A Data-Driven Framework for Optimal Covariate Clustering in Genome-wide Association Interaction Studies

Volker Neff[1 [0009−0001−4402−6523]], Lars Wienbrandt[1 [0000−0001−5685−2032]], and David Ellinghaus[1 [0000−0002−4332−6110]]

Institute of Clinical Molecular Biology (IKMB), Kiel University and University Medical Center Schleswig-Holstein, Rosalind-Franklin-Straße 12, 24105 Kiel, Germany.
{v.neff, l.wienbrandt, d.ellinghaus}@ikmb.uni-kiel.de

Abstract. Genome-wide epistasis detection is computationally extremely demanding and difficult to scale to modern genome-wide association study (GWAS) datasets. We recently introduced an efficient population structure-adjusted logistic regression algorithm for genome-wide association interaction studies (GWAIS) based on proxy-covariates derived from clustering principal component analysis (PCA) covariates (Neff et al., 2025). In this approach, we substantially reduced the computational burden while retaining genetic ancestry adjustment, but finding the optimal configuration for covariate clustering remained unsolved.
Here, we present a data-driven framework for automated testing of clustering configurations for covariate-adjusted epistasis screening. Starting from per-sample PCA coordinates treated as ground truth, we systematically evaluate K-means and Gaussian mixture model (GMM) clustering across candidate parameter settings, and evaluate these so-created proxy covariates on a reduced data subset. Graphical scatter plots help to highlight the best configuration minimizing the mean relative error (MRE) of interaction p-values compared to the ground truth.
Using the real-world GWAS datasets previously analyzed via genome-wide screening, we show that evaluating only 1 % to 5 % of the original data reliably reproduces or improves the previously identified configurations. In subsequent epistasis screening, our guided proxy-covariate adjustment achieves a speedup of up to 151x while maintaining a low MRE relative to the per-sample covariate-adjusted reference model.

Keywords: Epistasis detection · Covariate adjustment · K-means · Gaussian mixture models · Logistic regression

1 Introduction

Genome-wide association interaction studies (GWAIS) are a key statistical tool for detecting epistasis (statistical genetic interactions). In particular, pairwise SNP-SNP interactions play a significant role in the discovery of missing heritability in genetic diseases, but their detection suffers from a quadratic search space of

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

M^2, where M is the number of single-nucleotide polymorphisms (SNPs). Methods exist that try to reduce the immense number of SNP-pairs to be tested [9], but they risk overlooking significant result pairs. Therefore, an exhaustive search is essential to cover the complete scope, which implies the necessity to perform each individual interaction test as efficiently as possible. In the past, several tools emerged to perform an exhaustive screening using different metrics, whereas the most popular one is probably BOOST [16]. However, as logistic regression (as it is implemented in the *PLINK* toolkit [3]) is regarded as the gold-standard, SNP-SNP interaction detection becomes infeasible already for smaller datasets comprising several thousand samples and SNPs. The application of algorithmic transformations and hardware acceleration leads to a satisfactory remedy [18]. However, the plain logistic regression approach implemented in those tools does not account for population stratification, which disturbs epistasis scores significantly, especially in inhomogeneous datasets.

We recently successfully tackled this problem by introducing covariate clustering and using proxy-covariates instead of performing covariate correction for each sample individually, as shown by Neff et al. [13]. On a dataset with 141,621 genetic markers obtained from German individuals where 3,520 were diagnosed with inflammatory bowel disease (IBD) and 4,288 were healthy controls ([15]), our method reduced the mean relative error (MRE) when compared to a test run without covariate correction by 50.6% while the speedup was 70-fold compared to the ground truth (GT) computation that applied individual covariate correction for each sample. To find the optimal individual values, the same experiment had to be performed multiple times across different settings on the complete dataset, as selecting the optimal solution for the clustering method and parameters is not straightforward and impractical for large GWAIS datasets.

Therefore, we address this problem in this paper by introducing a new framework that quickly evaluates several clustering algorithms and parameters to determine accuracy and runtime predictions for a complete epistasis screening using cluster-derived proxy-covariates. We show that a small subset of the GWAS input dataset is sufficient to reliably predict the optimal configuration, and demonstrate that the framework is able to reproduce or improve our previous findings.

The remaining sections of this paper are organized as follows: Section 2 describes our new framework with its individual three-step pipeline, followed by a detailed explanation of the underlying logistic regression algorithm in Section 2.1 and the used K-means and Gaussian mixture model in Section 2.2 and 2.3. Section 3 describes the evaluation process, the datasets and experimental settings used, and the recorded results, which are finally concluded in Section 4.

2 Methods

Our new framework evaluates different clustering parameters and algorithms on a sub-dataset to determine the optimal proxy-covariates for a logistic regression, as shown in Equation 3 in subsection 2.1 below. The process is divided into three major steps:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

Preprocessing: The first step prepares a reduced sub-dataset from the input data. It randomly extracts a selectable percentage of the SNPs from the original dataset using *PLINK* [3]. To ensure reproducibility, a fixed seed can be set for the random number generator.

Experiments: In the second phase, multiple epistasis detection runs across several different configurations are performed on the reduced dataset. First, a ground truth (GT) is generated, which is regarded as the most accurate achievable using a logistic regression model with 10-dimensional per-sample covariate correction.

