Post Selection Inference with Kernels

Makoto Yamada[1,2,4] **Yuta Umezu**[3] **Kenji Fukumizu**[1,4] **Ichiro Takeuchi**[1,3]
[1]RIKEN AIP, [2]JST PRESTO, [3]Nagoya Institute of Technology, [4] Institute of Statistical Mathematics

Abstract

Finding a set of *statistically significant* features from complex data (e.g., nonlinear and/or multi-dimensional output data) is important for scientific discovery and has a number of practical applications including biomarker discovery. In this paper, we propose a kernel-based post-selection inference (PSI) algorithm that can find a set of *statistically significant* features from non-linearly related data. Specifically, our PSI algorithm is based on independence measures, and we call it the Hilbert-Schmidt Independence Criterion (HSIC)-based PSI algorithm (`hsicInf`). The novelty of `hsicInf` is that it can handle non-linearity and/or multi-variate/multi-class outputs through kernels. Through synthetic experiments, we show that `hsicInf` can find a set of *statistically significant* features for both regression and classification problems. We applied `hsicInf` to real-world datasets and show that it can successfully identify important features.

1 Introduction

Finding a set of features in high-dimensional data is important with many real-world applications such as biomarker discovery (Xing et al., 2001), document categorization (Forman, 2008), and prosthesis control (Shenoy et al., 2008). In particular, finding a set of *statistically significant* features is crucial for scientific discovery, and linear methods including the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) are extensively used. However, LASSO is focused on finding a set of *linearly* related features. Thus, if an input and output pair has a *non-linear* relationship, it is difficult to select a set of important features.

To select non-linearly related features, a feature-screening approach, which is based on ranking features with respect to the association score between each feature and its output, is widely used (Fan & Lv, 2008). Typically, the correlation coefficient (linear) and mutual information (non-linear) are used as an association measure (Cover & Thomas, 2006). Recently, kernel-based independence measures such as the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) and its normalized variant (NOCCO) were proposed and have started being used as surrogates of mutual information (Song et al., 2012; Balasubramanian et al., 2013; Fukumizu et al., 2008). The key advantage of kernel-based approaches is that they can deal with non-linearity and/or multi-variate and multi-class data through kernels. Besides association approaches, kernel- and sparse-regularization-based approaches including the sparse additive model (SpAM) (Ravikumar et al., 2009) and HSIC LASSO (Yamada et al., 2014) are extensively used in feature-selection communities. While these kernel-based approaches select a set of nonlinear related features, it is not clear whether the selected features are *statistically significant.* We may consider a naive two-step approach that involves first selecting features then testing the selected features without any adjustment. However, since the *selection event* needs to be taken into account for statistical inference, a naive two-step approach cannot control the desired false positive rates (FPRs).

The problem of testing the significance of the selected features is known as *selective inference* (Taylor & Tibshirani, 2015; Hastie et al., 2015). A basic approach to selective inference is data splitting. The key idea is to divide a training dataset into two disjoint sets: one of the sets is used for feature selection and the other for statistical inference. Since the selection event and statistical inference are independent due to splitting, we can avoid the issue of uncontrolled FPRs in selecting statistically important features. A drawback is the degraded detection power since the latter inference is based on only the half of the data.

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

Recently, selective inference algorithms called post-selection inference (PSI) have been proposed (Hastie et al., 2015; Lockhart et al., 2014; Lee et al., 2016). Since PSI uses an entire dataset for both feature selection and statistical inference, PSI algorithms tend to have higher detection power than the data-splitting algorithms. However, current PSI algorithms are limited to *linear* approaches built upon LASSO or other similar linear feature-selection methods. Since real-world datasets often have a non-linear relationship, the current *linear* approaches may fail to find a set of important features; this is a critical problem in practice. Moreover, current PSI algorithms are only applicable to the case of uni-variate output, which significantly limit applications.

In this paper, we propose a kernel-based PSI algorithm that can find *statistically significant* features from non-linear and/or multi-dimensional output. Specifically, our PSI algorithm is based on the independence measure HSIC, and we call it `hsicInf`. A clear advantage of `hsicInf` over current algorithms is that it can easily handle non-linearity and multi-dimensional output through kernels. Namely, it can be used for a wider range of applications including multi-class classification and multi-variate regression. Through synthetic and real-world experiments, we show that the proposed algorithm finds a set of *statistically significant* features for both regression and classification problems.

2 Proposed algorithm

In this section, we give details of our PSI algorithm with kernels.

2.1 Problem Formulation

Let an input vector be denoted as $\boldsymbol{x} = [x^{(1)}, \ldots, x^{(d)}]^\top \in \mathbb{R}^d$ and the corresponding target vector as $\boldsymbol{y} \in \mathbb{R}^{d_y}$. i.i.d. samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ have been drawn from a joint probability density $p(\boldsymbol{x}, \boldsymbol{y})$. The final goal of this study was to first screen $k < d$ features of input vector $\boldsymbol{x}$ then test whether the selected features are of *statistically significant* association to its output $\boldsymbol{y}$.

2.2 Marginal screening and post-selection inference with independence measure

We use an estimate $\widehat{I}(X_m, Y)$ of the independence measure $I(X_m, Y)$, which measures the discrepancy from the independence between the m-th random variable X_m and its output variable Y. The vector of the estimates is denoted as $\boldsymbol{z} = [\widehat{I}(X_1, Y), \ldots, \widehat{I}(X_d, Y)]^\top$. In this subsection, we assume that $\boldsymbol{z}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$:

$$\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Under this normality assumption, we exploit the post-selection inference approach developed recently by Lee et al. (2016) (see Theorem 1). In this approach, the selection event is expressed in a form of linear inequality on the input variables: this includes LASSO as a selection procedure.

