kernelPSI: a Post-Selection Inference Framework for Nonlinear Variable Selection

Lotfi Slim, Clement Chatelain, Chloé-Agathe Azencott, Jean-Philippe Vert

kernelPSI: a Post-Selection Inference Framework for Nonlinear Variable Selection

Lotfi Slim [1 2] **Clément Chatelain** [1] **Chloé-Agathe Azencott** [2 3] **Jean-Philippe Vert** [2 4]

Abstract

Model selection is an essential task for many applications in scientific discovery. The most common approaches rely on univariate linear measures of association between each feature and the outcome. Such classical selection procedures fail to take into account nonlinear effects and interactions between features. Kernel-based selection procedures have been proposed as a solution. However, current strategies for kernel selection fail to measure the significance of a joint model constructed through the combination of the basis kernels. In the present work, we exploit recent advances in post-selection inference to propose a valid statistical test for the association of a joint model of the selected kernels with the outcome. The kernels are selected via a step-wise procedure which we model as a succession of quadratic constraints in the outcome variable.

1. Introduction

Variable selection is an important preliminary step in many data analysis tasks, both to reduce the computational complexity of dealing with high-dimensional data and to discard nuisance variables that may hurt the performance of subsequent regression or classification tasks. Statistical inference about the selected variables, such as testing their association with an outcome of interest, is also relevant for many applications, such as identifying genes associated with a phenotype in genome-wide association studies. If the variables are initially selected using the outcome, then standard statistical tests must be adapted to correct for the fact that the variables tested after selection are likely to exhibit strong association with the outcome, because they were selected for that purpose.

This problem of *post-selection inference* (PSI) can be solved by standard data splitting strategies, where we use different samples for variable selection and statistical inference (Cox, 1975). Splitting data is however not optimal when the total number of samples is limited, and alternative approaches have recently been proposed to perform proper statistical inference after variable selection (Taylor & Tibshirani, 2015). In particular, in the *conditional coverage* setting of Berk et al. (2013), statistical inference is performed conditionally to the selection of the model. For linear models with Gaussian additive noise, Lee et al. (2016); Tibshirani et al. (2016) show that proper statistical inference is possible and computationally efficient in this setting for features selected by lasso, forward stepwise or least angle regression. In these cases it is indeed possible to characterize the distribution of the outcome under a standard null hypothesis model conditionally to the selection of a given set of features. This distribution is a Gaussian distribution truncated to a particular polyhedron. Similar PSI schemes were derived when features are selected not individually but in groups (Loftus & Taylor, 2015; Yang et al., 2016; Reid et al., 2017).

Most PSI approaches have been limited to linear models so far. In many applications, it is however necessary to account for nonlinear effects or interactions, which requires nonlinear feature selection. This requires generalizing PSI techniques beyond linear procedures. Recently, Yamada et al. (2018) took a first step in that direction by proposing a PSI procedure to follow kernel selection, where kernels are used to generalize linear models to the nonlinear setting. However, their approach is limited to a single way of selecting kernels, namely, marginal estimation of the Hilbert-Schmidt Independent Criterion (HSIC) independence measure (Song et al., 2007). In addition, it only allows to derive post-selection statistical guarantees for one specific question, that of the association of a selected kernel with the outcome.

In this work we go one step further and propose a general framework for kernel selection, that leads to valid PSI procedures for a variety of statistical inference questions. Our main contribution is to propose a large family of statistics

[1]Translational Sciences, SANOFI R&D, France [2]MINES ParisTech, PSL Research University, CBIO - Centre for Computational Biology, F-75006 Paris, France [3]Institut Curie, PSL Research University, INSERM, U900, F-75005 Paris, France [4]Google Brain, F-75009 Paris, France. Correspondence to: Lotfi Slim <lotfi.slim@mines-paristech.fr>, Jean-Philippe Vert <jpvert@google.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

that estimate the association between a given kernel and an outcome of interest, that can be formulated as a quadratic function of the outcome. This family includes in particular the HSIC criterion used by Yamada et al. (2018), as well as a generalization to the nonlinear setting (a "kernelization") of the criterion used by Loftus & Taylor (2015); Yang et al. (2016) to select a group of features in the linear setting. When these statistics are used to select a set of kernels, by marginal filtering or by forward or backward stepwise selection, we can characterize the set of outcomes that lead to the selection of a particular subset as a conjunction of quadratic inequalities. This paves the way to various PSI questions by sampling-based procedures.

2. Settings and Notations

Given a data set of n pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where for each $i \in [1, n]$ the data $x_i \in \mathcal{X}$ for some set $\mathcal{X}$ and the outcome $y_i \in \mathbb{R}$, our goal is to understand the relationship between the data and the outcome. We denote by $Y \in \mathbb{R}^n$ the vector of outcomes ($Y_i = y_i$ for $i \in [1, n]$). We further consider a set of S positive definite kernels $\mathcal{K} = \{k_1, \ldots, k_S\}$ defined over $\mathcal{X}$, and denote $K_1, \ldots, K_S$ the corresponding $n \times n$ Gram matrices (i.e., for any $t \in [1, S], i, j \in [1, n], [K_t]_{ij} = k_t(x_i, x_j)$). We refer to the kernels $k \in \mathcal{K}$ as *local* or *basis* kernels. Our goal is to select a subset of S' local kernels $\{k_{i_1}, \cdots, k_{i_{S'}}\} \subset \mathcal{K}$ that are most associated with the outcome Y, and then to measure the significance of their association with Y.

