The resulting high computational costs are caused by the calculation of the HSIC-estimates, where H grows as $\mathcal{O}(p)$ and M as $\mathcal{O}(p^2)$. Therefore, we introduce an upstream screening stage identifying a subset of potentially influential features so that HSIC-Lasso only has to deal with these. Following the approach of Yamada et al. (2018), we compute

the HSIC-estimates $\widehat{\text{HSIC}}(Y, X^{(j)}), j \in \{1, \dots, p\}$, and select a pre-fixed number $p' < p$ of the covariates having the highest estimates. We call this HSIC-ordering.
In order to ensure valid inference results, we have to adjust for the screening step as well because it affects feature selection. To do so, we split the data into two folds, one dedicated to screening, the other dedicated to HSIC-Lasso selection among the screened variables, cf. (Cox, 1975). Thus, potentially distorting effects of the screening step are separated from inference on the second fold. Moreover, unbiased HSIC-estimates can be used for screening, which are more precise than block or incomplete U-statistic estimates.

Remark. In future applications, random Fourier features could be used to speed up the kernel computation of the objective function (5), allowing for a larger p'. However, it is not immediately clear whether we can recover the theoretical guarantees for our method when using approximated kernel functions.

3.3.3. Hyper-parameter Choice

In practice, a suitable choice of the regularisation parameter λ and the weight vector w is key for meaningful results. Since the data generating process is unknown, we have to estimate these hyper-parameters. In order to prevent this from affecting inference results, we use the first fold for hyper-parameter selection. In doing so, we can apply any estimation method for λ, such as cross-validation, e.g. (Stone, 1974), or the AIC, cf. (Akaike, 1974), and get a valid procedure that is easy to implement. Moreover, we can employ Zou's adaptive Lasso penalty (2006), that uses the weight vector $w = 1/|\hat{\beta}|^\gamma$. γ is typically set to 1.5 or 2 and $\hat{\beta}$ is a $\sqrt{n}$-consistent estimator, e.g. the ordinary least squares estimator. Contrary to the vanilla Lasso, this method satisfies the oracle property, that is the sparsity-based estimator recovers the true underlying sparse model and has an asymptotically normal distribution. Yet, this property was only proven for a covariance matrix of the form $\Sigma = \sigma^2 \text{Id}$. Hence, we have to evaluate the usefulness of the adaptive Lasso in empirical simulations.

3.3.4. Algorithm

We summarise our proposed PSI method for a normal (weighted) HSIC-Lasso selection procedure as follows. To begin with, we split the data into two subsets. On the first fold, we compute the HSIC-estimates between all covariates and the response, determine the most influential p' features (screening) and estimate the hyper-parameters λ and w with the methods previously specified. On the second fold, we compute the estimates of H, M and $\widehat{\Sigma}$ for the screened features and solve the optimisation problem (6). For all $j \in \{1, \dots, p\}$ such that $\hat{\beta}_j > 0$, we find the truncation points for the specified targets with Theorem 10 and Lemma 1 and test with the asymptotic pivot (7) if the targets are significant at a given level α. The supplementary material contains a more detailed description of the algorithm in pseudocode.

Remark. Although Theorem 9 requires independence between the covariance- and the (H, M)-estimate we could not observe any detrimental influence if $\widehat{\Sigma}$ is computed as outlined above.

4. Experiments

In this section, we illustrate our theoretical contribution and the proposed algorithm on artificial and real-world data. The source code for the experiments was implemented in Python and relies on Lim et al.'s `mskernel`-package (2020). We use the Lasso optimisation routines of `scikit-learn` which implements the cyclical gradient descent algorithm, cf. (Friedman et al., 2007), and the least angle regression algorithm (LARS) (2004), proposed by Efron et al.. The source code for the following experiments is available on Github: `tobias-freidling/hsic-lasso-psi`.