Theorem 1 *(Lee et al., 2016). Consider a stochastic data-generating process $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If a feature-selection event is characterized by $\boldsymbol{A}\boldsymbol{z} \le \boldsymbol{b}$ for a matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$ that do not depend on $\boldsymbol{z}$, then, for any fixed vector $\boldsymbol{\eta} \in \mathbb{R}^d$,*

$$F_{\boldsymbol{\eta}^\top \boldsymbol{\mu}, \boldsymbol{\eta}^\top \boldsymbol{\Sigma} \boldsymbol{\eta}}^{[V^-(\boldsymbol{A},\boldsymbol{b}), V^+(\boldsymbol{A},\boldsymbol{b})]}(\boldsymbol{\eta}^\top \boldsymbol{z}) \quad | \quad \boldsymbol{A}\boldsymbol{z} \le \boldsymbol{b} \quad \sim \quad \mathrm{Unif}(0,1),$$

where $F_{t,u}^{[v,w]}(\cdot)$ is the cumulative distribution function of the uni-variate truncated normal distribution with the mean t, variance u, and lower and upper truncation points v and w, respectively. Furthermore, using $\boldsymbol{c} := \frac{\boldsymbol{\Sigma}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top \boldsymbol{\Sigma} \boldsymbol{\eta}}$, the lower and upper truncation points are given as

$$V^-(\boldsymbol{A}, \boldsymbol{b}) := \max_{j:(\boldsymbol{A}\boldsymbol{c})_j < 0} \left\{ \frac{b_j - (\boldsymbol{A}\boldsymbol{z})_j}{(\boldsymbol{A}\boldsymbol{c})_j} \right\} + \boldsymbol{\eta}^\top \boldsymbol{z}, \quad (1)$$

$$V^+(\boldsymbol{A}, \boldsymbol{b}) := \min_{j:(\boldsymbol{A}\boldsymbol{c})_j > 0} \left\{ \frac{b_j - (\boldsymbol{A}\boldsymbol{z})_j}{(\boldsymbol{A}\boldsymbol{c})_j} \right\} + \boldsymbol{\eta}^\top \boldsymbol{z}. \quad (2)$$

Note that a random variable W has a cumulative distribution function F if and only if $F(W)$ follows $\mathrm{Unif}(0,1)$. Thus, Theorem 1 roughly states that, given the selection event, the variable $\boldsymbol{\eta}^\top \boldsymbol{z}$ distributes as a truncated normal.

Our aim was to develop a PSI algorithm based on the independence measure vector $\boldsymbol{z}$. In applying Theorem 1, we need to confirm that the problem of selecting top k features in the decreasing order of $\widehat{I}(X_\cdot, Y)$ can be represented as a linear selection event in the form of $\boldsymbol{A}\boldsymbol{z} \le \boldsymbol{b}$. Let $\mathcal{S}$ and $\bar{\mathcal{S}}$ denote the index set of the selected k features and the unselected $\bar{k} = d - k$ features, respectively. The fact that k features in $\mathcal{S}$ are selected and $\bar{k}$ features in $\bar{\mathcal{S}}$ are not selected is rephrased by

$$\widehat{I}(X_m, Y) \ge \widehat{I}(X_\ell, Y), \quad \text{for all } (m, \ell) \in \mathcal{S} \times \bar{\mathcal{S}}. \quad (3)$$

This is in fact a set of linear inequalities with respect to $\boldsymbol{z}$, consisting of $k\bar{k}$ constraints in total. The truncation points in Theorem 1 can be also derived, and the truncated normal distribution in Theorem 1 for the marginal screening with $\boldsymbol{z}$ can be stated as follows.

Theorem 2 *Let $\theta \in [k\bar{k}]$ be the index of the first $k\bar{k}$ affine constraints in Eq. (3) and $C := \{1, \ldots, k\bar{k}\}$. Moreover, for notational simplicity, assume that first k features are selected and remaining $\bar{k} = d - k$ features are unselected. Then, the marginal screening event in Eq. (3) is written as*

$$(\boldsymbol{A}\boldsymbol{z})_\theta = \widehat{I}(X_{\ell(\theta)}, Y) - \widehat{I}(X_{m(\theta)}, Y) \leq 0$$
$$m(\theta) := \lceil \theta / \bar{k} \rceil, \ell(\theta) := k + (\theta \bmod \bar{k})\ \theta \in C.$$

The lower and upper truncation points for the m-th feature are written as

$$V^{-}(\boldsymbol{A}, \boldsymbol{0})$$
$$:= \max_{\theta \in \mathcal{D}} \left\{ \frac{[\boldsymbol{\Sigma}]_{m,m}\Big(\widehat{I}(X_{m(\theta)}, Y) - \widehat{I}(X_{\ell(\theta)}, Y)\Big)}{[\boldsymbol{\Sigma}]_{\ell(\theta),m} - [\boldsymbol{\Sigma}]_{m(\theta),m}} \right\} + \widehat{I}(X_m, Y),$$
$$V^{+}(\boldsymbol{A}, \boldsymbol{0})$$
$$:= \min_{\theta \in \bar{\mathcal{D}}} \left\{ \frac{[\boldsymbol{\Sigma}]_{m,m}\Big(\widehat{I}(X_{m(\theta)}, Y) - \widehat{I}(X_{\ell(\theta)}, Y)\Big)}{[\boldsymbol{\Sigma}]_{\ell(\theta),m} - [\boldsymbol{\Sigma}]_{m(\theta),m}} \right\} + \widehat{I}(X_m, Y),$$

where

$$\mathcal{D} = \{\theta \ \mid \ [\boldsymbol{\Sigma}]_{\ell(\theta),m} < [\boldsymbol{\Sigma}]_{m(\theta),m}\},$$
$$\bar{\mathcal{D}} = \{\theta \ \mid \ [\boldsymbol{\Sigma}]_{\ell(\theta),m} > [\boldsymbol{\Sigma}]_{m(\theta),m}\}.$$

