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, Slim et al. (2019) published the theoretical foundations of kernelPSI, a post-selection inference (PSI) framework for nonlinear variable selection. It develops quadratic kernel association scores. They 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 effect on the phenotype is tested in the second step. The selection bias introduced by the first step is accounted for by sampling constrained replicates of the response vector. The statistics of the response are compared to those of the replicates to obtain valid p-values.

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). This extension also generalizes kernelPSI to any non-normally distributed continuous phenotypic outcome.

To illustrate the use of kernelPSI on real GWAS datasets, we study BMI and its fluctuations (ΔBMI) in the UK BioBank. The UK BioBank (Bycroft et al., 2018) is one of the largest available sources of data for the investigation of the contribution of genetic predisposition to a variety of physiological and disease phenotypes. We study both BMI and ΔBMI because of the suspicion that different genetic mechanisms might be governing the two phenotypes (Sandholt et al., 2013). Our study yielded a number of putative genes for BMI and ΔBMI, along with a list of causal *loci* within. Our use case has also showed the better statistical performance of kernelPSI in comparison to other gene-level association tools, with the unique benefit of pinpointing the causal *loci*.

We propose an eponymous R package that implements the

[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>.

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 genotype, with p the number of SNPs considered. In this study, we define x_i as a set of p SNPs mapped to a gene (see Section 3.1.3), 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):

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

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

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

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

Testing the association between the kernels in M and Y needs to account for the statistical bias introduced by the selection event. 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.

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:

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

where $y \in \mathbb{R}$, $\mathrm{rank}(y)$ 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.

The accuracy of the transformation in Equation (4) depends on the accuracy of the estimation of the regularized quantile $(\text{rank}(y) - 1/2)/(n+1)$, and thus on the number of participants n. Thankfully, many recent GWAS, in particular for physiological measurements, boast a large number of participants.

Other outcome normalization methods have been proposed, such as the Lambert $W \times F$ (Goerg, 2011), or Box-Cox (Box & Cox, 1964) and Yeo-Johnson (Yeo, 2000) transformations. In practice, we found the Van der Waerden transformation in Equation (4) to be the most consistent approach across different types of outcome distributions. All the above transformations are implemented in the R package bestNormalize (Peterson & Cavanaugh, 2019).

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. More statistical power is to be expected.

To define these clusters, we use the R package BALD (Dehman et al., 2015) which implements adjacent hierarchical clustering (AHC) in conjunction with the gap statistic (Tibshirani et al., 2001). Following AHC, the optimal number of clusters S is estimated using the gap statistic.

Here, we apply adjacent hierarchical clustering coupled with the gap statistic to split genomic regions into contiguous groups of SNPs. This approach is illustrated in Figure 1, and 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 MAFs and 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

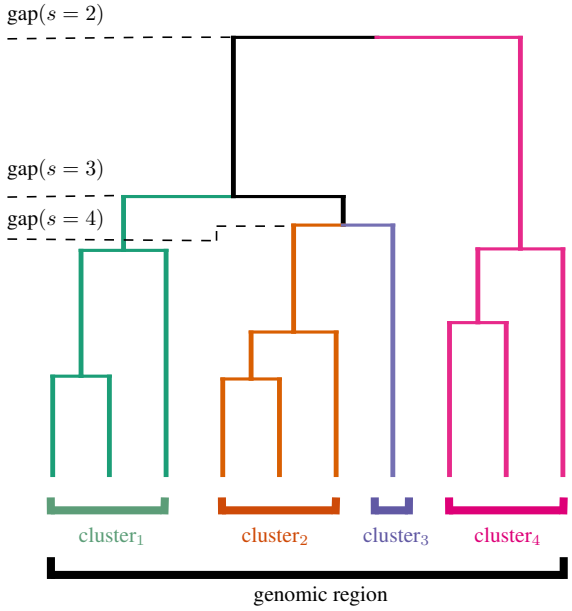

Figure 1. Clustering methodology: adjacent hierarchical clustering coupled with the gap statistic to determine the appropriate number of clusters.

by:

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

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

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

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 $(\alpha_q, \beta_q) = (0.5, 0.5)$. Such a parameterization still assigns higher weights to rare variants, but the difference is more moderate. For instance, for $(\alpha_q, \beta_q) = (0.5, 0.5)$, we have $w_q^2 = 0.63$ for $m_q = .5$ and for $m_q = 0.01$, $w_q^2 = 3.2$. To get a better grasp of these choices, we compare in Figure 2 the Beta densities for different values of (α_q, β_q).

