Nonlinear post-selection inference for genome-wide association studies

Lotfi Slim [1 2] **Clément Chatelain** [1] **Chloé-Agathe Azencott** [2 3]

Abstract

Association testing in genome-wide association studies (GWAS) is often performed at either the SNP level or the gene level. The two levels can bring different insights into disease mechanisms. In the present work, we provide a novel approach based on nonlinear post-selection inference to bridge the gap between them. Our approach selects, within a gene, the SNPs or LD blocks most associated with the phenotype, before testing their combined effect. Both the selection and the association testing are conducted nonlinearly. We apply our tool to the study of BMI and its variation in the UK BioBank. In this study, our approach outperformed other gene-level association testing tools, with the unique benefit of pinpointing the causal SNPs.

1. Introduction

Lack of statistical power is a major limitation in GWAS. If the analysis is performed at the SNP level, lack of statistical power may stem from small effect sizes and linkage disequilibrium, among others. By modeling the overall association signal, gene level analysis can address this limitation. Being the functional entity, genes have the potential to shed light on yet undiscovered biological and functional mechanisms. However, the incorporation of all mapped SNPs, including non-causal ones, can mask the association signal. An alternative strategy would be to select the SNPs most associated with the phenotype within a given gene, and then test their joint effect. If we do not account for the fact that these SNPs were selected in a first step based on the same data, their overall joint effect is likely to be overestimated. Post-selection inference (PSI) (Lee et al., 2016) was specifically developed to correct for selection bias. In addition, such a framework would also benefit from the incorporation of nonlinearities to model epistatic interactions between neighboring SNPs.

In previous work, we published the theoretical foundations of kernelPSI, a post-selection inference (PSI) framework for nonlinear variable selection (Slim et al., 2019). We introduced quadratic kernel association scores, which are quadratic forms of the response vector which can measure nonlinear association between a group of features and the response. Here, we extend kernelPSI to the demanding setting of GWAS, where the kernel association scores can model nonlinear effects and epistatic interactions among neighboring SNPs. A number of putative loci are selected in the first step according to a selected kernel association score, and their aggregated phenotypic effect is tested in the second step.

The extension of kernelPSI to GWAS required several modifications to improve scalability. Most importantly, we developed a GPU-version of the constrained sampling algorithm to speed up linear algebra operations. The rest of the code was also accelerated thanks to a more efficient C++ backend. In particular, we implemented a rapid estimator of the HSIC criterion (Gretton et al., 2005) based on quadratic-time rank-1 matrix multiplications. HSIC is an example of quadratic kernel association scores (Slim et al., 2019). To illustrate this extension of kernelPSI on real GWAS datasets, we study BMI and its fluctuations (ΔBMI) in the UK BioBank (Bycroft et al., 2018).

We propose an eponymous R package that implements the full pipeline of kernelPSI. The CPU-only version is directly available from CRAN. The enhanced GPU-version can be downloaded from the development branch of the GitHub repository https://github.com/EpiSlim/kernelPSI.git.

2. KernelPSI: post-selection inference for big genomic data

Before covering the modifications we implemented to extend kernelPSI to GWAS data, we start with a brief overview of the framework in the context of GWAS. For further details, we refer the reader to Slim et al. (2019).

We model a GWAS dataset as a set of n pairs $\{(x_1, y_1), \cdots, (x_n, y_n)\}$. For each sample $i \in [\![1, n]\!]$, $y_i \in \mathbb{R}$ represents the phenotype and $x_i \in \mathcal{X}^p$ the geno-

[1]SANOFI R&D, Translational Sciences, Chilly Mazarin, 91385 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. Correspondence to: Lotfi Slim <lotfi.slim@mines-paristech.fr>.

type, with p the number of SNPs considered. In this study, we define x_i as a set of p SNPs mapped to a gene, and $\mathcal{X} = \{0, 1, 2\}$ following the dosage encoding of SNPs. We denote by $Y \in \mathbb{R}^n$ the vector of phenotypes, where $Y_i = y_i$ for $i \in [\![1, n]\!]$. We further consider a partition of the genotype in a set of S contiguous SNP clusters $\{\mathcal{S}_1, \cdots, \mathcal{S}_S\}$ (see Section 2.2). For each $t \in [\![1, S]\!]$, we define a kernel $\mathcal{K}_t : \{0, 1, 2\}^{|\mathcal{S}_t|} \times \{0, 1, 2\}^{|\mathcal{S}_t|} \to \mathbb{R}$ and the corresponding Gram matrix K_t (see Section 2.3 for examples of such kernels). For any $i, j \in [\![1, n]\!]$, $[K_t]_{ij} = \mathcal{K}_t(x_{i,\mathcal{S}_t}, x_{j,\mathcal{S}_t})$, where $x_{i,\mathcal{S}_t}$ contains the values of the SNPs in $\mathcal{S}_t$ for sample i, that is to say, x_i restricted to its entries in $\mathcal{S}_t$.

