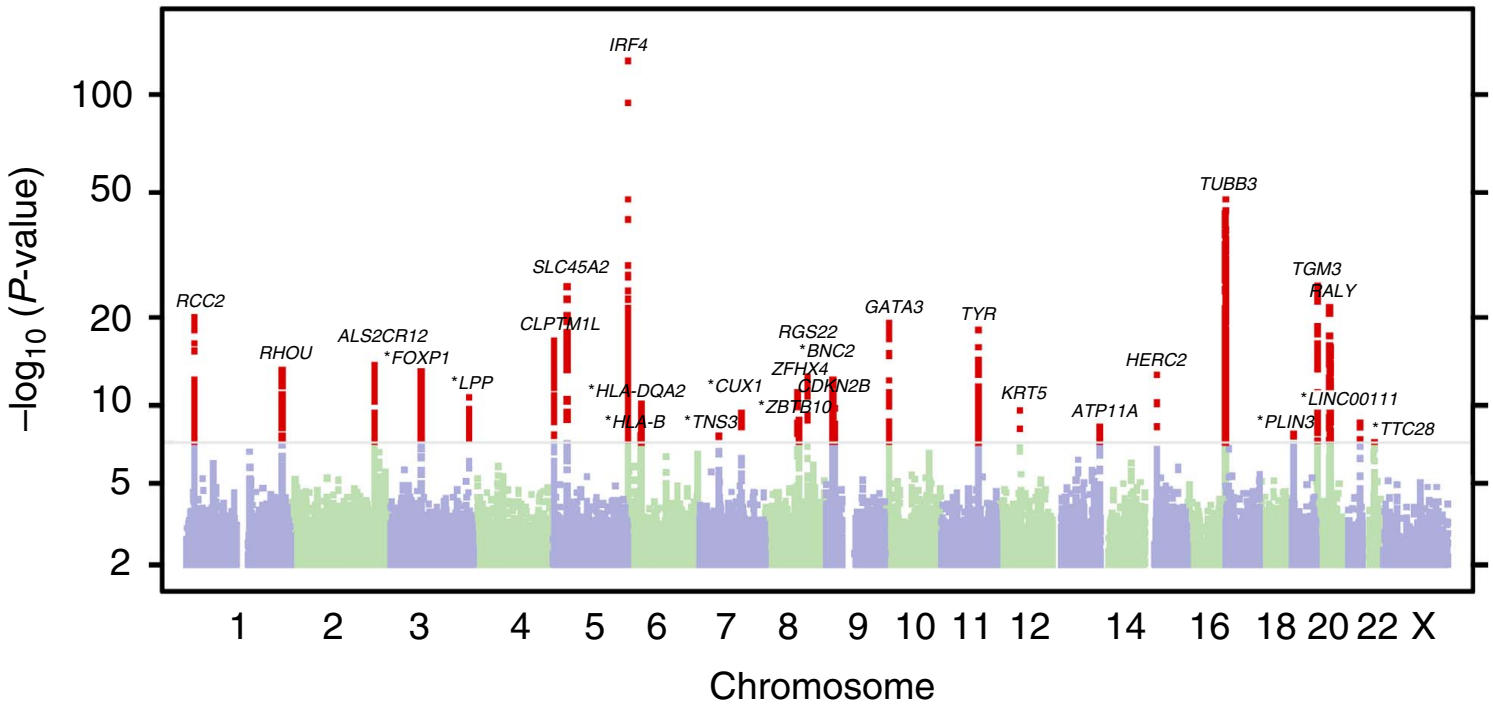
Post-Selection Inference for Nonlinear Feature Selection

L. Slim, C. Chatelain, C.A. Azencott & J.P. Vert

SANOFI / MINES ParisTech / Google

Motivation



- Nonlinear feature selection to identify genes
- Followed by valid statistical inference (P-value, confidence interval for association...)

Challenge: file drawer effect (aka publication bias)

Typical lab experiment :

- Measurement of n different variables of interest $(Y_i)_{i=1,\cdots,n}$. Each variable is normally distributed, $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
- Since we are interested in large effects, we only select such ones, e.g.:

$$\hat{\mathcal{I}} = \left\{ i \in \{1, \cdots, n\} \text{ s. t.} |Y_i| > 1 \right\}$$

- Hypothesis testing for $H_0 : \mu_i = 0, \ \forall i \in \hat{\mathcal{I}}$
  - Reject H_0, if $|Y_i| > 1.96$ (confidence interval for $\alpha = 0.05$) ?