4.1. Artificial Data

We examine the achieved type-I error of the proposed algorithm, compare its power to other approaches for post-selection inference and briefly discuss additional experiments.
For continuous data, we use Gaussian kernels where the bandwidth parameter is chosen according to the median heuristic, cf. (Schölkopf & Smola, 2018); for discrete data with n_c samples in category c, we apply the normalised delta kernel which is given by

$$l(y, y') := \begin{cases} 1/n_c, & \text{if } y = y' = c, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, we use a quarter of the data for the first fold, select the hyper-parameter λ applying 10-fold cross-validation with MSE, use a non-adaptive Lasso-penalty and do not conduct screening as the number of considered features is already small enough. On the second fold, we estimate M with the block estimator of size $B = 10$ as it is computationally less expensive than the unbiased estimator and leads to similar results. The covariance matrix Σ of H is estimated based on the summands of the block (3) and incomplete U-statistic (4) estimator, respectively. To this end, we use the oracle approximating shrinkage (OAS) estimator (2010), which was presented by Chen et al. and is particularly tailored for high-dimensional Gaussian data. We fix the significance level at $\alpha = 0.05$ and simulate 100 datasets for each considered sample size.

4.1.1. TYPE-I ERROR

In order to simulate the achieved type-I error we use the toy models

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

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

where $0_{50} \in \mathbb{R}^{50}$ and $\Xi \in \mathbb{R}^{50\times 50}$, to generate the data. These are clearly non-linear and cover categorical and continuous responses. In (M2), we choose σ^2 such that the variance of ε is a fifth of the variance of the X-dependent terms of Y amounting to a noise-to-signal ratio of 0.2. As for the covariance matrix Ξ, two cases are considered: we either set $\Xi = \mathrm{Id}$ or use decaying correlation, i.e. $\Xi_{ij} = 0.5^{|i-j|}$.

We simulate datasets with sample sizes $n \in \{400, 800, 1200, 1600\}$ for all different settings of models and covariance matrices and estimate H with block estimators of sizes 5 and 10 as well as with an incomplete U-statistics estimator of size $l = 1$. Since the partial target both depends on the entire set of selected variables and its value cannot be directly inferred from the data-generating mechanism, it is inherently hard to rigorously assess the type-I error of any given partial target. (However, the false positive rate for testing different partial targets hints that the type-I error is probably close to 0.05). For this reason, we concentrate on the HSIC-target in our analysis. For both models $\text{HSIC}(Y, X^{(j)}) = 0, j \in \{11, \dots, 50\}$, holds which allows us to estimate the type-I error as the ratio of null hypothesis rejections and tests among the selected features with indices in $\{11, \dots, 50\}$. If influential variables are correlated with uninfluential ones, the HSIC-value is not precisely zero; nonetheless, the use of *decaying* correlation renders the bias of this effect ignorable. Figure 1 illustrates that the type-I error is close to 0.05 across all estimators and data generating mechanisms, even for small sample sizes.

4.1.2. POWER

In this set of experiments, we adapt the toy model (M1), replacing X_1 by θX_1 and setting $\Xi = \mathrm{Id}$, and denote it (M1'). Moreover, we introduce the following linear and modified linear model

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

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

where $\theta \in \mathbb{R}$. Our proposed algorithm is applied with both a block estimator, $B = 10$, and an incomplete U-statistics estimator, $l = 1$, and is compared with the so-called Multi PSI-approach, presented by Lim et al. (2020): We select k features with HSIC-ordering and carry out inference for the HSIC-targets. (Since M is not involved in the feature selection, we cannot define partial targets for HSIC-ordering.) It was empirically shown that multiscale bootstrapping (2004), which was first presented by Shimodaira and is abbreviated by Multi, is a more powerful PSI-approach for HSIC-ordering than truncated Gaussians. In our simulations, we set $k = 15$ and applied Multi with a block estimator, $B = 10$, as well as an incomplete U-statistics estimator, $l = 1$. Additionally, we applied Lee et al.'s original PSI-method (2016), that relies on Lasso-regularisation and assumes a linear regression setting, to (M3) and (M4). The inference target in this case is the partial regression coefficient.