The choice of basis kernels $\mathcal{K}$ allows us to model a wide range of settings for the underlying data. For example, if $\mathcal{X} = \mathbb{R}^d$, then a basis kernel can only depend on a single coordinate, or on a group of coordinates, in which case selecting kernels leads to variable selection (individually or by groups). Another useful scenario is to consider nonlinear kernels with different hyperparameters, such as a Gaussian kernel with different bandwidth, in which case kernel selection leads to hyperparameter selection.

3. Kernel Association Score

Our kernel selection procedure is based on the following general family of association scores between a kernel and the outcome:

Definition 1. *A quadratic kernel association score is a function $s : \mathbb{R}^{n\times n} \times \mathbb{R}^n \to \mathbb{R}$ of the form*

$$s(K, Y) = Y^\top Q(K) Y \,, \tag{1}$$

for some function $Q : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$.

If $s(K, Y)$ is a positive definite quadratic form in Y (i.e., if $Q(K)$ is positive semi-definite), we can rewrite it as:

$$s(K, Y) = \|\widehat{Y}_K\|^2 \,, \tag{2}$$

where $\widehat{Y}_K = H(K)Y$ is called a *prototype* for a "hat" function $H : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ (take for example $H = Q^{1/2}$). We borrow the term "prototype" from Reid et al. (2017), who use it to design statistical tests of linear association between the outcome and a group of features.

One reason to consider quadratic kernel association scores is that they cover and generalize several measures used for kernel or feature selection. Consider for example $H_{\text{proj}}(K) = KK^+$, where K^+ is the Moore-Penrose inverse of K. The score proposed by Loftus & Taylor (2015) for a group of d features encoded as $X_g \in \mathbb{R}^{n\times d}$ is a special case of H_{proj} with $K = X_g X_g^\top$. In this case, the prototype $\widehat{Y}$ is the projection of Y onto the space spanned by the features.

If $K = \sum_{i=1}^r \lambda_i u_i u_i^\top$ is the singular value decomposition of K, with $\lambda_1 \geq \ldots \geq \lambda_r > 0$, H_{proj} can be rewritten as

$$H_{\text{proj}}(K) = \sum_{i=1}^{r} u_i u_i^\top \,. \tag{3}$$

For a general kernel K, which may have large rank r, we propose to consider two regularized versions of Eq. (3) to reduce the impact of small eigenvalues. The first one is the *kernel principal component regression (KPCR) prototype*, where $\widehat{Y}$ is the projection of Y onto the first $k \leq r$ principal components of the kernel:

$$H_{\text{KPCR}}(K) = \sum_{i=1}^{k} u_i u_i^\top \,.$$

The second one is the *kernel ridge regression (KRR) prototype*, where $\widehat{Y}$ is an estimate of Y by kernel ridge regression with parameter $\lambda \geq 0$:

$$H_{\text{KRR}}(K) = K\,(K + \lambda I)^{-1} = \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_i + \lambda} u_i u_i^\top \,.$$

The ridge regression prototype was proposed by Reid et al. (2017) in the linear setting to capture the association between a group of features and an outcome; here we generalize it to the more general kernel setting.

In addition to these prototypes inspired by those used in the linear setting to analyze groups of features, we now show that empirical estimates of the HSIC criterion (Gretton et al., 2005), widely used to assess the association between a kernel and an outcome (Yamada et al., 2018), is also a quadratic kernel association score. More precisely, given two $n \times n$ kernel matrices K and L, Gretton et al. (2005) propose the following measure:

$$\widehat{\text{HSIC}}_{\text{biased}}(K, L) = \frac{1}{(n-1)^2} \operatorname{trace}(K\,\Pi_n\,L\,\Pi_n) \,, \tag{4}$$

where $\Pi_n = I_{n\times n} - \frac{1}{n}1_n 1_n{}^\top$. $\widehat{\text{HSIC}}_{\text{biased}}$ is a biased estimator which converges to the population HSIC measure when n increases.

A second, unbiased empirical estimator, which exhibits a convergence speed in $\frac{1}{\sqrt{n}}$, better than that of $\widehat{\text{HSIC}}_{\text{biased}}$, was developed by Song et al. (2007):

$$\widehat{\text{HSIC}}_{\text{unbiased}}(X,Y) = \frac{1}{n(n-3)}\Bigg[\text{trace}(\underline{K}\,\underline{L}) + \frac{1_n^T\underline{K}1_n\; 1_n^T\underline{L}1_n}{(n-1)(n-2)} - \frac{2}{n-2}1_n^T\underline{K}\,\underline{L}1_n\Bigg], \tag{5}$$

where $\underline{K} = K - \text{diag}(K)$ and $\underline{L} = L - \text{diag}(L)$.

Both empirical HSIC estimators fit in our general family of association scores:

Lemma 1. *The function*

$$s(K,Y) = \widehat{\text{HSIC}}(K, YY^\top)\,,$$

where $\widehat{\text{HSIC}}$ *is either the biased estimator (4) or the unbiased one (5), is a quadratic kernel association score. In addition, the biased estimator is a positive definite quadratic form on* Y *for any kernel* K*.*

Proof. For the biased estimator (4), we simply rewrite it as

$$\widehat{\text{HSIC}}_{\text{biased}}(K, YY^\top) = \frac{1}{(n-1)^2}Y^\top\Pi_n K\Pi_n Y\,,$$

which is a positive quadratic form in Y, corresponding to the hat matrix $K^{1/2}\Pi_n/(n-1)$. For the unbiased estimate, the derivation is also simple but a bit tedious, and is postponed to Appendix A. □

We highlight that this result is fundamentally different from the results of Yamada et al. (2018), who show that, asymptotically, the empirical block estimator of HSIC (Zhang et al., 2018) has a Gaussian distribution. Here we do not focus on the value of the empirical HSIC estimator itself, but on its dependence on Y, which will be helpful later to derive PSI schemes. We also note that Lemma 1 explicitly requires that the kernel L used to model outcomes be the linear kernel, while the approach of Yamada et al. (2018) that leads to a more specific PSI schemes is applicable to any kernel L.