Challenge: file drawer effect (aka publication bias)

Typical lab experiment :

- Measurement of n different variables of interest $(Y_i)_{i=1,\cdots,n}$. Each variable is normally distributed, $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
- Since we are interested in large effects, we only select such ones, e.g.:

$$\hat{\mathcal{I}} = \left\{ i \in \{1, \cdots, n\} \text{ s. t.} |Y_i| > 1 \right\}$$

- Hypothesis testing for $H_0 : \mu_i = 0, \ \forall i \in \hat{\mathcal{I}}$
  - Reject H_0, if $|Y_i| > 1.96$ (confidence interval for $\alpha = 0.05$) ? **Wrong** ! More than 5% of hypothesis will be rejected under H_0
  - Proper way: condition on the selection event,

$$P\left(|Y_i| > L_\alpha \mid |Y_i| > 1\right) = 0.05 \Rightarrow L_\alpha = 2.41 > 1.96$$

Post-selection inference (PSI)

- Observe data Y
- Select model $\hat{M}$ which depends on Y
  - e.g., a subset of features for sparse regression
- Derive the distribution of a statistics of interest $S_{\hat{M}}(Y)$ conditionally on $\hat{M}(Y) = \hat{M}$
  - e.g., weight of a given feature $i \in \hat{M}$ in a linear regression model restricted to $\hat{M}$

Example: PSI for lasso regression (Lee et al., 2016)

$$\hat{\beta} \in \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1 , \quad \hat{M} = \{i : \hat{\beta}_i \neq 0\}$$

- For any M,

$$\{y : \hat{M}(y) = M\} = \cup_s \{y : A(M,s)y \leq b(M,s)\}$$

- Statistics of the form $\eta_{\hat{M}}^\top y$
- Polyhedral lemma: if $Y = \mu + \sigma^2 I$, then for any vector η,

$$F^{[V^-,V^+]}_{\eta^\top\mu,\sigma^2\eta^\top\eta}(\eta^\top Y)|\{AY \leq b\} \sim Unif(0,1),$$

where $F^{[a,b]}_{\mu,\sigma^2}$ is the c.d.f of a truncated Gaussian distribution, and V^-, V^+ are constants that are functions of η, A, b.

Objective

Extend PSI to nonlinear feature selection. For that:

1. Define nonlinear association scores $s(i, y)$ between feature(s) i and outcome y
2. Define a procedure to select a group of features $\hat{M}$
3. Characterize $\{y : \hat{M} = M\}$
4. Deduce PSI distribution of a statistics of interest

Kernels

- Instead of "features", we assume a collection of kernels $K_1, \ldots K_S$
- Includes linear setting when K_i is the linear kernel on the i-th feature
- Generalize to nonlinear feature selection when K_i is a nonlinear kernel on the i-th features
- Generalization to non-numeric features
- Generalization to group selection

Association based on prototypes

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

where $\widehat{Y}_K = H(K)Y$ is called a *prototype* for a "hat" function $H : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ (Reid et al., 2017).

- *Kernel principal component regression* (KPCR)

$$H_{proj}(K) = KK^+ = \sum_{i=1}^{r} u_i u_i^\top,$$

  where $u_1, \ldots, u_r$ are the eigenvectors of K with nonzero eigenvalues (Loftus and Taylor, 2015).
- *Kernel principal component regression* (KPCR) for some $R < r$:

$$H_{KPCR}(K) = \sum_{i=1}^{R} u_i u_i^\top$$

- *Kernel ridge regression* (KRR) for some $\lambda > 0$

$$H_{KRR}(K) = K(K + \lambda I)^{-1}$$

Association based on HSIC

Take $s(K, Y) = \widehat{\text{HSIC}}(K, YY^\top)$ with one empirical estimator of HSIC (Gretton et al., 2005):

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

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

where $\Pi_n = I_{n\times n} - \frac{1}{n} 1_n 1_n^\top$, $\underline{K} = K - \text{diag}(K)$ and $\underline{L} = L - \text{diag}(L)$.