The goal is to select the SNP clusters, that is the kernels within $\{\mathcal{K}_1, \cdots, \mathcal{K}_S\}$, most associated with the phenotype, and then, to measure their overall association with the phenotype Y. In other words, we perform model selection and measure afterwards the significance of the constructed model.

In both selection and inference stages, a measure of association between a kernel K and a phenotype Y is needed. For this purpose, we define *quadratic kernel association scores* which are quadratic forms in Y:

$$\begin{aligned} s \colon \mathbb{R}^{n\times n} \times \mathbb{R}^n &\to \mathbb{R} \\ (K, Y) &\mapsto Y^\top Q(K) Y, \end{aligned} \tag{1}$$

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

Quadratic kernel association scores encompass a wide gamut of scores. For instance, empirical estimators of the HSIC criterion. In the present paper, we restrict ourselves to the unbiased empirical HSIC estimator, first proposed by Song et al. (2007):

$$\begin{aligned} \widehat{\mathrm{HSIC}}_{\text{unbiased}}(X, Y) = \frac{1}{n(n-3)} \Bigg[&\operatorname{trace}(\underline{K}\,\underline{L}) \\ &+ \frac{1_n^\top \underline{K} 1_n \; 1_n^\top \underline{L} 1_n}{(n-1)(n-2)} - \frac{2}{n-2} 1_n^\top \underline{K}\,\underline{L} 1_n \Bigg], \end{aligned} \tag{2}$$

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

A multitude of kernel selection strategies can be deployed. The kernels can be selected in a *forward* or *backward* stepwise fashion. The number of selected kernels can be either fixed, or adaptively determined. Here, we opt for an adaptive forward strategy, where the number of selected kernels S' is determined according to the maximum of $\widehat{\mathrm{HSIC}}_{\text{unbiased}}$ attained by iteratively adding the kernels.

Regardless of the kernel selection strategy, the selection of a subset of kernels $M \subseteq \mathcal{K}$ can be modeled as a conjunction of quadratic constraints: there exists $i_M \in \mathbb{N}$, and $(Q_{M,1}, b_{M,1}), \cdots, (Q_{M,i_M}, b_{M,i_M}) \in \mathbb{R}^{n\times n} \times \mathbb{R}$ such that

$$\{Y \,:\, \widehat{M}(Y) = M\} = \bigcap_{i=1}^{i_M} \{Y \,:\, Y^\top Q_{M,i} Y + b_{M,i} \geq 0\}. \tag{3}$$

For valid inference, we need to correct for the fact that the kernels were selected on the basis of their strong association with the outcome Y. As determining the exact distribution of $\mathrm{HSIC}_{\text{unbiased}}$ conditionally to the event $\{Y \,:\, \widehat{M}(Y) = M\}$ was impossible, we developed instead an efficient sampling algorithm to derive empirical p-values. Replicates of the outcome Y which satisfy the quadratic constraints in (3) are sampled. The values of their test statistics (in this case, $\widehat{\mathrm{HSIC}}_{\text{unbiased}}$) are then compared to the value of the statistic of the original outcome Y to obtain the desired p-values.

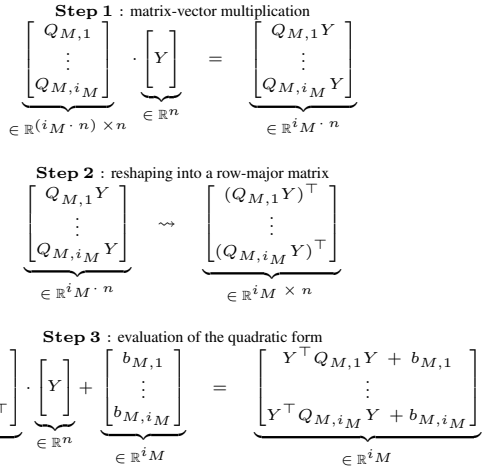

Figure 1. A GPU-accelerated pipeline for the evaluation of quadratic constraints.