4. Kernel Selection

Given any quadratic kernel association score, we now detail different strategies to select a subset of $S' \leq S$ of kernels among the initial set $\mathcal{K}$. We consider three standards strategies, assuming S' is given:

- *Filtering*: we compute the scores $s(K,Y)$ for all candidate kernels $K \in \mathcal{K}$, and select among them the top S' with the highest scores.
- *Forward stepwise selection*: we start from an empty list of kernels, and iteratively add new kernels one by one in the list by picking the one that leads to the largest increase in association score when combined with the kernels already in the list. This is formalized in Algorithm 1.
- *Backward stepwise selection*: we start from the full list of kernels, and iteratively remove the one that leads to the smallest decrease in association score, as formalized in Algorithm 2.

In addition, we consider *adaptive* variants of these selection methods, where the number S' of selected kernels is not fixed beforehand but automatically selected in a data-driven way. In adaptive estimation of S', we maximize over S' the association score computed at each step, potentially regularized by a penalty function that does not depend on Y. For example, for group selection in the linear regression case, Loftus & Taylor (2015) maximize the association score penalized by an AIC penalty.

Algorithm 1 Forward stepwise kernel selection

Input: set of kernels $\mathcal{K} = \{K_1, \ldots, K_S\}$; outcome $Y \in \mathbb{R}^n$; quadratic kernel association score $s(.,.)$; number of kernels to select $S' \leq S$.
Output: a subset of S' selected kernels.
Init: $\mathcal{I} \leftarrow \mathcal{K}$, $\mathcal{J} \leftarrow \emptyset$.
for $i = 1$ to S' **do**
  $\widetilde{K} \leftarrow \underset{K\in\mathcal{I}}{\text{argmax}} \;\; s\left(K + \sum_{K'\in\mathcal{J}} K', Y\right)$
  $\mathcal{I} \leftarrow \mathcal{I} \setminus \{\widetilde{K}\}$
  $\mathcal{J} \leftarrow \mathcal{J} \cup \{\widetilde{K}\}$
end for
return $\mathcal{J}$

The following result generalizes to the kernel selection problem a result that was proven by Loftus & Taylor (2015) in the feature group selection problem with linear methods.

Theorem 1. *Given a set of kernels* $\mathcal{K} = \{K_1, \ldots, K_S\}$*, a quadratic kernel association score* s*, and a method for kernel selection discussed above (filtering, forward or backward stepwise selection, adaptive or not), let* $\widehat{M}(Y) \subseteq \mathcal{K}$ *be the subset of kernels selected given a vector of outcomes* $Y \in \mathbb{R}^n$*. For any* $M \subseteq \mathcal{K}$*, there exists* $i_M \in \mathbb{N}$*, and* $(Q_{M,1}, b_{M,1}), \ldots, (Q_{M,i_M}, b_{M,i_M}) \in \mathbb{R}^{n\times n} \times \mathbb{R}$ *such that*

$$\{Y \,:\, \widehat{M}(Y) = M\} = \bigcap_{i=1}^{i_M}\{Y \,:\, Y^\top Q_{M,i} Y + b_{M,i} \geq 0\}.$$

Algorithm 2 Backward stepwise kernel selection

Input: set of kernels $\mathcal{K} = \{K_1, \dots, K_S\}$; outcome $Y \in \mathbb{R}^n$; quadratic kernel association score $s(.,.)$; number of kernels to select $S' \leq S$.
Output: a subset of S' selected kernels.
Init: $\mathcal{J} \leftarrow \mathcal{K}$.
for $i = 1$ to $S - S'$ **do**
  $\widetilde{K} \leftarrow \underset{K \in \mathcal{J}}{\operatorname{argmax}} \; s\left(\sum_{K' \in \mathcal{J} \setminus \{K\}} K', Y\right)$
  $\mathcal{J} \leftarrow \mathcal{J} \setminus \{\widetilde{K}\}$
end for
return $\mathcal{J}$

Again, the proof is simple but tedious, and is postponed to Appendix B. Theorem 1 shows that, for a large class of selection methods, we can characterize the set of outcomes Y that lead to the selection of any particular subset of kernels as conjunction of quadratic inequalities. This paves the way to a variety of PSI schemes by conditioning of the event $\widehat{M}(Y) = M$, as explored for example by Loftus & Taylor (2015); Yang et al. (2016) in the case of group selection.