Quadratic association

Theorem

All aforementioned assocation scores are quadratic kernel association score, i.e., functions $s : \mathbb{R}^{n\times n} \times \mathbb{R}^n \mapsto \mathbb{R}$ *of the form*

$$s(K, Y) = Y^\top Q(K) Y,$$

for a Gram matrix K *and some function* $Q : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$.

$Q(K)$ is positive semidefinite for all but $\widehat{\mathrm{HSIC}}_{\text{unbiased}}$

Stepwise kernel selection: selection strategies

For a fixed number of selected kernels S', we can deploy the following kernel selection strategies

- *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* (Song et al., 2007): 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.
- *Backward stepwise selection* (Song et al., 2007): we start from the full list of kernels, and iteratively remove the one that leads to the smallest decrease in association score.

For the adaptive equivalents, S' is automatically selected in a data-driven fashion. We maximize over S' the association score at each step.

Stepwise kernel selection: selection event

Theorem

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*

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

Statistical inference

- Model $Y = \mu + \sigma^2\epsilon$ and test:
  - $s(K, \mu) = 0$ for $K \in \hat{M}$ or $K = \sum_{K' \in \hat{M}} K'$ (Yamada et al., 2018)
  - $s\left(\sum_{K' \in \hat{M}} K', \mu\right) = s\left(\sum_{K' \in \hat{M}, K' \neq K} K', \mu\right)$ (Loftus and Taylor, 2015; Yang et al., 2016)
- Model $Y = \mu + \theta\hat{Y} + \sigma^2\epsilon$ and test $\theta = 0$ (Reid et al., 2017)

Besides a few cases, we need to computed *empirical* p-values,

- by approximating the distribution of the test statistic,
- by generating replicates of Y within the acceptance region $\{Y : \widehat{M}(Y) = M\}$.

Constrained sampling

- A simple rejection sampling algorithm is cumbersome for small acceptance regions in high-dimensional spaces
- The Hamiltonian Monte-Carlo algorithm from Pakman and Paninski (2014) is difficult to scale
- Closest thing in the literature: the Hypersphere Direction (Berbee et al., 1987): truncated **uniform** distributions on **bounded** space regions

To make it work, a smart trick is to use the c.d.f F of Y (see paper for details)

Experiments: statistical validity

- X an 100×50 design matrix
- The features are partitioned in $S = 10$ disjoint and mutually-independent subgroups of $p' = 5$ features.
- Within each group, we sample from normal distribution centered at 0 and with a covariance matrix $V_{ij} = \rho^{|i-j|}$
- $Y = \theta K_{1:3} U_1 + \epsilon$, where $K_{1:3} = K_1 + K_2 + K_3$, U_1 is the eigenvector corresponding to the largest eigenvalue of $K_{1:3}$
- $\theta \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$
- X is fixed, but Y is resampled 1000 times to create 1000 simulations.

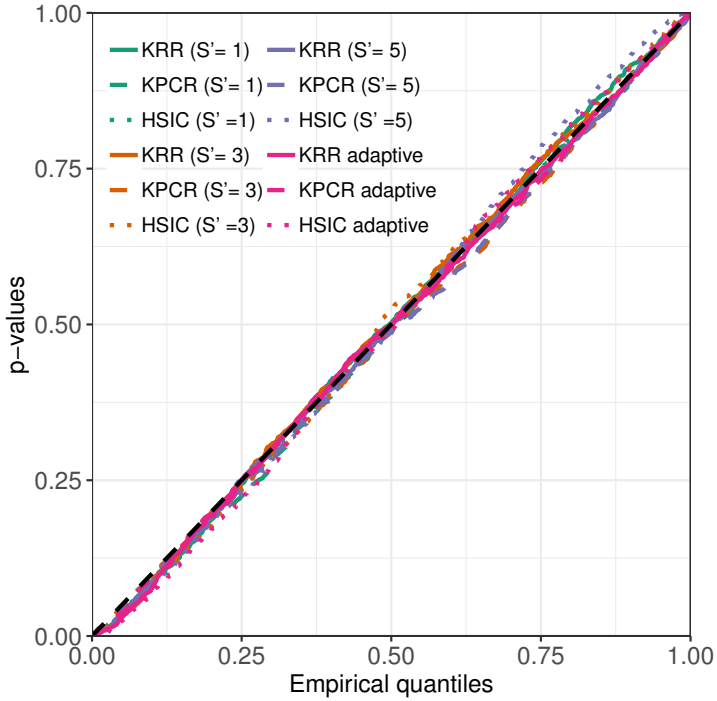


Figure: Q-Q plot comparing the empirical kernelPSI p-values distributions under the null hypothesis ($\theta = 0.0$) to the uniform distribution.