2.1. Outcome normalization

Our original proposal in Slim et al. (2019) is limited to normally-distributed outcomes. To expand kernelPSI to other continuous outcomes, it suffices to transform any continuous outcome Y into a vector of independent normally-distributed variables. A well-known transformation is the Van der Waerden (1952) quantile transformation given by:

$$g(y) = F_{0,1}^{-1}\left(\frac{\operatorname{rank}(y) - 1/2}{n+1}\right), \tag{4}$$

where $y \in \mathbb{R}$, $\operatorname{rank}(y)$ is the ranking of y in a descending order with respect to $y_1, \cdots, y_n$, and $F_{0,1}$ is the c.d.f of the standard normal distribution.

2.2. Contiguous hierarchical clustering for genomic regions

In GWAS, the true causal SNPs are often unmeasured, but exhibit a strong linkage disequilibrium (LD) with the lead SNPS. The classical strategy to approach this problem is fine mapping (Schaid et al., 2018), where we study the genomic region surrounding the lead SNPs to identify the causal SNPs. A better strategy would then be to directly select regions of strong LD patterns. This amounts to selecting clusters of strongly-correlated SNPs. Such a strategy also has the advantage of reducing the number of clusters/kernels to choose from, while simultaneously modeling the combined cluster effects on the outcome.

To define these clusters, we apply adjacent hierarchical clustering (AHC). Following AHC, the optimal number of clusters S is estimated using the gap statistic. This approach is readily available from the R package BALD (Dehman et al., 2015).

2.3. The IBS-kernels and nonlinear SNP selection

It is obviously possible to use linear kernels to define $\{\mathcal{K}_1, \cdots, \mathcal{K}_S\}$. However, such a representation does not take into account minor allele frequencies (MAFs) nor epistatic interactions between SNPs. To address this limitation, Wu et al. (2010) proposed identical-by-state (IBS) kernels, which measure the number of identical alleles between two individuals i and j. For a cluster t and two genotypes x_i, x_j, IBS kernels are given by:

$$\mathcal{K}_t(x_{i,\mathcal{S}_t}, x_{j,\mathcal{S}_t}) = \sum_{q=1}^{|\mathcal{S}_t|} w_q (2 - \left| [x_{i,\mathcal{S}_t}]_q - [x_{j,\mathcal{S}_t}]_q \right|), \quad (5)$$

where the weights $(w_q)_{1:|\mathcal{S}_t|}$ are a function of their respective MAFs $(m_q)_{1:|\mathcal{S}_t|}$:

$$\sqrt{w_q} = \mathrm{Beta}\left(m_q, \alpha_q, \beta_q\right), \quad (6)$$

where Beta is the density function of the Beta distribution.

The parameterization $(\alpha_q, \beta_q)_{1:|\mathcal{S}_t|}$ is chosen according to the scope of the GWAS study. For common variants, Ionita-Laza et al. (2013) recommend setting $(\alpha_q, \beta_q) = (0.5, 0.5)$.

2.4. Efficient nonlinear post-selection inference for high-dimensional data

In this section, we detail a number of modifications we included in order to improve the scalability of kernelPSI to the large sample sizes.

2.4.1. Rapid estimation of the HSIC criterion

We first recall the unbiased HSIC estimator in Equation (2):

$$\widehat{\mathrm{HSIC}}_{\mathrm{unbiased}}(X, Y) = \frac{1}{n(n-3)} \Bigg[\mathrm{trace}(\underline{K}\,\underline{L}) + \frac{1_n^\top \underline{K} 1_n \; 1_n^\top \underline{L} 1_n}{(n-1)(n-2)} - \frac{2}{n-2} 1_n^\top \underline{K}\,\underline{L} 1_n \Bigg]. \quad (7)$$