Further, a run without any covariate correction is conducted, which is expected to produce the least accurate results, followed by several experiments conducted with various proxy-covariates. We currently apply two clustering algorithms (either K-means or a Gaussian mixture model (GMM)) to cluster the available individual per-sample covariates. The centroids of each cluster are used as proxy-covariates to replace the individual covariate of each cluster member in the logistic regression. Both clustering methods are parameterized with a different value of K to generate a different number of clusters in each run. The range of cluster numbers (K) to be evaluated is 2 to 20 to cover a wide range of reasonable settings. From each experiment, a number of $S = 25,000$ top-scoring SNP pairs are stored together with their χ^2-score, odds-ratio (OR), and p-value. Further, the runtime for each test is recorded.

Evaluation: Once all experiments have been conducted on the reduced input dataset, the results are compared with respect to the ground truth (GT). First, a simple quality control is conducted on all results (including the GT results), i.e. improper values such as those with an odds ratio of zero or infinity ($OR = 0$ or $OR = \text{inf}$), are removed. Then, the results of the GT are compared to each single experiment by computing the mean relative error (MRE) of the p-values in logarithmic space (S is the number of the saved top results from the test, p_i is the p-value of the i-th top result and p_i^{GT} is the corresponding p-value of the same SNP-pair from the ground truth):

$$\text{MRE} = \frac{1}{S} \sum_{i=0}^{S-1} \frac{|\log_{10}(p_i^{\text{GT}}) - \log_{10}(p_i)|}{|\log_{10}(p_i^{\text{GT}})|} \tag{1}$$

We observed that, in general, several results from the GT run cannot be found in the top results of the test run. For these results, we used a best-case prediction, i.e. p_i is set to the p-value of the last (S-th) of the top reported SNP-pairs from the test run.

In addition to the prediction of the result accuracy, the process runtime for each experiment with the reduced input data is used to predict the runtime for a run with the complete dataset.

The evaluation of all results including graphical representations is presented in section 3.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

2.1 Logistic Regression with Contingency Tables and Proxy-Covariates

The additive two-locus logistic regression model (see Equation 2) is widely used to detect epistasis interaction statistically [4, 17]. Therefore, we assume a cohort of N study participants (samples), each assigned a binary trait (phenotype) Y, where $y_i = 1$ denotes a case (disease-positive patient) and $y_i = 0$ a control participant. A corresponding genotype vector $g_{i,A/B} \in G$ encodes the SNP-pair under testing with $G = \{0, 1, 2\}$ and the following encoding: (0) *homozygous reference*, (1) *heterozygous*, (2) *homozygous variant*. Finally, each sample is associated with a covariate vector $\overrightarrow{v_i} \in \mathbb{R}^C$, which are combined into $\mathcal{V} = \{\overrightarrow{v_0}, \ldots, \overrightarrow{v_{N-1}}\}$ as the set of covariate vectors from all samples (w.l.o.g. $\overrightarrow{v_i} \neq \overrightarrow{v_j}\ \forall i \neq j$).

$$\ln\left(\frac{P(Y=1 \mid X_A=g_A, X_B=g_B)}{P(Y=0 \mid X_A=g_A, X_B=g_B)}\right) = \beta_0 + \beta_1 g_A + \beta_2 g_B + \beta_3 g_A g_B + \sum_{l=0}^{|\overrightarrow{v}|-1} \beta_{4+l} v_l \quad (2)$$

This model can be fitted using a maximum-likelihood (ML) estimation over all N input samples in multiple iterations I. The interaction score (based on β_3), which indicates the degree of correlation between the two SNPs and the phenotype, follows a chi-squared distribution, allowing a p-value to be derived directly [18]. A commonly used implementation without covariate correction is available in the popular bioinformatics toolset *PLINK* [3]. The simplified runtime complexity for fitting this model is $\mathcal{O}(NI)$ for each pair of SNPs.

Wienbrandt et al. [18] showed how to reduce the runtime complexity for a scenario without covariate compensation to $\mathcal{O}(N + I)$ by using contingency tables. In [13] we described how to add support for covariates to this approach by using contingency tables in combination with proxy-covariates.

Each sample covariate vector $v_i \in \mathcal{V}$ is therefore substituted by one of K proxy-covariate vectors $\hat{v} \in \hat{\mathcal{V}}$ into $\hat{v}_i$. Based on this, two separate contingency tables are constructed, one for case samples and the other for control samples, for each proxy-covariate, resulting in $2K$ contingency tables. Each contingency table $\mathcal{N}_{\hat{v},3\times3} = (n_{\hat{v},j,k})_{3\times3}$ records the counts of genotype combinations across all samples assigned to the same proxy-covariate vector. The contingency tables are then used in a transformed version of Equation 2:

$$\hat{\mathbf{L}}_{\ln}(\beta) = \sum_{\hat{v}\in\hat{\mathcal{V}}} \sum_{g_A=0}^{2} \sum_{g_B=0}^{2} \left[n^{\text{case}}_{\hat{v},g_A,g_B} \ln(p_{\hat{v},g_A,g_B}) + n^{\text{ctrl}}_{\hat{v},g_A,g_B} \ln(1-p_{\hat{v},g_A,g_B})\right] \quad (3)$$

$$\text{with} \quad p_{v,g_A,g_B} = \frac{1}{1+e^{-z_{v,g_A,g_B}}}$$

$$\text{and} \quad z_{v,g_A,g_B} = \beta_0 + \beta_1 g_A + \beta_2 g_B + \beta_3 g_A g_B + \sum_{c=0}^{|\overrightarrow{v}|-1} \beta_{4+c} v_c$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

The new runtime complexity can be simplified with $\mathcal{O}(N + IK)$, which results in dramatically reduced analysis runtimes as long as the number of proxy-covariate vectors $K = |\hat{\mathcal{V}}|$ is substantially smaller than the number of samples N, i.e. $K \ll N$. For details, we refer to Neff et al. [13].