It is worth noting that Theorem 1 is valid in particular when an empirical HSIC estimator is used to select kernels, thanks to Lemma 1. In our setting, the kernel selection procedure proposed by Yamada et al. (2018) corresponds precisely to the filtering selection strategy combined with an empirical HSIC estimator. Hence Theorem 1 allows to derive an exact characterization of the event $\widehat{M}(Y) = M$ in terms of Y, which in turns allows to derive various PSI procedure involving Y, as detailed below. In contrast, Yamada et al. (2018) provide a characterization of the event $\widehat{M}(Y) = M$ not in terms of Y, but in terms of the vector of values $(s(K_i, Y))_{i=1,\dots,S}$. Combined with the approximation that this vector is asymptotically Gaussian when n tends to infinity, this allows Yamada et al. (2018) to derive PSI schemes to assess the values $s(K_i, Y)$ of the selected kernel. Theorem 1 therefore provides a result which is valid non-asymptotically, and which allows to test other types of hypotheses, such as the association of one particular kernel with the outcome, given other selected kernels.

5. Statistical Inference

Let us consider the general model

$$Y = \mu + \sigma^2 \epsilon \,, \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, I_n)$ and $\mu \in \mathbb{R}^n$. Characterizing the set $E = \{Y : \widehat{M}(Y) = M\}$ allows to answer a variety of statistical inference questions about the true signal μ and its association with the different kernels, conditional to the fact that a given set of kernels M has been selected.

For example, testing whether $s(K, \mu) = 0$ for a given kernel $K \in M$, or for the combination of kernels $K = \sum_{K' \in M} K'$, is a way to assess whether K captures information about μ. This is the test carried out by Yamada et al. (2018) to test each individual kernel after selection by marginal HSIC screening. Alternatively, to test whether a given kernel $K \in M$ has information about μ not redundant with the other selected kernels in $M \setminus \{K\}$, one may test whether the prototype of μ built from all kernels in M is significantly better that the prototype built without K. This can translate into testing whether

$$s\left(\sum_{K' \in M} K', \mu\right) = s\left(\sum_{K' \in M, K' \neq K} K', \mu\right) .$$

Such a test is performed by Loftus & Taylor (2015); Yang et al. (2016) to assess the significance of groups of features in the linear setting, using the projection prototype.

In general, testing a null hypothesis of the form $s(K, \mu) = 0$ for a positive quadratic form s can be done by forming the statistics $V = \|H(K)Y\|^2$, where H is the hat matrix associated with s, and studying its distribution conditionally on the event $Y \in E$. The fact that E is an intersection of subsets defined by quadratic constraints can be exploited to derive computationally efficient procedures to estimate p-values and confidence intervals when, for example, $H(K)$ is a projection onto a subspace (Loftus & Taylor, 2015; Yang et al., 2016). We can directly borrow these techniques in our setting, for example for the KPCR prototype, where $H(K)$ is a projection matrix. For more general $H(K)$ matrices, the techniques of Loftus & Taylor (2015); Yang et al. (2016) need to be adapted; another way to proceed is to estimate the distribution of V by Monte-Carlo sampling, as explained in the next section.

Alternatively, Reid et al. (2017) propose to test the significance of groups of features through prototypes, which they argue uses fewer degrees of freedom than statistics based on the norms of prototypes, which can increase statistical power. We adapt this idea to the case of kernels and show here how to test the association of a single kernel (whether one of the selected kernels, or their aggregation) with the outcome. We refer the reader to Reid et al. (2017) for extensions to several groups, that can be easily adapted to several kernels. Given a prototype $\widehat{Y} = H(K)Y$, Reid et al. (2017) propose to test the null hypothesis $H_0 : \theta = 0$ in the following univariate model:

$$Y = \mu + \theta \widehat{Y} + \sigma^2 \epsilon,$$

where again $\epsilon \sim \mathcal{N}(0, I_n)$, μ is fixed, and θ is the parameter of interest. One easily derives the log-likelihood:

$$\ell_Y(\theta) = \log|I - \theta H(K)| - \frac{1}{2\sigma^2}\|Y - \mu - \theta H(K)Y\|^2 \,,$$

which is a concave function of θ that can be maximized by Newton-Raphson iterations to obtain the maximum likelihood estimator $\widehat{\theta} \in \operatorname{argmax}_\theta \ell_Y(\theta)$. We can then form the likelihood ratio statistics

$$R(Y) = 2\left(\ell_Y(\widehat{\theta}) - \ell_Y(0)\right), \tag{7}$$

and study the distribution of $R(Y)$ under H_0 to perform a statistical test and derive a p-value. While $R(Y)$ asymptotically follows a χ^2_1 distribution under H_0 when we do not condition on Y (Reid et al., 2017), its distribution conditioned on the event $\widehat{M}(Y) = M$ is different and must be determined for valid PSI. As this conditional distribution is unlikely to be tractable, we propose to approximate it thanks to empirical sampling. This allows us to derive valid empirical PSI p-values as the fraction of samples Y_t for which $R(Y_t)$ is larger than the $R(Y)$ computed from the data.

6. Constrained Sampling

We now discuss how to sample T replicates $Y_1, \ldots, Y_T$ according to the Gaussian model (6) conditional to the event $\widehat{M}(Y) = M$. As explained in the previous section, this is needed to derive p-values for various statistical tests.

By Theorem 1, all replicates must be sampled within the acceptance region defined by a series of quadratic constraints on Y. Several strategies can be deployed to this end. The most straightforward one is rejection sampling, which consists in sampling independently Y_t from $\mathcal{N}\left(\mu, \sigma^2 I_n\right)$, and only retaining samples for which all quadratic constraints are satisfied, i.e., $Y_t^T Q_{M,i} Y_t + b_{M,i} \geq 0$, for $i \in \{1, \cdots, i_M\}$. Such a strategy can be time-consuming, especially if the volume of the acceptance region is small, leading to a high number of rejections. Alternatively, one could use the the Hamiltonian Monte Carlo algorithm of Pakman & Paninski (2014). In practice, we found that for large values of n, it does not scale well enough to generate a sufficient number of replicates T. Therefore, we propose a new hit-and run sampler below.