The computation of $1_n^\top \underline{K} 1_n$ and $1_n^\top \underline{L} 1_n$ can be performed in quadratic time $\mathcal{O}(n^2)$. As for $\mathrm{trace}(\underline{K}\,\underline{L})$ and $1_n^\top \underline{K}\,\underline{L} 1_n$, a $\mathcal{O}(n^3)$ complexity can ensue because of the matrix-matrix multiplication of $\underline{K}$ and $\underline{L}$. To avoid that, we decompose $\mathrm{trace}(\underline{K}\,\underline{L})$ as $\sum_{i,j=1}^n [\underline{K}]_{ij} [\underline{L}]_{ji}$, which results in a better $\mathcal{O}(n^2)$ complexity. The same complexity can be achieved for the quadratic form $1_n^\top \underline{K}\,\underline{L} 1_n$ by starting with the matrix-vector multiplication of either $\underline{K} 1_n$ or $\underline{L} 1_n$. Overall, we achieve a $\mathcal{O}(n^2)$ complexity, for which the HSIC criterion can be computed on a single CPU for thousands of samples in relatively little time. As an illustration, we performed 100 repetitive evaluations of the HSIC criterion for two matrices of size $5,000 \times 5,000$. On a 2.7 GHz intel core i5 processor, the average running time was 1.08s.

2.4.2. Accelerated replicates sampling

The gains achieved in Section 2.4.1 turned out to be insufficient because of the heavy computational workload involved in replicates' sampling. Our sampling algorithm in Slim et al. (2019) is partly a rejection sampling algorithm. At every iteration, we verify that the candidate replicate satisfies the constraints $Y : Y^\top Q_{M,i} Y + b_{M,i} \geq 0$ for $i \in [\![1, i_M]\!]$. For a large i_M, we observed a significant slow-down due to the overhead between the successive evaluations of the constraints. A single combined evaluation would then eliminate this overhead. We achieve this by encoding all computations in a matrix form, as illustrated in Figure 1.

For linear algebra operations, GPUs can dramatically speed-up computations (Krüger & Westermann, 2003). We used them here to accelerate the multiplications detailed in Figure 1. A major drawback in hybrid CPU-GPU calculations is the transfer time between the main memory and the GPU memory. To alleviate this problem, we transfer the matrices in $\mathcal{Q}_M$ to GPU memory once and for all before the sampling.

3. A study of BMI and its variation in the UK BioBank

The study of physiological phenotypes in GWAS has so far focused on basic anthropometric measures such as height, weight, and BMI. Their longitudinal fluctuations received

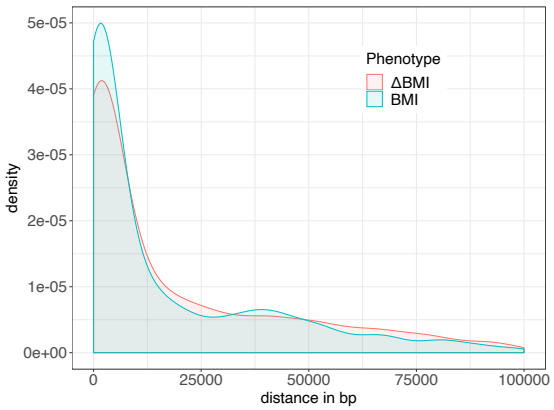

Figure 2. Distance between the SNPs of the GWAS Catalog and their closest neighbor among the SNPs in the clusters selected by kernelPSI.

little attention, mainly because of the lack of such data. To the best of our knowledge, the fluctuations of BMI have not been the subject of any specific GWA study. In fact, some studies (Sandholt et al., 2013) suggested that BMI and ΔBMI might be influenced by distinct sets of SNPs. We apply kernelPSI on the UK BioBank dataset (Bycroft et al., 2018) to separately study BMI and variations of BMI (ΔBMI).

To facilitate interpretation, we restricted ourselves to the genes already associated with BMI in the GWAS catalog (MacArthur et al., 2016). The scope of the narrower study is then gene prioritization. This is particularly interesting given the large number of genes associated with BMI (1811 genes).

A major strength of kernelPSI is its dual SNP-gene perspective. It presents the unique benefit of jointly performing SNP-level selection and gene-level significance testing. In this section, we evaluate the performance of kernelPSI in both steps.

3.1. Kernel selection

Because of the lack of a background truth for all genes, validating the results of statistical tools in GWAS has always been difficult. For our study, the validation task is relatively easier, though potentially biased. The genes were retrieved on the basis of their SNP-level association to BMI in the GWAS catalog. We can then compare the distance between the significant SNPs in each gene to their closest SNP neighbor in the clusters selected by kernelPSI. We provide in Figure 2 a histogram for the latter distances. The histogram is heavily skewed toward small distances. In other words, the GWAS catalog SNPs are often located near SNPs selected by kernelPSI. This confirms the capacity of kernelPSI to retrieve relevant genomic regions. Moreover, the selected clusters also surround significant SNPs. For BMI and ΔBMI, the selected clusters respectively included at least one significant SNP in 62.5% and 40.6% of genes.