2.2 *K*-Means Clustering

K-means is a widely used clustering algorithm that minimizes the *objective function*, i.e., the clusters $\{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_{K-1}\}$ with their centroids $\bar{x} \in \mathcal{S}_k$ are selected such that the within-cluster sum of squared distances is minimized.

The K-means algorithm has been implemented and extended in various versions (e.g., k-means++ [10], and K-median [2]). We use an implementation provided by the *mlpack* library [5] with a naive Lloyd's algorithm [12] for the minimization process, a random initialization of the centroids $\bar{x}$, and an Euclidean distance $\|x\| = \sqrt{\sum_{x=0}^{|x|-1} {x_i}^2}$ for the distance measurement. The Lloyd's algorithm is an iterative minimization method for K-means clustering. After the prior initialization of the centroids, it iterates over two steps until the system converges or a maximum number of 1,000 iterations is reached:

1. All distances between the samples v and all cluster centroids $\bar{x}$ are calculated. The samples are then assigned to the cluster with the smallest distance to its centroid.
2. After all samples have been assigned, the centroids are updated as follows: $\bar{x}_k = \frac{1}{|\mathcal{S}_k|} \sum_{x_j \in \mathcal{S}_k} x_j$.

Finally, the latest centroids are used as proxy-covariates ($\hat{v}_k = \bar{x}_k$) in the logistic regression algorithm in Equation 3.

2.3 Gaussian Mixture Model

A Gaussian mixture model (GMM) is a clustering method based on the assumption that each sampled data point was drawn from one of K multivariate Gaussian (normal) distributions $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$ with unknown parameters (μ as distribution mean and variance matrix $\Sigma_{C\times C}$ with parameter dimension C):

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{C}{2}} \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)\right) \tag{4}$$

Unlike K-means, in which clusters cannot overlap, multiple Gaussians in a GMM are allowed to overlap. By chance, the data point x could be generated by multiple distributions. To keep this into account, a weight factor π_k for each Gaussian is used. The full GMM probability model is defined as [7]:

$$p(x) = \sum_{k=0}^{K-1} \pi_k \, \mathcal{N}(x \mid \mu_k, \Sigma_k) \quad \text{with} \quad 0 \leq \pi_k \leq 1 \quad \text{and} \quad \sum_{k=0}^{K-1} \pi_k = 1. \tag{5}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

For clustering, we use the *Scikit-learn* [14] Python library. It uses K-means clustering for initialization and the expectation-maximization (EM) algorithm [14] for optimization. EM iteratively maximizes the likelihood for Equation 5 in two steps, the *expectation step* and the *maximization step*, until the likelihood value converges [7].

Expectation step: Compute the *responsibility matrix* $R_{N\times K} = (r_{n,\,k})_{N\times K}$:

$$r_{n,\,k}^{(t)} = \frac{\pi_k^{(t)}\ \mathcal{N}(x_n \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=0}^{K-1} \pi_j^{(t)}\ \mathcal{N}(x_n \mid \mu_j^{(t)}, \Sigma_j^{(t)})} \tag{6}$$

Maximization step: The parameters π_k, μ_k, and Σ_k are estimated based on R:

$$\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{n=0}^{N-1} r_{n,\,k}^{(t)}\ x_n \tag{7}$$

$$\Sigma_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{n=0}^{N-1} r_{n,\,k}^{(t)}\ (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T \tag{8}$$

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{N^{(t)}} \tag{9}$$

with $N_k^{(t)} = \sum_{n=0}^{N-1} r_{n,\,k}^{(t)}$ and $N^{(t)} = \sum_{k=0}^{K-1} N_k^{(t)}$.

After the algorithm converges, R contains the normalized responsibilities of the different Gaussians for each sample and therefore a soft cluster assignment. To achieve a hard cluster assignment, the cluster with the highest responsibility score is selected for each sample. For each hard-assigned cluster, the centroid is calculated and used as proxy-covariate $\hat{v}$.

3 Evaluation and Results

To evaluate our framework, we conduct benchmarks on two real-world GWAS datasets: an IBD *case-control* dataset and a COVID-19 *case-control* dataset on a compute node equipped with two AMD Epyc 7313 processors with a total number 32 cores at 3.0 GHz and 128 GB RAM. Each dataset was carefully selected and passed a quality control (QC) pipeline implemented in *BIGwas* [11]. This includes, but is not limited to, a sample and SNP missingness test, duplicate and relatedness tests, and a population structure analysis via principal component analysis (PCA). To resolve within-Europe relationships between study samples, the remaining quality-filtered samples were again tested using a PCA, as implemented in *FlashPCA2* [1]. Inspection of the top 10 PCs revealed no outliers with non-European ancestry or signs of population stratification. This process was crucial for avoiding sample outliers in the datasets, which can disrupt clustering (see sections 2.2 and 2.3).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

The first GWAS dataset (*IBD*) comprises 7,787 German participants, including 3,506 patients with inflammatory bowel disease (IBD) and 4,281 German controls, and 583,241 SNPs after QC. It was further reduced to 141,621 SNPs for this analysis by applying a (minor allele frequency) MAF filter of 0.2. The dataset was first published in 2022 by Sazonovs et al. [15] who identified multiple new genetic risk factors for Crohn's disease.