Proof: For deriving the lower and upper bounds, we simply need to plug $\boldsymbol{\eta} = \boldsymbol{e}_m$ (the unit vector whose m-th element is one, and zero otherwise), $\boldsymbol{b} = \boldsymbol{0}$, $-(\boldsymbol{A}\boldsymbol{z})_\theta = \widehat{I}(X_{m(\theta)}, Y) - \widehat{I}(X_{\ell(\theta)}, Y)$, and $(\boldsymbol{A}\boldsymbol{c})_\theta = \left[\frac{[\boldsymbol{\Sigma}]_{\ell(\theta),m} - [\boldsymbol{\Sigma}]_{m(\theta),m}}{[\boldsymbol{\Sigma}]_{m,m}}\right]$ into Eqs. (1) and (2). □

2.3 HSIC-based Post-Selection Inference

Our PSI algorithm is based on the HSIC, which is a previously proposed independence criterion Gretton et al. (2005) and has been used successfully in various problems such as feature selection (Song et al., 2012) and causal inference (Mooij et al., 2009). We aimed to use HSIC for I, as stated in the previous subsection. As detailed below, the most standard estimator of the HSIC does not satisfy asymptotic normality required in the development, as mentioned in Section 2.2; thus, we need a more recent version of the HSIC.

Hilbert-Schmidt Independence Criterion: The HSIC is defined by the Hilbert-Schmidt norm of the covariance operator and known to have an explicit integral form (Gretton et al., 2005) as:

$$\begin{aligned}\mathrm{HSIC}(X, Y) = &\ \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}',\boldsymbol{y},\boldsymbol{y}'}[K(\boldsymbol{x}, \boldsymbol{x}')L(\boldsymbol{y}, \boldsymbol{y}')] \\ &+ \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[K(\boldsymbol{x}, \boldsymbol{x}')]\mathbb{E}_{\boldsymbol{y},\boldsymbol{y}'}[L(\boldsymbol{y}, \boldsymbol{y}')] \qquad (4) \\ &- 2\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{x}'}[K(\boldsymbol{x}, \boldsymbol{x}')]\mathbb{E}_{\boldsymbol{y}'}[L(\boldsymbol{y}, \boldsymbol{y}')]\right],\end{aligned}$$

where $K(\boldsymbol{x}, \boldsymbol{x}')$ and $L(\boldsymbol{y}, \boldsymbol{y}')$ are positive definite kernels, and $\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}',\boldsymbol{y},\boldsymbol{y}'}$ denotes the expectation over independent pairs $(\boldsymbol{x}, \boldsymbol{y})$ and $(\boldsymbol{x}', \boldsymbol{y}')$ drawn from $p(\boldsymbol{x}, \boldsymbol{y})$. With the use of characteristic kernels (Fukumizu et al., 2004; Sriperumbudur et al., 2011), the HSIC serves as an independence criterion: it takes zero if X and Y are independent and takes positive values otherwise. We can thus expect to select important features by ranking HSIC scores $\{\mathrm{HSIC}(X_m, Y)\}_{m=1}^{d}$ in descending order, where X_m is the random variable of the m-th feature, as done in a previous study (Song et al., 2012)

Based on the expression 4, the standard estimators of HSIC are given by a V-statistic or U-statistic (Gretton et al., 2005; Song et al., 2012). Note, however, that those estimators are degenerate (i.e., under the independence between X and Y, HSIC has the asymptotic order $1/n$, not $1/\sqrt{n}$, as in the usual asymptotic theory), and do not satisfy the asymptotic normality. We therefore propose to use the block HSIC estimator, which has been introduced by Zhang et al. (2017).

Empirical Block HSIC: Assume that n samples can be disjointly divided into $\frac{n}{B}$ blocks, where B is the number of samples in each block. The samples are accordingly denoted as $\{\{(\boldsymbol{x}_i^{(b)}, \boldsymbol{y}_i^{(b)})\}_{i=1}^{B}\}_{b=1}^{n/B}$.

An empirical estimate of the *unbiased* block HSIC (Zhang et al., 2017) is given by

$$\widehat{\mathrm{HSIC}}(X, Y) = \frac{B}{n}\sum_{b=1}^{n/B} \widehat{\eta}_b,$$
$$\begin{aligned}\widehat{\eta}_b = &\ \frac{1}{B(B-3)}[\mathrm{tr}(\bar{\boldsymbol{K}}^{(b)}\bar{\boldsymbol{L}}^{(b)}) + \frac{\boldsymbol{1}_B^\top \bar{\boldsymbol{K}}^{(b)} \boldsymbol{1}_B \boldsymbol{1}_B^\top \bar{\boldsymbol{L}}^{(b)} \boldsymbol{1}_B}{(B-1)(B-2)} \\ &- \frac{2}{B-2}\boldsymbol{1}_B^\top \bar{\boldsymbol{K}}^{(b)}\bar{\boldsymbol{L}}^{(b)}\boldsymbol{1}_B],\end{aligned}$$

where $\boldsymbol{K}^{(b)} \in \mathbb{R}^{B \times B}$ is the input Gram matrix, $\boldsymbol{L}^{(b)} \in \mathbb{R}^{B \times B}$ is the output Gram matrix, $[\bar{\boldsymbol{K}}^{(b)}]_{ij} = [\boldsymbol{K}^{(b)}]_{ij} - \delta_{ij}[\boldsymbol{K}^{(b)}]_{ij}$ and $[\bar{\boldsymbol{L}}^{(b)}]_{ij} = [\boldsymbol{L}^{(b)}]_{ij} - \delta_{ij}[\boldsymbol{L}^{(b)}]_{ij}$, δ_{ij} takes 1 when $i = j$, and 0 otherwise, and $\boldsymbol{1}_B \in \mathbb{R}^B$ is the vector whose elements are all one. The $\widehat{\eta}_b$ is known to be an unbiased estimator of the HSIC with the samples in the b-th block. Note that $\widehat{\eta}_b$ $(b = 1, \ldots, n/B)$ are i.i.d. random variables, as they are disjointly computed with a partition of i.i.d. samples.