We simulate datasets for values of θ in $\{0.00, 0.33, 0.67, 1.00, 1.33, 1.67, 2.00, 2.33\}$ and a sample size of $n = 800$, and compute the ratio of rejections of the null-hypothesis, i.e. the respective inference target corresponding to X_1 is zero, and the number of tests carried out. Plotting the obtained ratios against θ does not correspond to the usual depiction of the power function as θ is not the inference target for all considered procedures. However, this allows for an intuitive understanding of how strong X_1 needs to influence Y in order to be detected.

Figure 2 exhibits that the power of our proposed algorithm is similar to the Multi procedure, especially when using the block estimator. This confirms that, even without a manual choice of the number of selected features and costly bootstrap sampling, it is possible to match the best performing model-free PSI-methods. Moreover, we observe that regularised linear regression clearly outperforms our procedure as well as Multi for small values of θ if the data-generating process is indeed linear. However, when X_1 influences the outcome Y not only through a linear, but also through a cubic term, that is (M4), we discern that PSI based on a linear model has no power at all whereas model-agnostic methods still achieve noticeable power. This exemplifies that PSI procedures, built upon a certain model, can only be confidently used if there is limited uncertainty about the underlying data generation.

4.1.3. ADDITIONAL EXPERIMENTS

In our simulations, we focused on the statistical properties of the proposed approach but also started investigating the behaviour of the algorithm for different feature selection set-ups. The conclusions we drew deserve a few comments. ***Screening*** HSIC-ordering includes the influential features into the screened set with high probability when p' and the size of the first fold are sufficiently large. Nonetheless,

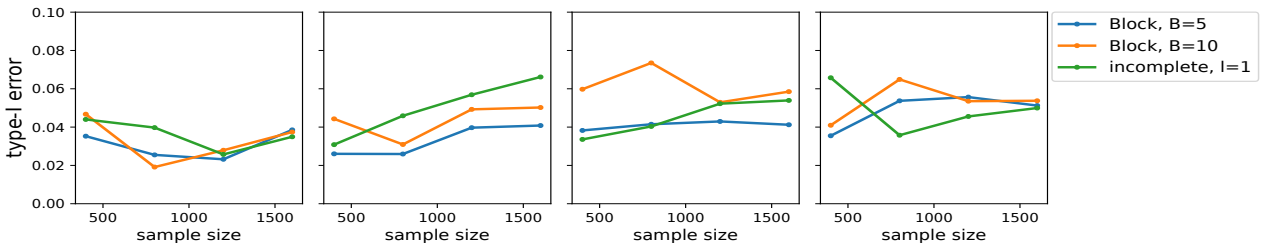

Figure 1. Type-I error for the HSIC-target in different toy models (from left to right): (M1) with $\Xi = \mathrm{Id}$, (M1) with $\Xi_{ij} = 0.5^{|i-j|}$, (M2) with $\Xi = \mathrm{Id}$, (M2) with $\Xi_{ij} = 0.5^{|i-j|}$

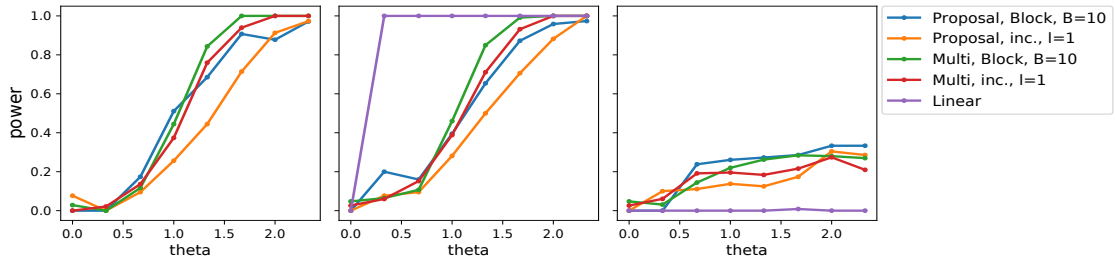

Figure 2. Power of detecting X_1 as influential feature in different toy models (from left to right): (M1'), (M3), (M4)

screening is merely a method to reduce computational complexity and can potentially weaken the performance of the downstream HSIC-Lasso procedure. Therefore, the number of screened features p' should be set as high as computational resources allow.