The second GWAS dataset was collected during the COVID-19 pandemic in early 2020 and includes a broader range of ancestries. It was first published by Ellinghaus and Degenhardt, et al. [8] and later extended in [6]. After the application of the *BIGwas* pipeline (see above), the dataset contains 16,739 samples with 443,401 SNPs from five countries (Austria, Germany, Italy, Norway, and Spain), each with cases (hospitalized SARS-CoV-2 patients) and healthy controls (except the Austrian samples, which had only 28 cases and no controls). The detailed counts of cases and controls by country are shown in Table 1.

Our new framework aims to evaluate the optimal proxy-covariates for a logistic regression model as in Equation 3 and Neff et al. [13] with a small subset of the input data. Thus, it uses a fraction of available SNPs for regression models with different configurations to quickly predict the accuracy of the corresponding complete run. We test with different fractions of the available SNPs (1 %, 5 %, and 10 %) per dataset, complemented by a complete run (100 %), to analyze and compare the explanatory power of each of these settings. Each test of our framework conducts 40 experiments: one GT run with per-sample covariates, one without covariates, and two groups of test runs with proxy-covariates determined by different clustering methods. The first group uses K-means for proxy-covariate generation, whereas the second group uses Gaussian mixture models (GMMs). In each group, 19 cluster sizes ranging from 2 to 20 are tested. The covariates were chosen as the 10 most relevant PCs from a PCA generated with *FlashPCA2* [1] separately for each dataset. The accuracy of the individual settings is measured as mean relative error (MRE) relative to a ground truth (GT) that uses per-sample covariates based on the top 25,000 results. The MREs of each test are presented in Table 2.

Table 1. Number of study participants in the two GWAS datasets used for evaluation, stratified by nationality and case-control status after quality control. Dataset (a) comprises an IBD cohort from Germany [15] with 141,621 SNPs. Dataset (b) represents a multi-national COVID-19 cohort from Austria, Germany, Italy, Norway, and Spain [6, 8] with 443,401 SNPs.

(a) *IBD*

	Germany
Cases	3,506
Controls	4,281
$\sum$	7,787

(b) *COVID-19*

	Austria	Germany	Italy	Norway	Spain	$\sum$
Cases	28	306	1,543	81	2,150	4,108
Controls	0	3,243	4,735	280	4,373	12,631
$\sum$	28	3,549	6,278	361	6,523	16,739

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

Table 2. Mean relative errors (MREs) of SNP-SNP interaction p-values obtained with proxy-covariate logistic regression (see Equation 3). We tested different numbers of clusters (K and G) and SNP subset sizes (1%, 5%, 10% and 100%) in two evaluation GWAS datasets from *IBD* and *COVID-19* cohorts (Table 1). MRE is computed relative to a ground truth (GT) derived from logistic regression with sample-wise covariate correction using 10 per-sample principal components (Equation 2). For each subset size, the minimal MRE among evaluated cluster sizes per clustering algorithm is underlined; the globally best-performing setting per dataset is highlighted in **bold**.