Our proposed sampler is based on the Hypersphere Directions (HD) algorithm, first proposed by Berbee et al. (1987) to detect nonredundant constraints in a system of linear inequalities. The main assumption in the HD algorithm is that the acceptance region is open and bounded. In our case, the boundedness assumption does not necessarily hold. For example, if $b_{M,i} = 0$ for all $i = 1, \ldots, i_M$, then the acceptance region is clearly an unbounded cone, that is, if $Y \in E$ then $\lambda Y \in E$ for any $\lambda \geq 0$. To use the HD algorithm nevertheless, we apply the reparametrization $Z = F(Y)$, where $F : \mathbb{R}^n \to]0,1[^n$ is given by $F(Y)_i = F_{\mu_i,\sigma^2}(Y_i)$ for $i = 1, \ldots, n$. Here $F_{\mu_i,\sigma^2}(Y_i)$ denotes the cumulative distribution function (c.d.f.) of the normal distribution $\mathcal{N}\left(\mu_i, \sigma^2\right)$. Without conditioning, Z is uniformly distributed over $]0,1[^n$, and when we condition on $Y \in E$, Z is uniformly distributed on the truncated space region $\mathcal{M}$ given by the quadratic constraints:

$$F^{-1}(Z) Q_{M,i} F^{-1}(Z) + b_{M,i} > 0, \ \forall i \in \{1, \cdots, i_M\}.$$

We use strict inequalities so that $\mathcal{M}$ is both open and bounded; this does not affect the probabilities we estimate.

Algorithm 3 Hypersphere Directions hit-and-run sampler

Input: Y an admissible point, T the total number of replicates and B the number of burn-in iterations.
Output: a sample of T replicates sampled according to the conditional distribution.
Init: $Z_0 \leftarrow F^{-1}(Y),\ t \leftarrow 0$
repeat
  $t \leftarrow t + 1$
  Sample uniformly θ_t from $\{\theta \in \mathbb{R}^n,\ ||\theta|| = 1\}$ [1]
  $a_t \leftarrow \max\left\{ \max_{\theta_t^{(i)} > 0} -\frac{Z_{t-1}}{\theta_t};\ \max_{\theta_t^{(i)} < 0} \frac{1 - Z_{t-1}}{\theta_t} \right\}$
  $b_t \leftarrow \min\left\{ \min_{\theta_t^{(i)} < 0} -\frac{Z_{t-1}}{\theta_t};\ \min_{\theta_t^{(i)} > 0} \frac{1 - Z_{t-1}}{\theta_t} \right\}$
  repeat
    Sample uniformly λ_t from $]a_t, b_t[$
    $Z_t \leftarrow Z_{t-1} + \lambda_t \theta_t$
    $Y_t \leftarrow F^{-1}(Z_t)$
  until $Z_t \in \mathcal{M}$
until $t = B + T$
return $\left\{Y_{B+1}, \cdots, Y_{B+T}\right\}$

Algorithm 3 presents our hit-and-run sampler (Blisle et al., 1993), based on iteratively sampling in the hypercube. In the HD algorithm, the unidimensional parameter λ_t is sampled according to the p.d. $f_t^\lambda(\lambda_t | Z_{t-1}, \theta_t) \propto f(Z_{t-1} + \lambda_t \theta_t)$, where f is the p.d. of $Z = F(Y)$. Given that Z is uniformly distributed on $\mathcal{M}' =]0,1[^n \cap \mathcal{M}$, λ_t is then uniformly distributed on the region $\Lambda = \{\lambda \text{ s.t. } Z_{t-1} + \lambda \theta_t \in \mathcal{M}'\}$. To sample λ_t, we first start by uniformly sampling on the interval $[a_t, b_t]$ to ensure that $Z_{t-1} + \lambda_t \theta_t \in]0,1[^n$. The sample λ_t is accepted if $Z_{t-1} + \lambda \theta_t \in \mathcal{M}$.

Though our sampling of λ_t is also a rejection sampling, the resulting hit-and-run sampler is faster than a mere rejection sampling of Y_t. Indeed, λ_t is unidimensional while each replicate Y_t is an n-dimensional normal variable. Moreover, the initial sampling on the interval $]a_t, b_t[$ reduces the total number of rejections.For a proof of the convergence of the HD sampler, we refer the reader to Smith (1984).

In hit-and-run samplers, to generate valid p-values, a large number of burn-in iterations and of replicates are needed. The burn-in period reduces the dependence on the original

[1] A classical technique to uniformly sample from the n-dimensional sphere is to first sample θ_t from $\mathcal{N}(0,1)$ and normalize, $\theta_t \leftarrow \theta_t / ||\theta_t||_2$

sample Y, while the large number of replicates addresses the correlation between consecutive replicates.

7. Experiments

In our experiments, we focus on the case where each kernel corresponds to a predefined group of features, and where we test the association of the sum of the selected kernels with the outcome. We use $\widehat{\text{HSIC}}_{\text{unbiased}}$ as a quadratic kernel association score for kernel selection in all our experiments.