Regularisation The use of an adaptive penalty term generally leads to fewer selected features than the vanilla Lasso regularisation. Moreover, cross-validation and the AIC often choose similar values for λ.

Dependence structure In datasets with strong correlation between influential and uninfluential features we observe that the partial target is capable of correctly detecting the uninfluential ones. How the dependence structure among influential features materialises in rejections of the null hypothesis, however, is a subtle and still open question.

Estimators In general, block estimators are computationally less expensive than incomplete U-statistics estimators. For the latter ones, the calculation costs, but also the power increases with the size l.

Split ratio In the experiments with artificial data, the size of the first fold was set to be a third of the second fold as this already suffices to obtain a decent estimate of the regularisation parameter. For most datasets it is advisable to dedicate more data to the HSIC-Lasso procedure than to the hyper-parameter selection; however, a reliable heuristic for the split ratio remains subject to future research.

4.2. Benchmarks

Now we proceed to applying our proposed algorithm to benchmark datasets from the UCI Repository and the Broad Institute's Single Cell Portal, respectively. We provide an additional, more in-depth experiment in the supplementary materials.

4.2.1. Turkish Student Dataset

This dataset contains 5 820 course evaluation scores provided by students from Gazi University, Ankara, who answered 28 questions on a five-level Likert scale, see further (Gunduz & Fokoue, 2013). For our experiment we use the perceived difficulty of the course as response variable.

This data was previously evaluated by Yamada et al. (2018) who selected features with HSIC-ordering and used the familiar framework of the polyhedral lemma and truncated Gaussians for PSI, denoted by Poly. We use the block estimator of size 10 and set $k = 10$ for the Multi approach to accord with Yamada et al. and report their findings along with ours. For our proposal, the first fold contains 20% of the data, we select λ with 10-fold cross-validation and do not carry out screening as the number of features ($p = 28$) is manageable. Table 1 summarises our findings.

First, we notice that Multi and Poly pick different features despite sharing the same selection procedure which can probably be attributed to randomisation, carried out by Yamada et al. (2018). Moreover, we observe that HSIC-Lasso

Table 1. P-values of the HSIC-target for selected features in the Turkish student dataset

FEATURE	P-VALUE		
	PROPOSAL	MULTI	POLY
Q2	**0.021**	-	0.452
Q3	-	0.782	-
Q11	**0.004**	-	-
Q13	-	-	**0.018**
Q14	-	**0.001**	-
Q15	-	0.095	-
Q17	**<0.001**	**<0.001**	**0.033**
Q18	-	-	0.186
Q19	-	**<0.001**	-
Q20	-	**0.004**	0.463
Q21	-	**0.032**	**0.033**
Q22	-	**<0.001**	**0.042**
Q23	-	-	**0.037**
Q25	-	**0.002**	-
Q26	-	-	0.176
Q28	**0.004**	**0.041**	**<0.001**

chooses a very parsimonious model with only four covariates whose associated HSIC-targets are highly significant. Among the tested approaches, there is a moderate agreement on the influential covariates where only 'Q17: The Instructor arrived on time for classes.' and 'Q28: The Instructor treated all students in a right and objective manner.' are unanimously chosen and found to be significant. Considering the partial targets for HSIC-Lasso, we find that only $\hat{\beta}^{\text{par}}_{17,S}$ appears significant.