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.

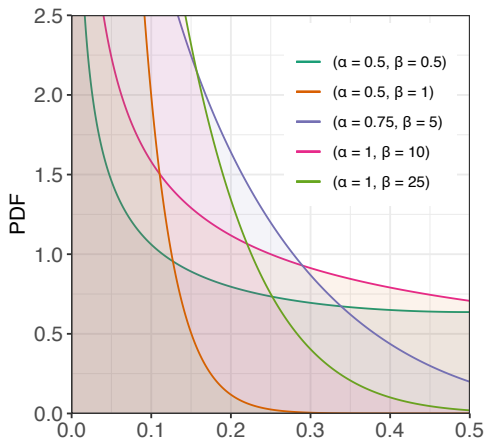

Figure 2. Comparison of the Beta densities for different values of the shape parameters (α, β)

.

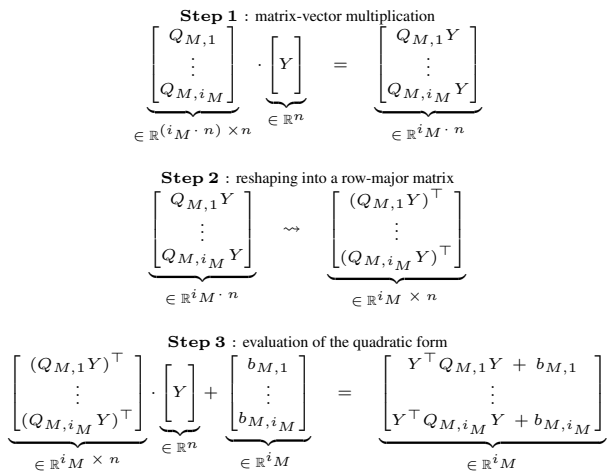

Figure 3. A GPU-accelerated pipeline for the evaluation of quadratic constraints.

2.4.1. Rapid estimation of the HSIC criterion

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

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

The computation of $1_n^\top \underline{K} 1_n$ and $1_n^\top \underline{L} 1_n$ can be performed in quadratic time $\mathcal{O}(n^2)$. However, 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 3.

For linear algebra operations, GPUs can dramatically speed-up computations (Krüger & Westermann, 2003). We used them here to accelerate the multiplications detailed in Figure 3. More specifically, we used the ViennaCL library (Rupp et al., 2016) which enables a simple, high-level access to GPU resources.

A major drawback in hybrid CPU-GPU calculations is the transfer time between the main memory and the GPU memory. With most Nvidia GPUs, the theoretical bandwidth limit is 8 Gb/s. For $i_M = 12$ and $n = 10^4$, the approximate memory size of all $\mathcal{Q}_M = \{Q_{M,1}, \cdots, Q_{M,i_M}\}$ matrices is 8.94 Gb in a double representation. If we sample $5 \cdot 10^4$ replicates, a repeated data transfer of $\mathcal{Q}_M$ would in best cases last 15 hours and 30 minutes. Such transfer times are prohibitive. To circumvent this problem, we transfer the matrices in $\mathcal{Q}_M$ to GPU memory once and for all before the sampling. However, because of memory size limitations, this imposes an upper limit on the number of matrices i_M in $\mathcal{Q}_M$, and consequently on the number of clusters $S = i_M/2 + 1$.

Finally, we give a rough estimation of the complexity of our sampling algorithm. If we denote by $N_{\text{replicates}}$ the number of replicates, the overall complexity can be approximated as $\mathcal{O}\big(N_{\text{replicates}} i_M n^2/\tau(n)\big)$. $\tau(n)$ is a decreasing function of n which corresponds to the probability of sampling a

replicate in the acceptance region. The average number of iterations to obtain a valid replicate is then $1/\tau(n)$ (mean of a geometric distribution). We are currently unable to propose a closed form for $\tau(n)$.