7.1. Statistical Validity

We first demonstrate the statistical validity of our PSI procedure, which we refer to as kernelPSI. We simulate a design matrix X of $n = 100$ samples and $p = 50$ features, partitioned in $S = 10$ disjoint and mutually-independent subgroups of $p' = 5$ features, drawn from a normal distribution centered at 0 and with a covariance matrix $V_{ij} = \rho^{|i-j|}, \quad i, j \in \{1, \cdots p'\}$. We set the correlation parameter ρ to 0.6. To each group corresponds a *local* Gaussian kernel K_i, of variance $\sigma^2 = 5$.

The outcome Y is drawn as $Y = \theta K_{1:3} U_1 + \epsilon$, where $K_{1:3} = K_1 + K_2 + K_3$, U_1 is the eigenvector corresponding to the largest eigenvalue of $K_{1:3}$, and ϵ is Gaussian noise centered at 0. We vary the effect size of θ across the range $\theta \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, and resample Y 1 000 times to create 1 000 simulations.

In this particular setting where the *local* kernels are additively combined, the three kernel selection strategies in Section 4 are equivalent. Along with the *adaptive* variant, we consider 3 variants with a predetermined number of kernels, $S' \in \{1, 3, 5\}$. For inference, we compute the likelihood ratio statistics for KPCR or KRR prototypes, or directly use $\widehat{\text{HSIC}}_{\text{unbiased}}$ as a test statistic (see Section 5). Finally, we used our hit-and-run sampler to provide empirical p-values (see Section 6), fixing the number of replicates at $T = 5 \times 10^4$ and the number of burn-in iterations at 10^4.

Figure 1 shows the Q-Q plot comparing the distribution of the p-values provided by kernelPSI with the uniform distribution, under the null hypothesis ($\theta = 0.0$). All variants give data points aligned with the first diagonal, confirming that the empirical distributions of the statistics are uniform under the null.

Figure 2 shows the Q-Q plot comparing the distribution of the p-values provided by kernelPSI with the uniform distribution, under the alternative hypothesis where $\theta = 0.3$. We now expect the p-values to deviate from the uniform. We observe that all kernelPSI variants have statistical power, reflected by low p-values and data points located towards the bottom right of the Q-Q plot. The three strategies (KPCR, KRR and HSIC) enjoy greater statistical power for smaller

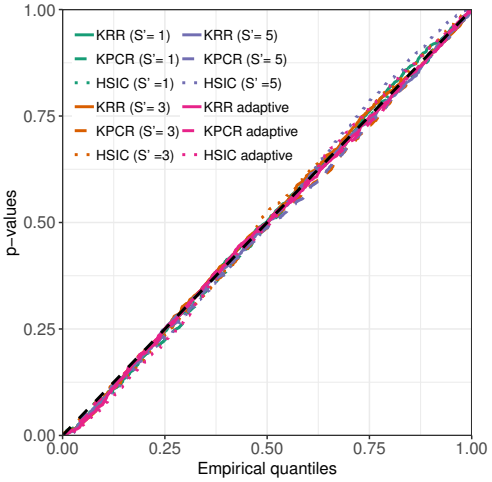

Figure 1. Q-Q plot comparing the empirical kernelPSI p-values distributions under the null hypothesis ($\theta = 0.0$) to the uniform distribution.

number of selected kernels. Because of the selection of irrelevant kernels, statistical power decreases when S' increases. The same remark holds for the adaptive variants, which performs similarly to the fixed variant with $S' = 5$. In fact, the average support size for the adaptive kernel selection procedure is $\overline{S'} = 5.05$. We also observe that HSIC has more statistical power than the KRR or KPCR variants, possibly because we used an HSIC estimator for kernel selection, making the inference step closer to the selection one.

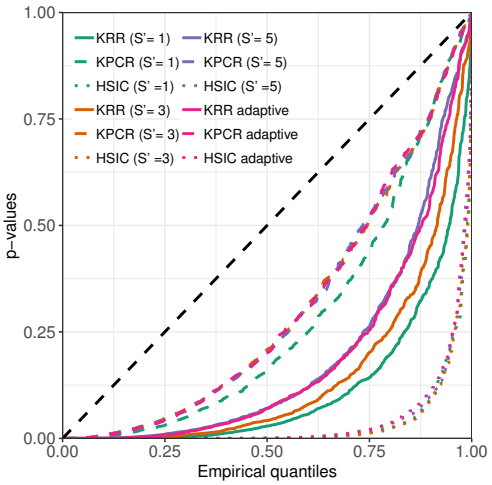

Figure 2. Q-Q plot comparing the empirical kernelPSI p-values distributions under the alternative hypothesis ($\theta = 0.3$) to the uniform distribution.

7.2. Benchmarking

We now evaluate the performance of the kernelPSI procedure against a number of alternatives:

- *protoLasso:* the original, linear prototype method for post-selection inference with L_1-penalized regression (Reid et al., 2017);
- *protoOLS:* a selection-free alternative, where the prototype is obtained from an ordinary least-squares regression, and all variables are retained;
- *protoF:* a classical goodness-of-fit F-test. Here the prototype is constructed similarly as in protoOLS, but the test statistic is an F-statistic rather that a likelihood ratio;
- *KPCR, KRR, and HSIC:* the non-selective alternatives to our kernelPSI procedure. KPCR and KRR are obtained by constructing a prototype over the sum of all kernels, without the selection step. HSIC is the independence test proposed by Gretton et al. (2008);
- *SKAT:* The Sequence Kernel Association Test (Wu et al., 2011) tests for the significance of the joint effect of all kernels in a non-selective manner, using a quadratic form of the residuals of the null model.