The different results that we obtain may hint that the Turkish student data set is intrinsically noisy or that methods based on HSIC-estimation and the polyhedral lemma are unstable. However, the Lasso-selection of HSIC-Lasso, unlike Multi or Poly, penalises correlated features which affects PSI as well. Hence, different results for the compared methods do not necessarily indicate incorrectness of either approach.

4.2.2. SINGLE-CELL RNASEQ DATA

Villani et al. (2017) isolated around 2 400 blood cells, enriched in two particular kinds of leukocytes: dendritic cells (DCs) and monocytes. Then, they measured the gene expression on every cell using single-cell RNAseq aiming to describe the diversity between, and within those two cell types based on their gene expression profile. They end up defining 10 different subclasses: 6 types of DCs and 4 types of monocytes. In our experiment, we use 1 078 samples of this data aspiring to find the genes that separate these 10 classes among the 26 593 genes. We standardise the single-cell RNAseq data gene-wise and impute missing gene expressions with MAGIC, see further (van Dijk et al., 2018). Since the response is categorical, we use the normalised delta kernel. Unlike the Turkish student dataset, we are now confronted with a considerably high-dimensional problem where the number of features greatly exceeds the sample size. Therefore, screening becomes more challenging and we consequently split the data evenly into first and second fold. We screen 1 000 potentially influential features and apply the incomplete U-statistic estimator with a large size of 20, hoping to better capture the potentially involved dependence structure. The remaining parameters were set as in the previous experiment. For an in-depth analysis of the dataset, we recommend to conduct a sensitivity analysis which investigates the behaviour of the method for different values of the parameters, such as the split ratio, the number of screened features or the size of the estimator.

Table 2. P-Values of the partial targets corresponding to selected features in the single-cell RNAseq dataset

GENE	P-VALUE	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**		

We find that HSIC-Lasso selects 13 features and that all of the associated HSIC-targets and most of the partial targets are significant, cf. Table 2. One of the traditionally defining characteristics of monocytes is the expression of the CD14 protein; encouragingly, HSIC-Lasso selected this gene as a discriminating feature. In fact, it also selected six other genes which Villani et al. used in multiple cell signatures: *FCGR3A* (DC4), *FCER1A* (DC2 and DC3), *FTL* (DC4), *IFI30* (cDC-like), *IGJ* (pDC-like), and *LYZ* (cDC-like). Jointly, this shows the ability of HSIC-Lasso to recover multiple genes used to define the classes.

More exciting, however, are the genes which are selected by HSIC-Lasso and were not used in the original study. These might point to new molecular signatures and functions that differentiate these cell types. For instance, one of these genes play a role in immunity: *HLA-DPA1*, which presenting cells like DCs use to present exogenous proteins to other immune cells. The proposed PSI framework adds nuance to this picture by providing a soft ranking of the selected genes. Notoriously, *FTL* and *ACTB* have p-values close to 1, which suggests that most of the information they provide is already captured by other selected genes. Hence, other genes should be prioritised in hypothetical downstream experiments.

Our method contributes to the detection of discriminatory features inasmuch it facilitates to formulate confidence statements about the obtained results.

5. Discussion and Outlook

Post-selection inference was initially proposed for linear regression models with Lasso regularisation and subsequently expanded to generalised regression situations and Lasso

variants, see for example (Taylor & Tibshirani, 2018) and (Hyun et al., 2018). However, knowledge of an underlying model seemed to be necessary in order to properly account for the selection-process. Yamada et al. (2018) overcame this limitation by capturing the unknown dependence between variables via the model-agnostic Hilbert-Schmidt independence criterion and embedding the estimates into the formerly proposed PSI-framework using asymptotic normality. However, this approach still requires the user to decide how many features to select. Our method chooses features in a data-driven manner and correctly accounts for the selection process and is thus ideal for situations with limited knowledge about the structure of the data.