As discussed in Zhang et al. (2017), from the standard asymptotic theory on i.i.d. variables, the empirical block HSIC score asymptotically follows normal distribution when B is finite and n goes to infinity. We can use the block HSIC for PSI based on Theorem 2 in the asymptotic regime. Note that, to ensure Gaussian assumption, we need to have a relatively large number of samples n with a finite block size B.

Bagging Block HSIC: The block HSIC heavily depends on a partition of $\{\{(\boldsymbol{x}_i^{(b)}, \boldsymbol{y}_i^{(b)})\}_{i=1}^{B}\}_{b=1}^{n/B}$. To mitigate this problem, we propose the Bagging Block HSIC:

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

where $\widehat{\eta}_{\ell,b}$ is computed using the blocked sample $\{(\boldsymbol{x}_i^{(\ell,b)}, \boldsymbol{y}_i^{(\ell,b)})\}_{i=1}^{B}$, $\boldsymbol{x}_i^{(\ell,b)}$ and $\boldsymbol{y}_i^{(\ell,b)}$ are samples of the b-th block with ℓ-th random permutation, and L is the number of permutations. We can easily show that $\mathbb{E}[\widehat{\mathrm{HSIC}}(X,Y)] = \mathrm{HSIC}(X,Y)$.

Variance of the Bagging Block HSIC: We will derive the variance of the bagging block HSIC and relate it with that of the full U-statistics and (unbagged) single block HSIC.

Let $Z_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)$ $(i = 1, 2, \ldots, n)$ be i.i.d. samples. The HSIC can be expressed in the form of U-statistics of 4th degree:

$$U_n = \frac{1}{\binom{n}{4}}\sum_{S\in\mathfrak{S}_{n,4}} h(Z_S),$$

where $h(z_1, z_2, z_3, z_4)$ is the U-statistic kernel corresponding to the HSIC (see Song et al. (2012)), $\mathfrak{S}_{n,k}$ is the set of all k-tuples (in this case $k = 4$) of $\{1, \ldots, n\}$, and Z_S is an abbreviation of $(Z_{i_1}, \ldots, Z_{i_k})$ for $S = (i_1, \ldots, i_k)$.

Consider the block HSIC with block size B. For simplicity, let $M := n/B$ denote the number of blocks for a single block HSIC estimator and assume that n is taken so that M in an integer. The block HSIC is then defined by

$$W_B = \frac{1}{M}\sum_{b=1}^{M} U_B^{(b)}, \tag{5}$$

where $U_B^{(b)}$ is the U-statistics corresponding to the empirical HSIC computed from only the B samples in the b-the block, namely,

$$U_B^{(b)} = \frac{1}{\binom{B}{4}}\sum_{S} h(Z_S),$$

where the sum is taken for all the quadruplets from the b-th block. Note that W_B converges in law to a normal distribution as $n \to \infty$ with B fixed since $U_B^{(b)}$ $(b = 1, \ldots, M)$ are i.i.d. samples.

Recall that the bagging block HSIC with L random permutations is defined by

$$\xi_{L,B} := \frac{1}{L}\sum_{\ell=1}^{L} W_{\ell,B}, \tag{6}$$

where $W_{\ell,B}$ is defined similarly to W_B in Eq. (5), but with a random permutation of $Z_1, \ldots, Z_n$. We generate L independent uniform random permutations of $\{1, \ldots, n\}$ and make copies of W_B. Note that, by the independence of the random permutations, given $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$, $W_{\ell,B}$ and $W_{\ell',B}$ are independent for $\ell \neq \ell'$ but can be dependent unconditionally. The bagging block HSIC is simply the average over these L copies.

We rewrite $\xi_{L,B}$ with the indicator of index. Let $\mathfrak{I}_{\ell,b}$ be the index set of the b-th block in the ℓ-th permutation. For an arbitrary quadruplet $S = (i_1, i_2, i_3, i_4)$, define $\theta_{\ell,b}(S)$ by $\theta_{\ell,b}(S) = \begin{cases} 1 & \text{if } S \in \mathfrak{I}_{\ell,b} \\ 0 & \text{otherwise.}\end{cases}$

Then, we have

$$\xi_{L,B} = \frac{1}{LM\binom{B}{4}}\sum_{\ell=1}^{L}\sum_{b=1}^{M}\sum_{S\in\mathfrak{S}_{n,4}}\theta_{\ell,b}(S)h(Z_S). \tag{7}$$

Since $\mathbb{E}[h(S)] = \mathrm{HSIC}(X,Y)$, it is obvious that

$$\mathbb{E}[\xi_{L,B}] = \mathrm{HSIC}(X,Y);$$

thus, $\xi_{L,B}$ is an unbiased estimator of the HSIC.

It is not difficult to see that, as estimators of the HSIC, the standard U-statistics U_n has less variance than the block HSIC $\xi_{1,B}$, which is regarded as a special type of incomplete U-statistic. The next proposition asserts that the variance of $\xi_{L,B}$, under the assumption of independence $X \perp\!\!\!\perp Y$, interpolates the variances of these two estimators.

Proposition 3 *Assume X and Y are independent. Then, we have*

$$\mathrm{Var}[\xi_{L,B}] = \Big(1 - \frac{1}{L}\Big)\mathrm{Var}[U_n] + \frac{1}{L}\mathrm{Var}[\xi_{1,B}].$$

Proof: See the supplementary material.

Choice of Kernel: For regression problems, we use the Gaussian kernel or the Laplacian kernel for both input and output as $\boldsymbol{K}^{(b)} \in \mathbb{R}^{B\times B}$ and $\boldsymbol{L}^{(b)} \in \mathbb{R}^{B\times B}$:

$$[\boldsymbol{K}^{(b)}]_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i^{(b)} - \boldsymbol{x}_j^{(b)}\|_2^2}{2\tau_x^2}\right), [\boldsymbol{L}^{(b)}]_{ij} = \exp\left(-\frac{\|\boldsymbol{y}_i^{(b)} - \boldsymbol{y}_j^{(b)}\|_2^2}{2\tau_y^2}\right),$$

$$[\boldsymbol{K}^{(b)}]_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i^{(b)} - \boldsymbol{x}_j^{(b)}\|_2}{\tau_x}\right), [\boldsymbol{L}^{(b)}]_{ij} = \exp\left(-\frac{\|\boldsymbol{y}_i^{(b)} - \boldsymbol{y}_j^{(b)}\|_2}{\tau_y}\right),$$

where $\tau_x > 0$ and $\tau_y > 0$ are kernel parameters.