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 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. Only rare variants impacting weight loss through gene-diet interaction are referenced in the literature (Qi, 2014). Recent biobanks such as HUNT (m.fl, 2011), ALSPAC (Fraser et al., 2012) and BiB (Raynor & Group, 2008) are finally making such data available. Another notable biobank is the UK BioBank (Bycroft et al., 2018) which provides extensive phenotypic and health-related information for over $500,000$ British participants. We apply kernelPSI on the UK BioBank dataset to separately study BMI and variations of BMI (ΔBMI).

3.1. Data and experiments

3.1.1. Quality control

Preprocessing in any GWA study is a manadatory step. Our preprocessing pipeline for the UK BioBank dataset is closely similar to the pipeline of the Neale lab [1] who provides exhaustive summary statistics for over $2,000$ phenotypes in the UK BioBank. We detail below the sample quality-control we conducted.

- Heterozygosity and missing rates: Discarding outliers for both criteria.
- Sex chromosome aneuploidy: only individuals with sex chromosome configurations XX or XY are retained.
- Prior use in phasing: whether sample was selected as input for the phasing of autosomal chromosomes
- Kinship to other participants: we only select participants with no identified relatives in the dataset.
- Ethnic grouping: we subset samples identified as 'white British' to avoid any potential population structure effects.
- Prior use in principal components analysis (PCA): we discard all samples not included in the PCA. The analysis is used for population stratification (see next step).
- Homogeneity: additional population structure artifacts are detected used genomic dispersion (GD). We approximated it through the normalized squared distance of the first six principal components (PCs):

$$\mathrm{GD}(X) = \frac{1}{6}\sum_{i=1}^{6}\langle \mathrm{PC}_i, X\rangle^2$$

The application of the above pipeline yielded $n = 266,679$ final samples. In contrast, Neale lab obtained $337,000$ samples by implementing less stringent thresholds.

We directly extracted the SNPs of the UK BioBank Axiom array from the imputed genotypes provided by the UK BioBank consortium. As for SNP quality control, we focused on bi-allelic SNPs located on autosomal chromosomes. Moreover, We filtered out the SNPs with a MAF < 0.01 or not in a Hardy-Weinberg equilibrium ($p < 1\mathrm{e}{-10}$). Out of caution, we also incorporated two additional filters: we shed the SNPs with an internal UK BioBank information score < 0.8 and a missing proportion rate $> 1/n$. The QC pipeline resulted in $577,811$ SNPs.

3.1.2. Phenotypes

ΔBMI is not directly available. We computed it from the participants who attended both the initial assessment visit and the first repeat assessment visit. In total, we obtained $11,992$ samples for ΔBMI. More precisely, we use the average yearly variations $\Delta\mathrm{BMI}/\Delta t_{\text{years}}$. The reason for this is that the time span between the visits is not the same for all participants. For simplicity of notation, we use ΔBMI to denote yearly variations in the rest of this paper.

As for BMI, we use the measurements of the initial visit. The joint analysis of results stemming from datasets with different samples is not straightforward and requires utmost attention. We mention two-sample problems as an example of tools providing statistically-principled methods to tackle this problem. Our case is even more delicate: BMI and ΔBMI share a number of samples with a huge discrepancy in the total number of samples. To avoid this issue, we restrict ourselves in both phenotypes to the samples for which the ΔBMI measurement is available.

As explained in Section 2.1, we apply the Van der Waerden transformation. We illustrate the accuracy of the transformation in Figure 4, by visually comparing the empirical c.d.fs of BMI and ΔBMI to the c.d.f of a standard normal distribution. We notice a complete overlap between the c.d.fs. We attribute this good performance to the total number of samples and the low number of ties ($11,933$ unique values).