Experiments: benchmarking

We benchmark against the following methods

- *protoLasso:* the original, linear prototype method for PSI with L_1-penalized regression Reid et al. (2017);
- *protoOLS:* a selection-free OLS prototype
- *protoF:* a classical goodness-of-fit F-test for the OLS prototype
- *KPCR, KRR, and HSIC:* the non-selective alternatives to our kernelPSI procedure.
- *SKAT (Wu et al., 2011):* a non-selective quadratic kernel association score.

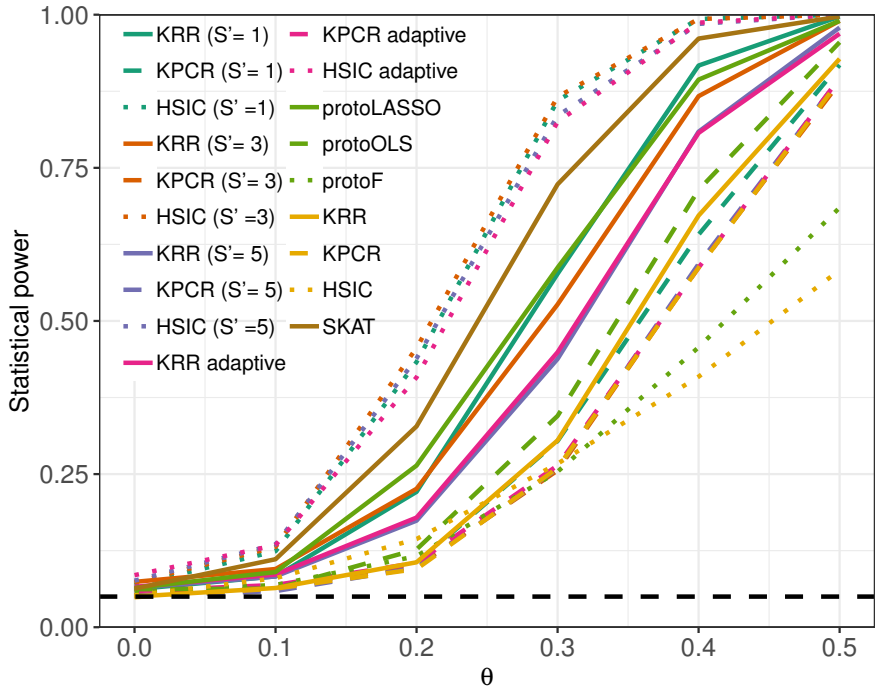


Figure: Statistical power of kernelPSI variants and benchmark methods, using Gaussian kernels for simulated Gaussian data.

Conclusion

- Nonlinear feature selection with valid PSI.
- Open questions: better association measures for nonlinear variable selection, constrained sampling, PSI beyond linear models, large-scale kernel methods, MKL.
- `https://github.com/EpiSlim/kernelPSI`

References I

H. C. P. Berbee, C. G. E. Boender, A. H. G. Rinnooy Ran, C. L. Scheffer, R. L. Smith, and J. Telgen. Hit-and-run algorithms for the identification of nonredundant linear inequalities. *Mathematical Programming*, 37(2): 184–207, jun 1987. doi: 10.1007/bf02591694. URL `https://doi.org/10.1007/bf02591694`.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, jun 2016. doi: 10.1214/15-aos1371. URL `https://doi.org/10.1214/15-aos1371`.

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

References II

Ari Pakman and Liam Paninski. 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. URL https://doi.org/10.1080/10618600.2013.788448.

Stephen Reid, Jonathan Taylor, and Robert Tibshirani. 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. URL https://doi.org/10.1080/01621459.2016.1246368.

Le Song, Alex Smola, Arthur Gretton, Karsten M. Borgwardt, and Justin Bedo. 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. URL https://doi.org/10.1145/1273496.1273600.

References III

Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 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. URL `https://doi.org/10.1016/j.ajhg.2011.05.029`.

Makoto Yamada, Yuta Umezu, Kenji Fukumizu, and Ichiro Takeuchi. Post selection inference with kernels. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 152–160, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pages 2469–2477, 2016.