		IBD				*COVID-19*			
	method	1 %	5 %	10 %	100 %	1 %	5 %	10 %	100 %
	w/o covar	0.1563	0.0997	0.0867	0.1089	0.0768	0.0589	0.0539	0.0703
K-means	K=2	0.1457	0.0944	0.0823	0.0826	0.0766	0.0588	0.0538	0.0701
	K=3	0.1160	0.0786	0.0699	0.0648	0.0436	0.0354	0.0332	0.0347
	K=4	0.1161	0.0789	0.0704	0.0656	0.0441	0.0360	0.0337	0.0355
	K=5	0.1206	0.0806	0.0716	0.0714	**0.0371**	**0.0304**	**0.0284**	**0.0283**
	K=6	0.1202	0.0815	0.0716	0.0749	0.0377	0.0311	0.0291	0.0293
	K=7	0.1200	0.0812	0.0719	0.0738	0.0391	0.0322	0.0302	0.0307
	K=8	0.1216	0.0817	0.0719	0.0797	0.0397	0.0326	0.0306	0.0312
	K=9	0.1223	0.0813	0.0717	0.0661	0.0575	0.0464	0.0432	0.0510
	K=10	0.1226	0.0821	0.0722	0.0774	0.0587	0.0475	0.0440	0.0521
	K=11	0.1198	0.0807	0.0715	0.0715	0.0598	0.0484	0.0447	0.0532
	K=12	0.1194	0.0802	0.0705	0.0759	0.0565	0.0463	0.0428	0.0498
	K=13	0.1179	0.0795	0.0706	0.0744	0.0607	0.0498	0.0459	0.0549
	K=14	0.1171	0.0791	0.0699	0.0831	0.0587	0.0482	0.0450	0.0523
	K=15	0.1175	0.0798	0.0695	0.0817	0.0415	0.0339	0.0313	0.0315
	K=16	0.1134	0.0772	0.0691	0.0746	0.0413	0.0339	0.0313	0.0314
	K=17	0.1165	0.0797	0.0706	0.0784	0.0412	0.0337	0.0312	0.0311
	K=18	0.1156	0.0786	0.0700	0.0787	0.0411	0.0337	0.0313	0.0308
	K=19	0.1161	0.0789	0.0696	0.0771	0.0409	0.0335	0.0311	0.0306
	K=20	0.1156	0.0790	0.0699	0.0830	0.0414	0.0339	0.0314	0.0310
GMM	G=2	0.1562	0.0996	0.0866	0.1088	0.0488	0.0393	0.0367	0.0394
	G=3	0.1160	0.0789	0.0699	0.0679	0.0439	0.0356	0.0336	0.0351
	G=4	0.1222	0.0822	0.0723	0.0630	0.0385	0.0319	0.0296	0.0302
	G=5	0.1210	0.0814	0.0717	0.0636	0.0452	0.0366	0.0344	0.0366
	G=6	**0.1121**	**0.0769**	**0.0671**	**0.0538**	0.0443	0.0362	0.0337	0.0355
	G=7	0.1239	0.0827	0.0727	0.0661	0.0534	0.0432	0.0405	0.0453
	G=8	0.1223	0.0818	0.0721	0.0646	0.0457	0.0371	0.0350	0.0367
	G=9	0.1201	0.0810	0.0713	0.0626	0.0413	0.0336	0.0311	0.0323
	G=10	0.1181	0.0799	0.0699	0.0776	0.0515	0.0420	0.0390	0.0439
	G=11	0.1196	0.0806	0.0706	0.0698	0.0530	0.0434	0.0401	0.0453
	G=12	0.1199	0.0804	0.0710	0.0563	0.0483	0.0392	0.0366	0.0387
	G=13	0.1267	0.0843	0.0742	0.0631	0.0585	0.0472	0.0436	0.0503
	G=14	0.1199	0.0803	0.0702	0.0588	0.0681	0.0546	0.0507	0.0632
	G=15	0.1374	0.0929	0.0837	0.0851	0.0597	0.0480	0.0454	0.0514
	G=16	0.1214	0.0812	0.0718	0.0620	0.0735	0.0588	0.0545	0.0688
	G=17	0.1463	0.0942	0.0818	0.0873	0.0697	0.0560	0.0510	0.0646
	G=18	0.1365	0.0905	0.0801	0.0869	0.0711	0.0562	0.0520	0.0641
	G=19	0.1403	0.0935	0.0811	0.0776	0.0714	0.0574	0.0527	0.0657
	G=20	0.1397	0.0936	0.0815	0.0785	0.0792	0.0630	0.0571	0.0762

For the *IBD* dataset, the configuration without covariates performs worst, independent of a fraction (1 %, 5 %, 10 %) or the complete dataset (100 %), which is expected. The optimal K-means result for the complete (100 %) run on the *IBD* dataset is achieved with three clusters ($K = 3$) with 40.5 % improvement when compared to the run without covariate correction, but the overall best performance is achieved by a GMM with $G = 6$ clusters. In this run the im-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

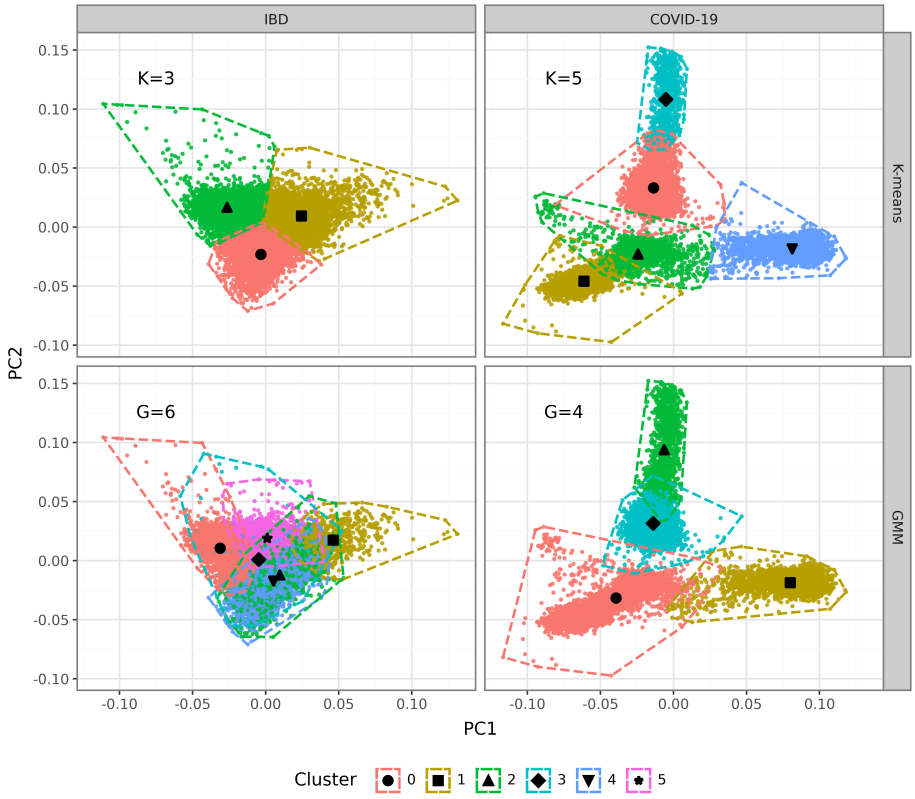

Fig. 1. Scatter plot of the first two principal components (PC1 and PC2) of the *IBD* and *COVID-19* GWAS datasets. PCs are clustered using K-means (top row) and Gaussian mixture models (GMM; bottom row) with the number of clusters (K or respectively G) selected based on minimal mean relative error (MRE) according to Table 2. Cluster centroids are marked with black symbols, and cluster boundaries (convex hulls) are indicated by dotted lines.