[1] More details are provided on their website https://www.nealelab.is/uk-biobank

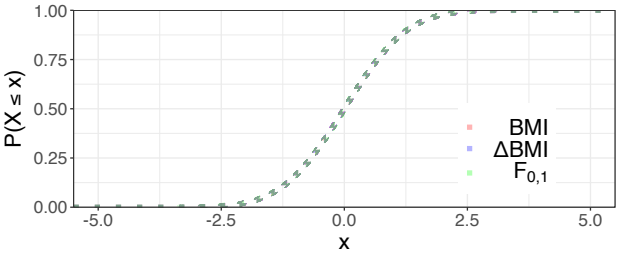

Figure 4. Comparison of the empirical c.d.fs of BMI and ΔBMI to the c.d.f of a standard normal distribution.

3.1.3. Gene selection

Because of the computation time and resource requirements, a full genome-wide study is impossible. We instead 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).

Genes are included in the GWAS catalog, if they at least contain one significant SNP, which can possibly result in a high number of false positive genes. A major strength of kernelPSI is its dual SNP-gene perspective. The gene-level association testing in kernelPSI can assess whether the SNP-level association translates into a gene-level association.

To define genic boundaries, we used the biomaRt tool (Durinck et al., 2009), which provided a genomic interval for 1774 genes. Moreover, the intervals were converted from the GRCh38 coordinate system to the GRCh37 one, since the SNP positions in the UK BioBank are given in the GRCh37 system. We point out that the conversion can result in several noncontiguous intervals (see Hinrichs (2006); Lawrence et al. (2009) for further explanation).

An immediate use of the resulting intervals led to a number of genes without any SNPs within. As a result, we added a downstream/upstream 50kb buffer to cover more SNPs. The same buffer size was also opted for by several other authors (Nakka et al., 2016; Shah et al., 2018).

3.1.4. Hierarchical clustering

Despite the extended 50kb buffer, several genes still contained a handful of SNPs. 1215 genes contained at most 3 SNPs. In particular, if only one SNP only is mapped to a given gene, the kernel selection step becomes irrelevant. Nonetheless, we still perform hypothesis testing by directly using the HSIC statistic in Eq. (2) to measure the association between the gene and the phenotype. If 2 or 3 SNPs mapped to a gene, we associate a distinct cluster/kernel to each one of them. This allows for a more accurate SNP selection without entailing a dramatic increase in computational complexity.

For all other genes (with more than 4 SNPs), we applied AHC, as explained in Section 2.2. The optimal number of clusters S is determined by the gap statistic. In the adaptive kernel selection strategy we use here, this leads to $i_M = 2\,(S-1)$ constraints. To avoid the issues encountered for a large i_M (see Section 2.4.2), we set the maximum number of clusters to 5. This leads to a ~ 9.2Gb maximum GPU memory occupancy for the matrices $\mathcal{Q}_M$.

3.2. Results

KernelPSI 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.

3.2.1. 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 5 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.

If kernelPSI turned out to be selecting all clusters, the above results would be irrelevant. The clusters would always contain significant SNPs, and a few selected SNPs would also be located near the significant ones. In our application, kernelPSI conservatively selected the number of associated clusters S' (see Table 1). For BMI, kernelPSI selected one cluster in $75,9\%$ of the genes for which $S = 3$ and at most 2 clusters in $73,6\%$ of the genes for which $S = 5$. Similar results were obtained for ΔBMI.

Overall, the conservative kernel selection combined with the proximity of the selected kernels to the GWAS catalog SNPs demonstrate the selection performance of kernelPSI.

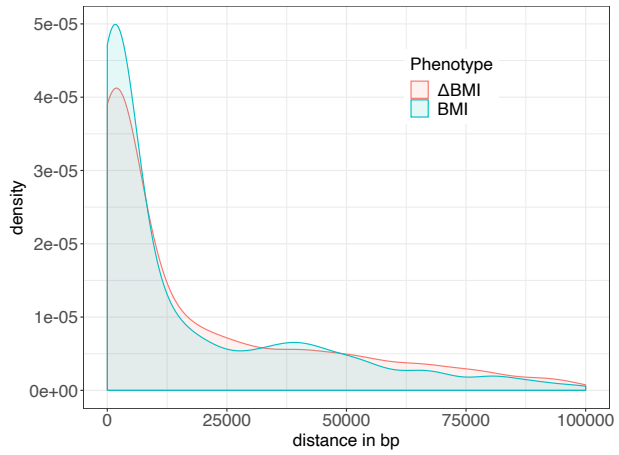

Figure 5. Distance between the SNPs of the GWAS Catalog and their closest neighbor among the SNPs in the clusters selected by kernelPSI.