3.2. Hypothesis testing

For association testing, we benchmark kernelPSI against two state-of-the-art gene-level baselines. The first one is SKAT (Wu et al., 2011), and can described as a non-selective variant of kernelPSI. Furthermore, it is a quadratic kernel association score which can be incorporated into the framework of kernelPSI. The second baseline is MAGMA (de Leeuw et al., 2015) which implements the principal components regression gene analysis model. More specifically, it implements an F-test in which the null hypothesis corresponds to absence of effects of all genotype PCs.

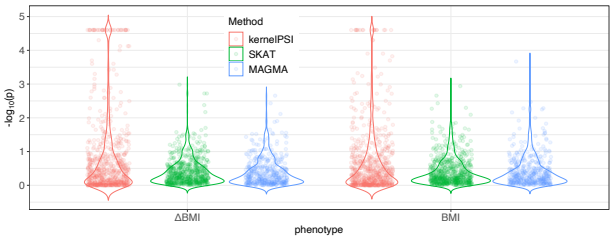

Figure 3. A violin plot comparing the p-values of kernelPSI for BMI and ΔBMI to two benchmarks.

To compute the empirical p-values in kernelPSI, we sampled $40,000$ replicates in addition to $10,000$ burn-in replicates. The comparison of the distributions of the resulting p-values to those of SKAT and MAGMA shows that kernelPSI clearly enjoys more statistical power than the two baselines for both phenotypes (Figure 3). The p-values were altogether significantly lower. Thanks to the large number of replicates, we attribute this performance, not to the lack of accuracy of the empirical p-values, but to the discarding of non-causal clusters in the selection stage.

4. Conclusion

Most GWAS restricted themselves to SNP-level association testing. In the present work, we presented a tool that still enables SNP selection, but ascends to the gene-level to perform association testing. The combination of the SNP and gene levels was possible through the use of post-selection inference which properly accounts for the SNP selection bias to perform valid gene inference. A major novelty in our work is the use of kernel methods which can model nonlinear effects and interactions among SNPs. The broad GWAS community can benefit from tools like kernelPSI which combine statistical performance with interpretability.

References

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T. et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018. doi: 10.1038/s41586-018-0579-z.

de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLOS Computational Biology*, 11(4):e1004219, April 2015. doi: 10.1371/journal.pcbi.1004219.

Dehman, A., Ambroise, C. and Neuvial, P. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16(1), May 2015. doi: 10.1186/s12859-015-0556-6.

Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *Lecture Notes in Computer Science*, pp. 63–77. Springer Berlin Heidelberg, 2005. doi: 10.1007/11564089_7.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D. and Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, 92(6):841–853, June 2013. doi: 10.1016/j.ajhg.2013.04.015.

Krüger, J. and Westermann, R. Linear algebra operators for GPU implementation of numerical algorithms. *ACM Transactions on Graphics*, 22(3):908, July 2003. doi: 10.1145/882262.882363.

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

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P. et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research*, 45(D1):D896–D901, November 2016. doi: 10.1093/nar/gkw1133.

Sandholt, C.H., Allin, K.H., Toft, U., Borglykke, A., Ribel-Madsen, R. et al. The effect of GWAS identified BMI loci on changes in body weight among middle-aged danes during a five-year period. *Obesity*, 22(3):901–908, August 2013. doi: 10.1002/oby.20540.

Schaid, D.J., Chen, W. and Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, May 2018. doi: 10.1038/s41576-018-0016-z.

Slim, L., Chatelain, C., Azencott, C.A. and Vert, J.P. kernelPSI: a post-selection inference framework for nonlinear variable selection. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5857–5865, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

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.

Van der Waerden, B. Order tests for the two-sample problem and their power. In *Indagationes Mathematicae (Proceedings)*, volume 55, pp. 453–458. Elsevier, 1952.

Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J. and Lin, X. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, June 2010. doi: 10.1016/j.ajhg.2010.05.002.

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, July 2011. doi: 10.1016/j.ajhg.2011.05.029.