provement for 100 % of the SNPs achieves 50.6 % while all prediction settings with a lower fraction of SNPs (1 %, 5 %, and 10 %) underestimate the final improvements from the 100 % run (28.3 % for 1 % SNPs, 22.9 % for 5 % SNPs, and 22.6 % for 10 % SNPs). This is not a problem as the best predicted configuration also leads to the best result from a complete run. Although we further observe that for K-means the predicted results with 1 %, 5 % and 10 % of the SNPs are slightly better for $K = 16$ than for $K = 3$, which means that our framework is not able to predict the best configuration for K-means clustering correctly from the chosen fractions of SNPs, the overall best solution is determined with a GMM with $G = 6$ clusters anyway (see above). However, the results for all runs with $K > 2$ are highly similar, and if the user chose K-means clustering with $K = 16$ for the complete run, the accuracy improvement would still be 31.5 %.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

For the *COVID-19* dataset, the best MRE results over all tested (sub-)datasets are predicted and achieved from the proxy-covariate run with five K-means clusters ($K = 5$). The improvement in MRE for the complete (100 %) run achieves 59.7 % when compared to the run without covariate correction. Other K-means runs show improvements of at least 49.5 % with the exception of $K = 2$ and between $K = 9$ and $K = 14$. $K = 2$ shows almost no improvement, while for the other runs the improvement is at least 21.8 %. In contrast, proxy-covariate runs with GMM clustered covariates tend to continuously reduce accuracy improvement with increasing G. The best GMM runs are observed with $G = 4$ and a 57.0 % improvement, while the worst GMM runs with $G = 20$ introduce an error such that the accuracy even drops by -8.3 % when compared to the runs without covariate correction. It is important to note that independent of the clustering method the best predicted results also lead to the best result for the complete (100 %) run.

Scatter plots of the first two principal components (PCs) from PCA of the best configurations for both clustering methods and both evaluation GWAS datasets (*IBD* and *COVID-19*) are shown in Figure 1 (i.e. $K = 3$ and $G = 6$ for *IBD* and $K = 5$ and $G = 4$ for *COVID-19* (see Table 2)). For *IBD*, the K-means clustering (upper left subfigure) indicates three clearly distinct clusters, while the GMM with $G = 6$ does not show a clear distinction in the first two PC layers (lower left subfigure). However, GMM with $G = 6$ delivers a 17.0 % higher accuracy when compared to K-means with $K = 3$ (as discussed above). This example indicates that the clustering does not have to show a clear distinction in the first two PC dimensions to achieve a good improvement of GWAIS accuracy, and that the clustering must account for all higher PC layers as well. For *COVID-19*, the visualizations of K-means clustering with $K = 5$ (upper right subfigure) and GMM with $G = 4$ (lower right subfigure) follow the dense areas visible in the PC1-PC2 layer, which is also reflected in similar accuracy results (K-means with $K = 5$ improves GMM with $G = 4$ by only 6.3%).

Further, the accuracy of using proxy-covariates from covariate clustering in GWAIS with logistic regression, rather than no covariate correction at all, is demonstrated in scatter plots showing the best p-values from each of the evaluated configurations (including no covariates) against the ground truth (using sample-wise covariate correction). Figure 2 shows the scatter plots for both evaluation datasets, visualizing the best p-values from the run without covariates against the ground truth (upper subfigures) in comparison to the scatter plots from the run with the best covariate clustering from our framework for each dataset (lower subfigures). In the plots for the results without covariates a wide scattering of data points can be observed, indicating that the correlation of these results to the GT is weak. In contrast, the runs with proxy-covariates show the data points densely located around the ideal line (blue diagonal), which in turn indicates a strong correlation.

Besides accuracy, the choice for the optimal configuration for a complete proxy-covariate logistics regression run also depends strongly on the predicted runtime for these configurations. Table 3 shows the predicted wall clock runtimes