We consider the same setting as in Section 7.1, but now add benchmark methods and additionally consider linear kernels over binary features, a setting motivated by the application to genome-wide association studies, where the features are discrete. In this last setting, we vary the effect size θ over the range $\{0.01, 0.02, 0.03, 0.05, 0.07, 0.1\}$. We relegate to Appendix C.4.2 an experiment with Gaussian kernels over Swiss roll data.

Figures 3 and 4 show the evolution of the statistical power as a function of the effect size θ in, respectively, the Gaussian and the linear data setups. These figures confirm that kernel-based methods, particularly selective HSIC and SKAT, are superior to linear ones such as protoLASSO. We observe once more that the selective HSIC variants have more statistical power than their KRR or KPCR counterparts, that methods selecting fewer kernels enjoy more statistical power, and that adaptive methods tend to select too many kernels (closer to $S' = 5$ than to the true $S' = 3$). We also observe that the selective kernelPSI methods ($S' = 1, 3, 5$ or adaptive) have more statistical power than their non-selective counterparts.

Finally, we note that, in the linear setting, the KRR and KPCR variants perform similarly. We encounter a similar behavior in simulations (not shown) using a Wishart kernel. Depending on the eigenvalues of K, the spectrum of the transfer matrix $H_{\text{KRR}} = K(K + \lambda I_{n\times n})^{-1}$ can be concentrated around 0 and 1. H_{KRR} becomes akin to a projector matrix, and KRR behaves similarly to KPCR.

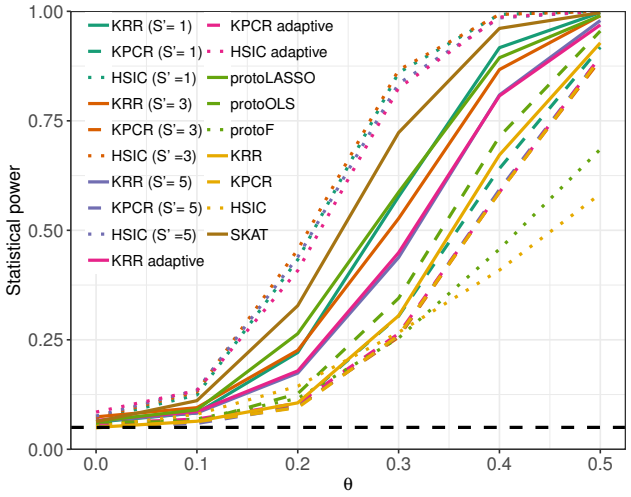

Figure 3. Statistical power of kernelPSI variants and benchmark methods, using Gaussian kernels for simulated Gaussian data.

In addition, we evaluate the ability of our kernel selection procedure to recover the three true causal kernels used to simulate the data. Table 1 reports the evolution of the precision and recall of our procedures, in terms of selected kernels, for increasing effect sizes in the Gaussian kernels and data setting. Note that when S' is fixed, a random selection method is expected to have a precision of $3/10$ (the proportion of kernels that are causal), and a recall of $S'/10$, which corresponds to the values we obtain when there is no signal ($\theta = 0$). As the effect size θ increases, both precision and recall increase.

When S' increases, the precision increases and the recall decreases, which is consistent with our previous observations that increasing S' increases the likelihood to include irrelevant kernels in the selection. Once again, the performance of the adaptive kernelPSI is close to that of the setting where the number of kernels to select is fixed to 5, indicating that the adaptive version tends to select too many kernels.

7.3. Case Study: Selecting Genes in a Genome-Wide Association Study

In this section, we illustrate the application of kernelPSI on genome-wide association study (GWAS) data. Here we study the flowering time phenotype "FT_GH" of the *Arabidopsis thaliana* dataset of Atwell et al. (2010). We are interested in using the 166 available samples to test the association of each of 174 candidate genes to this phenotype. Each gene is represented by the single-nucleotide polymorphisms (SNPs) located within $\pm$ 20-kilobases. We use hierarchical clustering to create groups of SNPs within each gene; these clusters are expected to correspond to

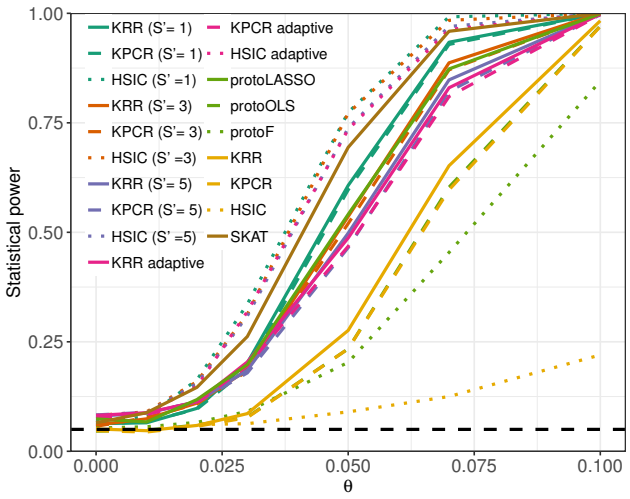

Figure 4. Statistical power of kernelPSI variants and benchmark methods, using linear kernels for simulated binary data.

Table 1. Ability of the kernel selection procedure to recover the true causal kernels, using Gaussian kernels over simulated Gaussian data.