For K-class classification problems, we use the delta kernel (Song et al., 2012).

Mean and covariance matrix estimation: Suppose that the mean and variance of the within-block estimator $\widehat{\eta}_b$ are $\widetilde{\mu}$ and $\widetilde{\sigma}$, respectively. Then, the vector of empirical HSICs $[\widehat{\text{HSIC}}(X_1, Y), \ldots, \widehat{\text{HSIC}}(X_d, Y)]^\top \in \mathbb{R}^d$ converges in distribution to a multi-variate normal by the central limit theorem, where the mean and covariance matrices are given by $\widetilde{\mu}$ and $B/n\widetilde{\sigma}$, respectively. In estimating $[\widetilde{\boldsymbol{\Sigma}}]_{m,m'}$, the standard covariance estimator is applied to the (n/B) within-block estimators for m and m'. Note that, when n/B is too small, we can use a high-dimensional covariance-estimation algorithm such as POET (Fan et al., 2013).

Sure Independence Screening: For a high-dimensional setting, since the `hsicInf` algorithm depends on the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ and the covariance matrix needs to be estimated from data samples, detection performance can be degraded when d is large. To address this issue, we incorporate the sure independence screening (SIS) (Fan & Lv, 2008) into `hsicInf`. More specifically, we use the HSIC variant of the SIS algorithm (Balasubramanian et al., 2013).

The key property of SIS is that the selected r features ($k \leq r \ll d$) by marginal screening include *true* features with high probability (Fan & Lv, 2008; Balasubramanian et al., 2013). This indicates that, once we select r features with an HSIC-based SIS, we can ignore the non-selected features. That is, under SIS setup, we need to run hsicInf from the $r \ll d$ features; we only need to estimate $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, which can be easily estimated.

Post-Selection Inference: As a post selection inference, we consider the following hypothesis tests:

- $H_{0,m}$: $\text{HSIC}(X_m, Y) = 0 \mid \mathcal{S}$ was selected,
- $H_{1,m}$: $\text{HSIC}(X_m, Y) \ \neq 0 \mid \mathcal{S}$ was selected.

Then, the p-value of the m-th feature is estimated using Theorem 2.

3 Related Work

In this section, we briefly review related work. It has long been recognized that *selection bias* must be corrected for statistical inference after feature selection. One of the most basic approaches to dealing with selection bias is data splitting, in which the dataset is divided into two disjoint sets: one is used for feature selection and the other for statistical inference. Since the inference phase is made independently of feature selection, we do not have to be concerned about the selection bias. An obvious drawback of data splitting is that the powers become lower both in feature selection and inference phases. Since only half the data are used in each of the two procedures, the risk of failing to select truly important features would increase, and the power of statistical inference (i.e., the probability of true positive finding) would decrease. In addition, different features might be selected if the dataset is divided differently. It is important to note that data splitting is also regarded as selective inference because the inference is made for the selected features, and the other unselected features are ignored.

In statistics, *simultaneous inference* has been studied traditionally for selection bias correction, where all possible subsets of features are considered. Let $\widehat{\mathcal{S}}$ represent the set of selected features and $T_j(\widehat{\mathcal{S}})$ be a test statistic for the j^{th} feature, which is within $\widehat{\mathcal{S}}$. In simultaneous inference, critical points l and u at level α are determined to satisfy

$$P(T_j(\widehat{\mathcal{S}}) \notin [l, u] \text{ for any subset } \widehat{\mathcal{S}} \text{ of the features}) \leq \alpha.$$

This probability is also written as

$$\begin{aligned} &P(T_j(\widehat{\mathcal{S}}) \notin [l, u] \text{ for any subset } \widehat{\mathcal{S}} \text{ of the features}) \\ &= \sum_{\mathcal{S}} P(T_j(\widehat{\mathcal{S}}) \notin [l, u] \mid \widehat{\mathcal{S}} = \mathcal{S}) P(\widehat{\mathcal{S}} = \mathcal{S}), \end{aligned}$$

where the summation of the right-hand side runs over all possible subsets of features. Unfortunately, unless the number of original features is fairly small, it is computationally challenging to consider all possible subsets of features $\widehat{\mathcal{S}}$ (Berk et al., 2013).

In selective inference, we only consider the case in which a certain $\mathcal{S}$ is selected, and we determine the critical points l and u so that *selective type I error* is controlled, i.e.,

$$P(T_j(\widehat{\mathcal{S}}) \notin [l, u] \mid \widehat{\mathcal{S}} = \mathcal{S}) \leq \alpha. \tag{8}$$

The selective inference framework in the form of (8) has been increasingly popular after the seminal work by Lee et al. (2016), in which the authors studied selective inference after feature selection with LASSO (Tibshirani, 1996). Their novel finding is that, in linear regression models with Gaussian noise, if the selection event can be represented by a set of linear inequalities with respect to the response variables (as in LASSO case), then any linear combination of the responses conditioned on the selection event is distributed according to a truncated normal distribution, as stated in Theorem 1. This result is known as a polyhedral lemma and very useful for deriving a null distribution of a test statistic in the context of selective inference. Following this work, the selective inference framework has been studied for several problems where the assumptions of polyhedral lemma are satisfied (see, e.g., (Lee & Taylor, 2014)).