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

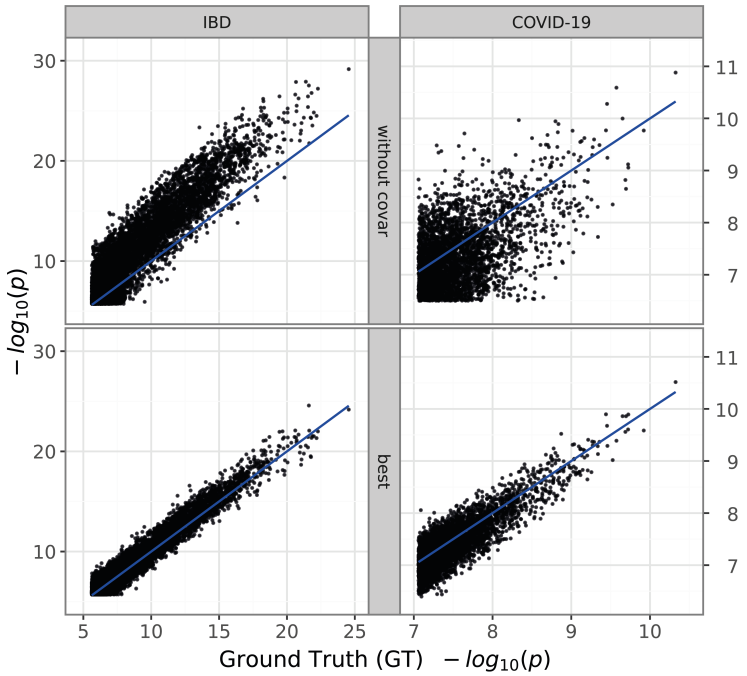

Fig. 2. Scatter plots comparing $-log_{10}(p)$-values from the genome-wide interactions analyses of the *IBD* and *COVID-19* GWAS datasets to the corresponding ground truth (GT). The GT is defined as logistic regression with sample-wise covariates using 10 per-sample principal components (x-axis). Upper panels show results without covariate adjustment; lower panels show results using the best-performing proxy-covariate configuration selected according to Table 2 (i.e. $G = 6$ for *IBD* and $K = 5$ for COVID-19). Each point represents one SNP-SNP interaction. The blue diagonal indicates perfect agreement with the GT.

of a complete run for chosen clustering configurations for subsets of 1 %, 5 %, and 10 % of the input data besides the measured runtime of the complete run. The table shows that the best runtime predictions can be made from a subset with 5 % of the available SNPs, while the most unreliable prediction is from the configuration with a subset of only 1 %, especially for the *IBD* dataset, as it contains only 1,416 SNPs. In general, a linear correlation between the runtime and the number of clusters is expected. Thus, the user should weigh how much trade-off should be made between the predicted runtime and the expected accuracy. For the *IBD* dataset, the overall most accurate proxy-covariate run using GMM with $G = 6$ is approximately three times slower than the best K-means run with $K = 3$, but in summary, all proxy-covariate runs are significantly faster than the complete run with sample-wise covariate correction, e.g., for *IBD* the $G = 6$ run is 35x times faster, and for *COVID-19* the $K = 5$ run is 154x faster.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

Table 3. Predicted and measured wall clock execution runtimes for the two evaluation datasets *IBD* and *COVID-19* (Table 1) for the runs with sample-wise covariate correction (ground truth) and chosen clustering configurations. Predictions are presented for test runs with 1 %, 5 %, 10 % of the available SNPs. The measured runtime is presented in the 100 % column. Times are displayed as HH:MM:SS.

	predicted from			measured
	1%	5%	10%	100%
IBD:				
w/ covar	216:52:57	215:08:35	213:34:20	216:25:00
K-means $K = 3$ (best K)	$\leq$ 00:00:01	02:06:41	02:25:01	01:55:17
K-means $K = 16$ (predicted best K)	02:46:50	06:33:24	05:30:02	05:29:07
GMM $G = 6$ (overall best)	08:20:30	06:20:03	06:31:42	06:11:46
COVID-19:				
w/ covar	5301:12:27	5273:34:59	5299:14:33	*5297:17:02
K-means $K = 5$ (overall best)	36:07:10	34:13:26	34:16:43	34:19:12
GMM $G = 4$ (best G)	30:33:45	32:00:05	30:35:03	32:00:10

*estimated from an independent 1 % test runs of the complete run

Further, the overhead introduced by the framework itself for testing the different settings to predict the optimal configuration must also be considered for runtime analysis. A test run with 5 % of the available SNPs adds 1 hour and 4 minutes to the best complete clustering run with $G = 6$ for the *IBD* dataset, resulting in a total runtime of 7 hours and 16 minutes, which is a speedup of 30x when compared to the run with sample-wise covariate correction. For *COVID-19* a 5 % test run adds 18 hours and 14 minutes to the best complete clustering run with $K = 5$, resulting in a total runtime of 52 hours and 34 minutes, which is still a speedup of more than 100x. However, as the *COVID-19* dataset contains more SNPs, the predictions from a 1 % test run are equally satisfying, and as the execution time is expected to rise approximately quadratically with the number of SNPs, this run takes only 45 minutes. This leads to a total speedup of 151x with a reasonable total runtime of only 32 hours and 45 minutes compared to an impractical runtime of more than 220 days for the complete run using sample-wise covariate adjustment.

4 Conclusion

We developed a new framework in order to tackle the issue of finding optimal parameters for covariate clustering in statistical epistasis detection (as presented in Neff et. al. [13]). It tests several clustering algorithms (K-means and Gaussian mixture model (GMM)) and their parameters on a reduced input dataset to quickly predict the runtime and mean relative error (MRE) compared to an experiment with per-sample covariates for a run with the complete dataset. From the tested parameters, the user may choose their favorite configuration, consid-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

ering their own tradeoff between expected runtime and accuracy, to conduct the epistasis screening process using logistic regression with covariate clustering [13].

We tested the framework with three different configurations (using 1 %, 5 % and 10 % of the input SNPs) on two real-world GWAS datasets: a dataset with 7,787 cases and controls from an IBD cohort in Germany and a dataset from a COVID-19 study [6, 8] comprising 16,739 participants from five different European countries (details in Table 1). We also generated the results from the complete datasets (100 %) for a comparison of the accuracy and runtime predictions to the three preliminary test runs using the input subsets.