Extending the theoretical framework to allow for a sequential application of HSIC-Lasso with different values of λ, similar to Tibshirani et al.'s least angle regression algorithm (2016), is an interesting and practically relevant step for future research. Moreover, Liu et al. (2018) hint that developing inference targets apart from the partial and HSIC-target can be useful to reduce the length of confidence intervals. Lastly, the incorporation of the choice of λ into the post-selection framework for HSIC-Lasso would allow to analyse the data on only one fold and render the proposed framework fully in line with the PSI philosophy. Loftus (2015) and Markovic et al. (2017) took first steps in this direction; albeit, the application to HSIC-Lasso is still an open issue. From an algorithmic point of view, establishing heuristics for a good choice of hyper-parameters, such as the size of the estimators or the split ratio, can as well be object of future research.

Acknowledgments

Tobias Freidling thanks the Max Weber Program and Erasmus+ for supporting his stay at Kyoto University with a scholarship and Mathias Drton for an insightful discussion on Theorem 9. Makoto Yamada was supported by MEXT KAKENHI 20H04243 and partly supported by MEXT KAKENHI 21H04874. Benjamin Poignard was supported by the Japanese Society for the Promotion of Science (Start-Support Kakenhi 19K23193).

References

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Chen, Y., Wiesel, A., Eldar, Y. C., and Hero, A. O. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Cox, D. R. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62:441–444, 1975.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. (With discussion). *The Annals of Statistics*, 32(2):407–499, 2004.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory. 16th international conference, ALT 2005, Singapore, October 8–11, 2005. Proceedings.*, pp. 63–77. Berlin: Springer, 2005.

Gunduz, N. and Fokoue, E. UCI machine learning repository, 2013.

Higham, N. J. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103 – 118, 1988.

Hocking, R. R. A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32(1):1–49, 1976.

Hyun, S., G'sell, M., and Tibshirani, R. J. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. Exact post-selection inference, with application to the Lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Leeb, H. and Pötscher, B. M. Model selection and inference: facts and fiction. *Econometric Theory*, 21(1):21–59, 2005.

Leeb, H. and Pötscher, B. M. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591, 2006.

Lim, J. N., Yamada, M., Jitkrittum, W., Terada, Y., Matsui, S., and Shimodaira, H. More Powerful Selective Kernel Tests for Feature Selection. volume 108 of *Proceedings of Machine Learning Research*, pp. 820–830. PMLR, 2020.

Liu, K., Markovic, J., and Tibshirani, R. More powerful post-selection inference, with application to the Lasso. *arXiv preprint*, 1801.09037, 2018.

Loftus, J. R. Selective inference after cross-validation. *arXiv preprint*, 1511.08866, 2015.

Markovic, J., Xia, L., and Taylor, J. Unifying approach to selective inference with applications to cross-validation. *arXiv preprint*, 1703.06559, 2017.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2018.

Shimodaira, H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *The Annals of Statistics*, 32(6):2616–2641, 2004.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Algorithmic learning theory. 18th international conference, ALT 2007, Sendai, Japan, October 1–4, 2007. Proceedings*, pp. 13–31. Berlin: Springer, 2007.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *Journal of Machine Learning Research (JMLR)*, 13:1393–1434, 2012.

Stone, M. Cross-validatory choice and assessment of statistical predictions. Discussion. *Journal of the Royal Statistical Society. Series B*, 36:111–147, 1974.

Taylor, J. and Tibshirani, R. Post-selection inference for ℓ_1-penalized likelihood models. *The Canadian Journal of Statistics*, 46(1):41–61, 2018.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174 (3):716 – 729.e27, 2018.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P. L., Rozenblatt-Rosen, O., Lane, A. A., Haniffa, M., Regev, A., and Hacohen, N. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), 2017.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. High-dimensional feature selection by feature-wise kernelized Lasso. *Neural Computation*, 26 (1):185–207, 2014.

Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. Post Selection Inference with Kernels. volume 84 of *Proceedings of Machine Learning Research*, pp. 152–160. PMLR, 2018.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.