The polyhedral lemma in Lee et al. (2016), however, can be used only when the responses are normally distributed. It is thus difficult to generalize the selective inference framework to other important problems such as classification and multi-task learning. To the best of our knowledge, there have only been a few attempts of selective inference in those generalized settings (Taylor & Tibshirani, 2016). The idea in those studies is to use asymptotic theory, but the underlying assumption used was somewhat restrictive: they required that at least one truncation point is bounded in probability tending to 1. This may not be natural because the truncation points are not bounded in the case of classical inference without feature selection.

The proposed algorithm hsicinf avoids such a restriction because we only use an asymptotic normality of $\widehat{\mathrm{HSIC}}(X_j, Y)$. By virtue of kernel methods, hsicinf enables us to apply selective inference for a variety of response types including classification and multivariate response.

4 Experiment

In this section, we discuss the experiments we conducted to determine the effectiveness of hsicinf in regression and classification problems.

4.1 Setup

We compared the performance of hsicinf with that of the *linear* PSI algorithm (Lee et al., 2016; Efron et al., 2004). Specifically, we used the larInf function in the R package selectiveInference. We additionally compared hsicinf with the data-splitting algorithm split with the block HSIC[1]. For both the proposed and current algorithms, we set the number of selected features to $k = 10$ and the significance level to $\alpha = 0.05$. For the HSIC-based approaches, we used the block parameter $B = \{5, 10\}$. Note that, the bagging block HSIC and SIS have not been theoretically guaranteed yet; thus, we mainly validated the original block HSIC + Theorem 1. Then, we empirically evaluated the bagging block HSIC and SIS ($r = 20$) for high-dimensional data and showed its effectiveness for selective inference. Giving the theoretical guarantee of asymptotic normality for the bagging block HSIC and SIS is important future work.

In the PSI frameworks, the covariance matrix $\mathbf{\Sigma}$ is assumed to be known. However, since the true covariance matrix is not available in practice, we need to estimate the covariance matrix from data. To this end, for hsicInf, we divided the samples into *two* disjoint sets; we used $\frac{n}{3}$ samples for estimating $\mathbf{\Sigma}$ and the rest of the $\frac{2n}{3}$ samples for selecting features and computing the HSIC. In this study, we used the POET algorithm for the covariance matrix estimation (Fan et al., 2013). For split, we divided the samples into *three* disjoint sets with sample size $\frac{n}{3}$. Then, we used each one for estimating $\mathbf{\Sigma}$, selecting and testing features, and computing the HSIC, respectively. Since the block HSIC is asymptotically normal and we need to test each feature independently, the p-value of the k-th feature is given by the 1 - cumulative distribution of $N(0, s_k^2)$, where s_k^2 is estimated from data.

In the regression setup, we used the Gaussian/Laplace kernel in which the kernel parameters are experimentally set to $(\tau_x, \tau_y) = (1.0, 1.0)$ for uni-variate setup and $(\tau_x, \tau_y) = (1.0, \mathrm{med}(\{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2\}_{i,j=1}^n))$ for multivariate output, respectively. Each feature was normalized to have mean zero and standard deviation 1. In the classification setup, we used the Gaussian/Laplace kernel for input and the delta kernel for output. We reported the true positive rate (TPR) $\frac{k'}{k^*}$, where k' is the number of truly relevant features that are reported to be positive, while k^* is the number of truly relevant features. We further computed the FPR $\frac{k''}{k}$, where k'' is the number of truly irrelevant features that are falsely reported to be positive. For synthetic datasets, we ran experiments 500 times with different random seeds and reported the average TPR and average FPR.

4.2 Synthetic Data

Uni-variate Regression: For this experiment, we first generated the input matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$, where $\boldsymbol{x} \sim N(\mathbf{0}, \bar{\mathbf{\Sigma}})$, $[\bar{\mathbf{\Sigma}}]_{ij} = 0.95\delta_{ij} + 0.05, i, j \in \{1, 2, 3, 4, 5\}, [\bar{\mathbf{\Sigma}}]_{ii} = \delta_{ij}, i, j \in \{6, \ldots, d\}$, $\delta_{ij} = 1$ if $i = j$, and 0 otherwise, $d = \{20, 500\}$, and $n = \{300, 600, \ldots, 3000\}$.

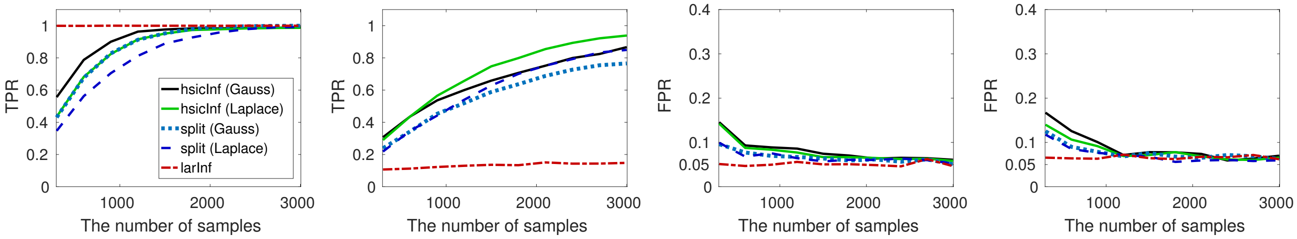
We generated the corresponding output variable as

- **Linear:** $Y = \sum_{i=1}^{5} X_i + 0.1E$,
- **Non-linear:** $Y = X_1 \exp(X_2) X_3 \exp(X_4) X_5 + 0.1E$,

where $E \sim N(0, 1)$ is an independent random noise.