We observed that by using only a subset of the input data, we were able to successfully predict the optimal clustering parameters for both complete datasets in all three configurations. Importantly, for the *IBD* dataset, the 1 % subset was already sufficient to identify a GMM configuration with $G = 6$, which achieved lower MRE than the K-means configuration with $K = 3$ previously obtained in Neff et al. [13] via complete genome-wide screening runs (but restricted to K-means clustering). For the *COVID-19* dataset, the previously reported configuration (K-means clustering with $K = 5$) was consistently recovered also using only 1 % of the input SNPs, demonstrating the robustness of the approach.

However, the runtime predictions from the 1 % configuration turned out to be unreliable for the *IBD* dataset due to its smaller size, which is why we recommend to use subsets of 5 % for datasets with a lower number of SNPs (e.g. below 400,000) while a 1 % configuration is sufficient for larger datasets.

By using our new framework to determine the most accurate clustering configuration with 5 % of the input SNPs first, following a complete proxy-covariate logistic regression run with this configuration ($G = 6$), we achieve a total speedup of 30x for the *IBD* dataset when compared to a logistic regression without clustering and sample-wise covariate correction. This is a runtime reduction from more than 9 days to 7 hours and 16 minutes. For the *COVID-19* dataset, the runtime advantage is a reduction from more than 220 days to less than one and a half days (151x speedup for a pilot run with our framework and 1 % of the input SNPs, and K-means clustering with $K = 5$ for the complete run).

We demonstrate that by using our new data-driven framework, optimal clustering and its parameters can be successfully and reliably predicted. This enables the practical use of epistasis detection with proxy-covariate adjustment as the challenge to select the optimal configuration for covariate clustering can be resolved with only a negligible overhead in advance. It is no longer necessary to run a computationally intensive screening with sample-wise covariate adjustment.

Future work includes a user-friendly tool that helps scientists select their optimal parameters, based on a clean PDF report containing all necessary scores and visualizations.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

Acknowledgments. Written informed consent was obtained from all study participants and approved by the Ethics Committee of the Medical Faculty of Kiel University and the University Medical Center Schleswig-Holstein. The authors would like to thank the COVID-19 GWAS Group [6, 8] for the chance to use the COVID-19 GWAS data for benchmarks. The project was funded by DFG Research Unit 5042: miTarget - The Microbiome as a Therapeutic Target in Inflammatory Bowel Diseases (RU 5042; Project-ID 426660215, Project: INF (EL 831/5-2)). The study received infrastructure support funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2167/2 - 390884018. This research was supported in part through high-performance computing resources available at the Kiel University Computing Centre.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Abraham, G., Qiu, Y., Inouye, M.: FlashPCA2: Principal component analysis of Biobank-scale genotype datasets. Bioinformatics **33**(17) (2017)
[2] Bradley, P., Mangasarian, O., Street, N.: Clustering via concave minimization. In: Advances in Neural Information Processing Systems 9 (1996)
[3] Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J.: Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **4** (2015)
[4] Cordell, H.J.: Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics **11**(20) (2002)
[5] Curtin, R.R., Edel, M., et al.: Mlpack 4: A fast, header-only C++ machine learninglibrary. Journal of Open Source Software **8**(82) (2023)
[6] Degenhardt, F., Ellinghaus, D., et al.: Detailed stratified GWAS analysis for severe COVID-19 in four European populations. Human Molecular Genetics **31**(23) (2022)
[7] Deisenroth, M.P., Faisal, A.A., Ong, C.S.: Mathematics for Machine Learning. Cambridge University Press (2020)
[8] Ellinghaus, D., Degenhardt, F., Bujanda, L., et al.: Genomewide Association Study of Severe Covid-19 with Respiratory Failure. The New England Journal of Medicine **383**(16) (Oct 2020)
[9] Hoffmann, M., Poschenrieder, J.M., Incudini, M., Baier, S., Fritz, A., et al.: Network medicine-based epistasis detection in complex diseases: ready for quantum computing. Nucleic Acids Research **52**(17) (2024)
[10] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: A local search approximation algorithm for k-means clustering. Computational Geometry **28**(2–3) (2004)
[11] Kässens, J.C., Wienbrandt, L., Ellinghaus, D.: BIGwas: Single-command quality control and association testing for multi-cohort and biobank-scale GWAS/PheWAS data. Gigascience **10**(6) (2021)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17

[12] Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory **28**(2) (1982)
[13] Neff, V., Wienbrandt, L., Ellinghaus, D.: Logistic regression with covariate clustering in genome-wide association interaction studies. In: Paszynski, M., Barnard, A.S., Zhang, Y.J. (eds.) Computational Science – ICCS 2025 Workshops, vol. 15907. Springer Nature Switzerland, Cham (2025)
[14] Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12** (2011)
[15] Sazonovs, A., Stevens, C.R., Venkataraman, G.R., Yuan, K., Avila, B., et al.: Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility. Nature Genetics **54**(9) (2022)
[16] Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., et al.: BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. American Journal of Human Genetics **87**(3) (2010)
[17] Wang, K.: Genetic association tests in the presence of epistasis or gene-environment interaction. Genetic Epidemiology **32**(7) (2008)
[18] Wienbrandt, L., Kässens, J.C., Hübenthal, M., Ellinghaus, D.: 1000× faster than PLINK: Combined FPGA and GPU accelerators for logistic regression-based detection of epistasis. Journal of Computational Science **30** (2019)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_17