	θ	$S' = 1$	$S' = 3$	$S' = 5$	Adaptive
Recall	0.0	0.102	0.302	0.505	0.435
	0.1	0.150	0.380	0.569	0.523
	0.2	0.263	0.528	0.690	0.678
	0.3	0.324	0.630	0.770	0.768
	0.4	0.332	0.691	0.830	0.822
	0.5	0.333	0.733	0.862	0.855
Precision	0.0	0.306	0.302	0.303	0.305
	0.1	0.450	0.380	0.341	0.352
	0.2	0.791	0.528	0.414	0.437
	0.3	0.974	0.630	0.462	0.485
	0.4	0.997	0.691	0.498	0.518
	0.5	1.000	0.733	0.517	0.548

linkage disequilibrium blocks. As is common for GWAS applications, we use the identical-by-state (IBS) kernel (Kwee et al., 2008) to create one kernel by group. We then apply our kernelPSI variants as well as the baseline algorithms used in Section 7.2. Further details about our experimental protocol are available in Appendix C.6.

We first compare the p-values obtained by the different methods using Kendall's tau coefficient τ to measure the rank correlation between each pair of methods (see Appendix C.7). All coefficients are positive, suggesting a relative agreement between the methods. We also resort to non-metric multidimensional scaling (NMDS) to visualize the concordance between the methods (see Appendix C.9). Altogether, we observe that related methods are located nearby (e.g. KRR near KPCR, protoLASSO near protoOLS, etc.), while selective methods are far away from non-selective ones.

Our first observation is that none of the non-selective methods finds any gene significantly associated with the phenotype ($p < 0.05$ after Bonferroni correction), while our proposed selective methods do. A full list of genes detected by each method is available in Appendix C.8. None of those genes have been associated to this phenotype by traditional GWAS (Atwell et al., 2010). We expect the most conservative methods ($S' = 1$) to yield the fewest false positive, and hence focus on those. KRR, KPCR and HSIC find, respectively, 2, 2, and 1 significant genes. One of those, AT5G57360, is detected by all three methods. It is interesting to note that this gene has been previously associated with a very related phenotype, FT10, differing from ours only in the greenhouse temperature (10°C vs 16°C). This is also the case of the other gene detected by KRR, AT5G65060. Finally, the second gene detected by KPCR, AT4G00650, is the well-known FRI gene, which codes for the FRIGIDA protein, required for the regulation of flowering time in late-flowering phenotypes. All in all, these results indicate that our proposed kernelPSI methods have the power to detect relevant genes in GWAS and are complementary to existing approaches.

8. Conclusion

We have proposed kernelPSI, a general framework for post-selection inference with kernels. Our framework rests upon quadratic kernel association scores to measure the association between a given kernel and the outcome. The flexibility in the choice of the kernel allows us to accommodate a broad range of statistics. Conditionally on the kernel selection event, the significance of the association with the outcome of a single kernel, or of a combination of kernels, can be tested. We demonstrated the merits of our approach on both synthetic and real data. In addition to its ability to select causal kernels, kernelPSI enjoys greater statistical power than state-of-the-art techniques. A future direction of our work is to scale kernelPSI to larger datasets, in particular with applications to full GWAS data sets in mind, for example by using the block HSIC estimator (Zhang et al., 2018) to reduce the complexity in the number of samples. Another direction would be to explore whether our framework can also incorporate Multiple Kernel Learning (Bach, 2008). This would allow us to complement our filtering and wrapper kernel selection strategies with an embedded strategy, and to construct an aggregated kernel prototype in a more directly data-driven fashion.

References

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J., and Nordborg, M. Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465 (7298):627–631, mar 2010. doi: 10.1038/nature08800.

Bach, F. R. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, June 2008. ISSN 1532-4435.

Berbee, H. C. P., Boender, C. G. E., Ran, A. H. G. R., Scheffer, C. L., Smith, R. L., and Telgen, J. Hit-and-run algorithms for the identification of nonredundant linear inequalities. *Mathematical Programming*, 37(2):184–207, jun 1987.

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

Blisle, C. J. P., Romeijn, H. E., and Smith, R. L. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993.

Cox, D. R. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Scholkopf, B., and Logothetis, N. Kernel constrained covariance for dependence measurement. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 1–8, January 2005.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. A Kernel Statistical Test of Independence. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 585–592. Curran Associates, Inc., 2008.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397, feb 2008. doi: 10.1016/j.ajhg.2007.10.010.

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, jun 2016. doi: 10.1214/15-aos1371.

Loftus, J. R. and Taylor, J. E. Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*, 2015.

Pakman, A. and Paninski, L. Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, apr 2014. doi: 10.1080/10618600.2013.788448.

Reid, S., Taylor, J., and Tibshirani, R. A general framework for estimation and inference from clusters of features. *Journal of the American Statistical Association*, 113(521):280–293, sep 2017. doi: 10.1080/01621459.2016.1246368.

Smith, R. L. Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions. *Operations Research*, 32(6):1296–1308, 1984.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M., and Bedo, J. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning - ICML '07*. ACM Press, 2007. doi: 10.1145/1273496.1273600.

Taylor, J. and Tibshirani, R. J. Statistical learning and selective inference. *Proc. Natl. Acad. Sci. U.S.A.*, 112: 7629–7634, June 2015.

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, apr 2016. doi: 10.1080/01621459.2015.1108848.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, jul 2011. doi: 10.1016/j.ajhg.2011.05.029.

Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. Post selection inference with kernels. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 152–160, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Yang, F., Barber, R. F., Jain, P., and Lafferty, J. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pp. 2469–2477, 2016.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018. ISSN 1573-1375. doi: 10.1007/s11222-016-9721-7.