Table 1. Distribution of the number of selected clusters S' depending on the total number of clusters S and the phenotype.

	BMI				
	1	2	3	4	5
1	100.0%				
2	82.9%	17.1%			
3	75.9%	18.5%	5.6%		
4	55.9%	30.9%	10.3%	2.9%	
5	43.9%	29.7%	17.6%	7.2%	1.5%
	ΔBMI				
	1	2	3	4	5
1	100.0%				
2	92.7%	7.3%			
3	70.4%	20.4%	9.3%		
4	50.0%	27.9%	19.1%	2.9%	
5	40.3%	28.4%	21.4%	7.6%	2.3%

3.2.2. 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 SKAT score is a variance-component score (Lin, 1997) given by $s_{\text{SKAT}}(K, Y) = Y^\top KY$, for a centered phenotype Y. 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.

A central hypothesis for our study is the different mechanisms involved in BMI and ΔBMI. The low rank correlations of the p -values between the two phenotypes (see Table 2) lend further credence to this hypothesis. Interestingly, we observed a similar range of values for kernelPSI and the two benchmarks SKAT and MAGMA. For all metrics and methods, the rank correlations are lower than 0.1.

Despite the low rank correlations between BMI and ΔBMI, we obtained 7 common significant genes [2] out of 64 significant genes for ΔBMI and 40 for BMI. The latter were determined after the application of the Benjamini-Hochberg procedure with an FDR threshold of 0.05. The existence of a number of separate mechanisms does not preclude the existence of common ones simultaneously regulating BMI and ΔBMI.

[2]The common genes are: CKB, EIF2S2, KSR2, MIR100HG, NRXN3, PDILT and RAB27B

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 6). 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. In the future, we look forward to developing exact variants of kernelPSI which forego the sampling step to directly determine the associated p-value. This can dramatically reduce computational times. Another area of major interest to us is the application of kernelPSI to other SNP sets such as cis-regulatory regions or whole pathways to obtain the significance of association of the pathway and the involved genes.

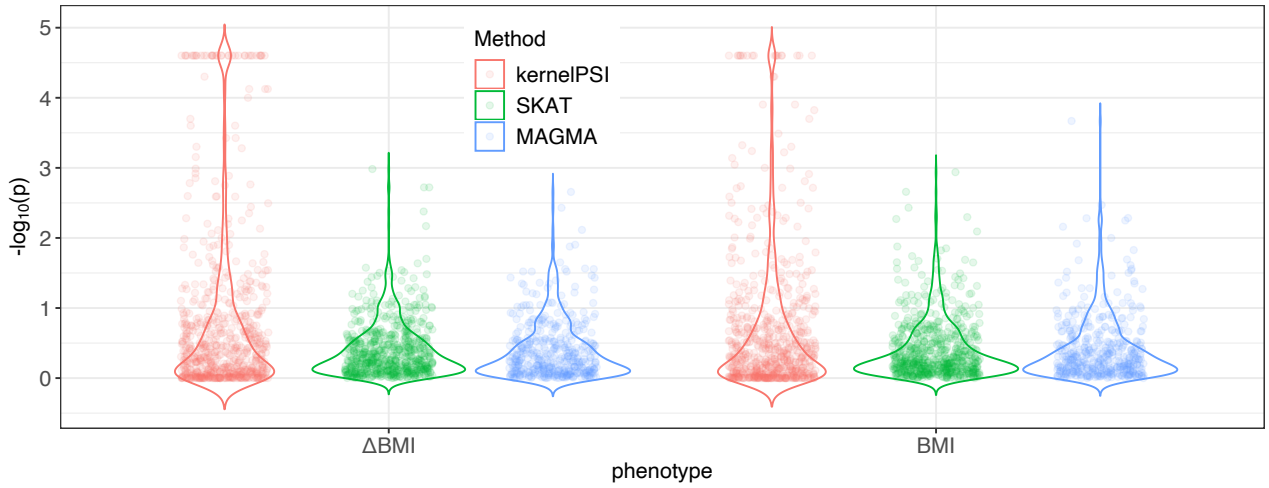

Figure 6. A violin plot comparing the p-values of kernelPSI for BMI and ΔBMI to two benchmarks.