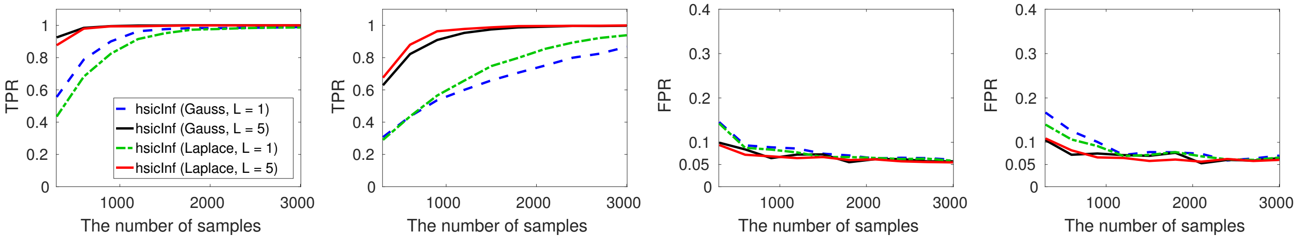
Figures 1 (a)-(b) show the TPRs of all the algorithms ($d = 20$). As we expected, hsicInf had higher TPRs compared to split since hsicinf can use a larger number of samples than split for selecting features. Figures 1 (c)-(d) show the FPRs of all the algorithms ($d = 20$). All the HSIC-based algorithms had larger FPRs than the significance level when the number of samples were small. This may be due to the violation

[1] To the best of our knowledge, the combination of the block HSIC and data splitting has not been proposed, and that combination is also our contribution.



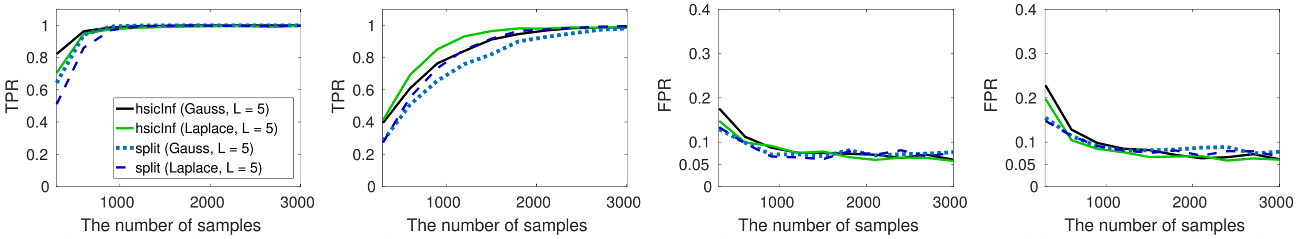
(a) Linear (TPR) (b) Non-linear (TPR) (c) Linear (FPR) (d) Non-linear (FPR)

Figure 1: Results for uni-variate regression setups ($d = 20$). We used $B = 10$ and $L = 1$ for HSIC-based approaches. (a)(b): TPRs. (c)(d): FPRs.



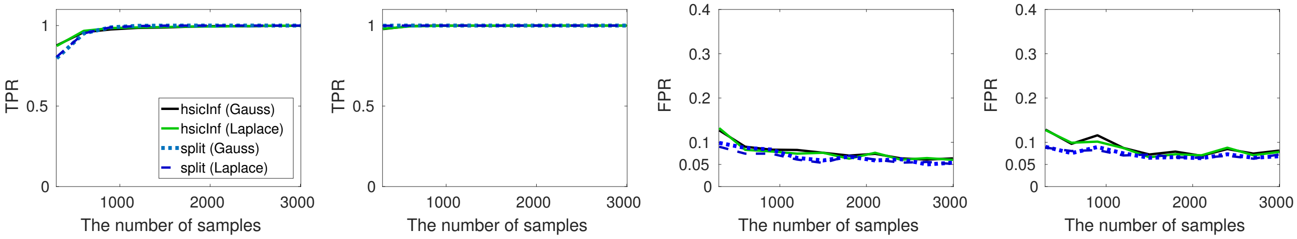
(a) Linear (TPR) (b) Non-linear (TPR) (c) Linear (FPR) (d) Non-linear (FPR)

Figure 2: Comparison between block HSIC ($L = 1$) and bagging block HSIC ($L = 5$) for low-dimensional data $d = 20$. We used $B = 10$. (a)(b): TPRs. (c)(d): FPRs.



(a) Linear (TPR) (b) Non-linear (TPR) (c) Linear (FPR) (d) Non-linear (FPR)

Figure 3: Comparison between hsicInfs and split for high-dimensional setting ($d = 500$). For hsicInfs, we used SIS algorithm in addition to bagging HSIC. We used $B = 10$ and $L = 5$ for all algorithms. (a)(b): TPRs. (c)(d): FPRs.



(a) Multi-variate (TPR) (b) Multi-class (TPR) (c) Multi-variate (FPR) (d) Multi-class (FPR)

Figure 4: Results for multi-variate regression and multi-class classification datasets for low-dimensional setting ($d = 20$). (a)(b): TPRs. (c)(d): FPRs.

in the Gaussian assumption or insufficient accuracy of covariance estimation. Our `hsicInf` with a small number of samples is important future work.

The linear algorithm `larInf` can select features for only linear setups and fails for non-linear counterparts. In contrast, `hsicinf`can successfully detect statistically significant features for all setups. See the supplementary material for comparisons of block size B.

Figure 2 shows the comparison between the block HSIC ($L = 1$) and bagging block HSIC ($L = 5$). The bagging block HSIC had high TPRs with low FPRs. Thus, it would be highly useful.

Figure 3 shows the TPRs and FPRs for high-

Table 1: p-values computed using `hsicInf` from Turkey Student Evaluation dataset

Feature description	p-value
Q28: The Instructor treated all students in a right and objective manner.	< 0.001
Q17: The Instructor arrived on time for classes.	0.033
Q13: The Instructor's knowledge was relevant and up to date.	0.018
Q22: The Instructor was open and respectful of the views of students about the course.	0.042
Q21: The Instructor demonstrated a positive approach to students.	0.033
Q18: The Instructor has a smooth and easy to follow delivery/speech.	0.186
Q23: The Instructor encouraged participation in the course.	0.037
Q26: The Instructor's evaluation system effectively measured the course objectives.	0.176
Q2: The course aims and objectives were clearly stated at the beginning of the period.	0.452
Q20: The Instructor explained the course and was eager to be helpful to students.	0.463

dimensional $d = 500$ cases. By combining the bagging block HSIC and SIS algorithm with `hsicInf`, `hsicInf` could outperform `split` in terms of the TPR. Note that, to compare fairly, we also used the bagging block HSIC with split.