Table 2. Concordance between BMI and ΔBMI by method, according to three Kendall rank correlation measures (standard, multiplicative, additive).

	Standard		Multiplicative		Additive	
	BMI	ΔBMI	BMI	ΔBMI	BMI	ΔBMI
kernelPSI	1.000	0.015	1.000	0.093	1.000	0.072
	0.015	1.000	0.093	1.000	0.072	1.000
SKAT	1.000	0.020	1.000	0.008	1.000	0.028
	0.020	1.000	0.008	1.000	0.028	1.000
MAGMA	1.000	0.036	1.000	0.058	1.000	0.083
	0.036	1.000	0.058	1.000	0.083	1.000

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number '48319'.

References

Box, G.E.P. and Cox, D.R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, July 1964. doi: 10.1111/j.2517-6161.1964.tb00553.x.

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.

Durinck, S., Spellman, P.T., Birney, E. and Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4:1184–1191, 2009.

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J. et al. Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *International Journal of Epidemiology*, 42(1):97–110, April 2012. doi: 10.1093/ije/dys066.

Goerg, G.M. Lambert w random variables—a new family of generalized skewed distributions with applications to risk estimation. *The Annals of Applied Statistics*, 5(3): 2197–2230, September 2011. doi: 10.1214/11-aoas457.

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.

Hinrichs, A.S. The UCSC genome browser database: update 2006. *Nucleic Acids Research*, 34(90001):D590–D598, January 2006. doi: 10.1093/nar/gkj144.

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.

Lawrence, M., Gentleman, R. and Carey, V. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25:1841–1842, 2009. doi: 10.1093/bioinformatics/btp328.

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.

Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, June 1997. doi: 10.1093/biomet/84.2.309.

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.

m.fl, J.H. The nord-trøndelag health study 1995-97 (HUNT 2). *Norsk Epidemiologi*, 13(1), October 2011. doi: 10.5324/nje.v13i1.305.

Nakka, P., Raphael, B.J. and Ramachandran, S. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics*, 204 (2):783–798, August 2016. doi: 10.1534/genetics.116.188391.

Peterson, R.A. and Cavanaugh, J.E. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, pp. 1–16, 2019. doi: 10.1080/02664763.2019.1630372.

Qi, L. Gene–diet interaction and weight loss. *Current Opinion in Lipidology*, 25(1):27–34, February 2014. doi: 10.1097/mol.0000000000000037.

Raynor, P. and Group, B.i.B.C. Born in Bradford, a cohort study of babies born in Bradford, and their parents: Protocol for the recruitment phase. *BMC Public Health*, 8(1):327, 2008. ISSN 1471-2458. doi: 10.1186/1471-2458-8-327.

Rupp, K., Tillet, P., Rudolf, F., Weinbub, J., Morhammer, A., Grasser, T., Jüngel, A. and Selberherr, S. ViennaCL—linear algebra library for multi- and many-core architectures. *SIAM Journal on Scientific Computing*, 38(5): S412–S439, January 2016. doi: 10.1137/15m1026419.

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.

Shah, R.L., Li, Q., Zhao, W., Tedja, M.S., Tideman, J.W.L. et al. A genome-wide association study of corneal astigmatism: The CREAM Consortium. *Molecular vision*, 24: 127–142, feb 2018.

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.

Tibshirani, R., Walther, G. and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, May 2001. doi: 10.1111/1467-9868.00293.

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.

Yeo, I.K. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, December 2000. doi: 10.1093/biomet/87.4.954.