Multi-variate Regression: We used the zero-mean multi-variate Gaussian input matrix with $[\bar{\boldsymbol{\Sigma}}]_{ij} = 0.95\delta_{ij} + 0.05, i, j \in \{1, 2, 3, 4\}, [\bar{\boldsymbol{\Sigma}}]_{ii} = \delta_{ij}, i, j \in \{5, \ldots, 20\}$. For the output variable, we generated the three dimensional output variables as $Y_1 = X_1 + 2X_2 + 0.1E$, $Y_2 = 2X_1 + X_2^2 + 0.1E$, $Y_3 = X_3 \exp(2X_4) + 0.1E$.. Note that, since current PSI algorithms cannot be used for multi-variate outputs, we only reported on the HSIC-based algorithms.

Figures 4(a)(c) show that `hsicInf` can successfully select statistically significant features.

Multi-class classification: In this experiment, we evaluated the algorithm/`hsicInf` using a three-class classification dataset (see the supplementary material for details). Again, there is no multi-class PSI algorithm; thus, we simply report on the performance of the HSIC-based algorithms. Figures 4(b)(d) show that `hsicInf` can perfectly detect the important features.

4.3 Real-world data

Turkey Student Dataset: The Turkey dataset consists of 5820 samples with 28 features, where variables take *integers* and *are non-Gaussian.* We used the "Level of difficulty of the course as perceived by the student $(\{1, 2, \ldots, 5\})$" as the output variable and selected features that significantly affected the difficulty of the course. In this experiment, we set the number of selected features k to 10 and block size B to 10, and used the Gaussian kernel for output. Note that, since the output variable takes integers and is non-Gaussian, `larInf` cannot be used for these data.

Table 1 lists the selected features from `hsicInf`. The difficulties of class are highly related to the attitude of teachers and teacher's support to students, and this result is reasonable.

Table 2: p-values computed with `hsicInf` from QSAR biodegradation dataset

Symbol	Descriptor type	p-value
SpMax_L	2D-matrix-based	0.009
SpPosA_B (p)	2D-matrix-based	0.004
SM6_B (m)	2D-matrix-based	0.002
SpMax_B (m)	2D-matrix-based	< 0.001
SpMax_A	2D-matrix-based	0.011
HyWi_B (m)	2D-matrix-based	0.018
C-026	atom centered fragments	0.086
SM6_L	2D-matrix-based	0.239
nN	constitutional indices	0.351
F04[C-N]	2D atom pairs	0.346

Quantitative Structure-Activity Relationship (QSAR) biodegradation dataset: This dataset consists of 1055 samples with 41 features, where each feature is a molecule descriptor. The task with this dataset is to classify the samples into two classes: "ready biodegradable" and "not ready biodegradable" (see Mansouri et al. (2013) for details). In this experiment, we set the number of selected features k to 10 and the block size B to 10.

Table 2 lists the selected features from `hsicInf`. The 2D matrix-based descriptors, which were previously reported as important Mansouri et al. (2013), were also selected with `hsicInf` as significant features. Thus, `hsicinf` successfully found important descriptors.

5 Conclusion

We proposed the post-selection inference (PSI) algorithm `hsicInf`. The key advantage of `hsicinf` is that it can select *statistically significant* features from nonlinear and/or multi-variate data and has high detection power. To the best of our knowledge, this is the first work to address a PSI algorithm for both nonlinear and multi-variate regression problems. Through several experiments, we showed that `hsicinf` overcomes the limitations of the state-of-the-art *linear* PSI algorithm and outperforms the splitting algorithm in large sample cases.

Acknowledgement

MY was supported by the JST PRESTO program JPMJPR165A and partly supported by MEXT KAKENHI 16K16114.IT was partially supported by MEXT KAKENHI (17H00758, 16H06538), JST CREST (JPMJCR1302, JPMJCR1502), RIKEN Center for Advanced Intelligence Project, and JST support program for starting up innovation-hub on materials research by information integration initiative.

References

Balasubramanian, K., Sriperumbudur, B., and Lebanon, G. Ultrahigh dimensional feature screening via RKHS embeddings. In *AISTATS*, 2013.

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

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition, 2006.

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

Fan, J. and Lv, J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Fan, J., Liao, Y., and Mincheva, M. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4): 603–680, 2013.

Forman, G. BNS feature scaling: An improved representation over TF-IDF for SVM text classification. In *CIKM*, 2008.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS*, 2008.

Gretton, A., Bousquet, O., Smola, Alex., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 2005.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.

Lee, J. D. and Taylor, J. E. Exact post model selection inference for marginal screening. In *NIPS*, 2014.

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.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R., and Consonni, V. Quantitative structure–activity relationship models for ready biodegradability of chemicals. *Journal of chemical information and modeling*, 53(4):867–878, 2013.

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. Regression by dependence minimization and its application to causal inference in additive noise models. In *ICML*, pp. 745–752, Montreal, Canada, 2009.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Shenoy, P., Miller, K. J., Crawford, B., and Rao, R. N. Online electromyographic control of a robotic prosthesis. *IEEE Transactions on Biomedical Engineering*, 55(3):1128–1135, 2008.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. RG. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

Taylor, J. and Tibshirani, R. Post-selection inference for l1-penalized likelihood models. *arXiv preprint arXiv:1602.07358*, 2016.

Taylor, J. and Tibshirani, R. J. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

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

Xing, E. P., Jordan, M. I., Karp, R. M., et al. Feature selection for high-dimensional genomic microarray data. In *ICML*, 2001.

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.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. Large-scale kernel methods for independence testing. *Statistics and Computing*, pp. 1–18, 2017.