Logistic Regression with 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. Logistic regression with covariates is the gold standard for detecting epistasis (statistical genetic interactions) when analysing case-control datasets from genome-wide association studies (GWAS) of diseases. Nevertheless, genome-wide interaction studies (GWAIS) are still performed without covariate correction for performance reasons as the analysis of modern GWAS datasets may lead to several weeks of computation time. However, omitting necessary covariate information causes a substantial statistical error in most studies requiring genetic ancestry adjustment via principal component analysis (PCA). Here, we present a novel approach that uses proxy covariates generated by k-means clustering in combination with contingency tables to reduce the runtime complexity of logistic regression from $\mathcal{O}(NI)$ to $\mathcal{O}(N+IK)$ and to minimize the statistical error to a ground truth (GT) implementation that uses per-sample covariate vectors from the PCA. By using GWAS data with 141,621 genetic markers from 3,520 German patients with inflammatory bowel disease (IBD) and 4,288 healthy controls, we demonstrated a 97-fold speed-up with two k-means clusters from PCA covariates compared to the GT implementation. At the same time, we improved the mean relative error (MRE) by more than 55 % when compared to logistic regression without covariate correction. Our developments enable logistic regression-based epistasis analysis with clustered PCA covariates for GWAS datasets on a genome-wide scale.

Keywords: Logistic Regression · Covariates · k-Means Clustering · Epistasis · GWAS · GWAIS · Bioinformatics

1 Introduction

Genome-wide association interaction studies (GWAIS) is a common tool to detect associations between Single Nucleotide Polymorphisms (SNPs) and disease status, typically via logistic regression [8]. In contrast to genome-wide association studies (GWAS) [9], where SNP markers are tested individually for assocation with disease status, GWAIS are conducted to detect (usually pairwise) gene-gene (GxG) or SNP-SNP interactions, also known as statistical epistasis, which

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_26

refers to the deviation from additive effects of the SNP markers. Consequently, for an exhaustive genome-wide search of all possible combinations of SNP-SNP interactions, the algorithmic complexity increases quadratically with the number of SNP markers to be tested. To obtain statistically reasonable results in GWAIS, large GWAS datasets with at least several thousands of cases and controls must be tested for association to achieve sufficient statistical power [13]. Recently, using GWAS data from UK Biobank for more than 70 human traits, it was estimated that epistatic (non-additive) effects between SNP markers make the second largest contribution to the heritability of complex traits, behind the contribution of additive effects of SNP markers [4]. Therefore, efficient, accurate, and interpretable methods for epistasis detection are needed to analyze the epistatic genetic component on a genome-wide level.

Several epistasis tools such as BOOST [11] (heuristic) or PLINK's [2] multiplicative logistic regression model enable an exhaustive (i.e. complete) search for SNP-SNP interactions on a genome-wide scale, but suffer from long runtimes, and thus need to be carried out using cloud computing [3] or high-performance clusters [10]. Other tools suggest machine learning approaches [1] or even quantum-computing [5] to avoid exhaustive genome-wide searches, but this carries the risk that significant interactions remain undetected. However, most existing tools are unable to handle covariates, which are important variables that should be taken into account in association studies, as they may confound the outcome variable (here case-control status) in GWAS and GWAIS. Omitting covariates produces lower quality results with less accuracy when compared to the same results achieved from fitting a logistic regression model using covariate correction [15]. As a standard procedure in GWAS and GWAIS quality control principal component analysis (PCA) is used to characterize the population structure (genetic ancestry) of the study participants, and the principal components (PCs) from PCA are usually used as covariates in the association analysis to correct for population stratification.

Here, we present a new algorithm to introduce covariate correction in logistic regression for epistasis detection. We reduce the runtime complexity to $\mathcal{O}(N + IK)$ by using contingency tables and *k-means* clustering for the covariates with subsequent sample classification based on the cluster affiliation of each sample (N is the number of input samples, I the number of fitting iterations, K the number of clusters). Evaluation on a real-world dataset shows a 97-fold speedup and 55% improvement in accuracy using only K=2 clusters.

2 Methods

The classical additive model for logistic regression in epistasis detection has a runtime complexity of $\mathcal{O}(NI)$ and can be optimized to $\mathcal{O}(N + I)$ using contingency tables (see [14], N is the number of input samples, I the number of fitting iterations). The approach provides a 10-fold speed advantage over the naive implementation, but lacks the application of covariates to enhance accuracy.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_26

Our new approach achieves a similar runtime complexity of $\mathcal{O}(N + IK)$ at a higher accuracy using K *proxy covariate* vectors. We demonstrate how to generate reasonable proxy covariates using k-means clustering.

For the derivation, we presume a cohort of N study participants (samples) and a binary trait (phenotype) Y, distributed in cases (y_i=1) and controls (y_i=0). For each SNP pair (A, B) to be tested, we know the corresponding genotypes $g_{i,A/B} \in G = \{0, 1, 2\}$ with the following encoding: (0) *homozygous reference*, (1) *heterozygous*, (2) *homozygous variant*. Furthermore, each sample is bound to a covariate vector $\overrightarrow{v_i} \in \mathbb{R}^L$. We define $\mathcal{V} \in \{\overrightarrow{v_0}, \ldots, \overrightarrow{v_{N-1}}\}$ as the set of all covariate vectors from all samples. W.l.o.g. we assume $\overrightarrow{v_i} \neq \overrightarrow{v_j}\ \forall i \neq j$.

2.1 Additive Two-locus Logistic Regression

A commonly used statistically model is an additive two-locus logistic regression as shown in Equation 1 [12, 15]. It uses the characteristic that SNPs occurs in only three different genotypes. Additionally, it uses the above-described environment.

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

The model is fitted by using an maximum likelihood (ML) estimation over all N input samples in multiple iterations I. A score is calculated to test for a correlation between the two SNPs and the phenotype (trait). As the score follows a chi-squared distribution with one degree of freedom, a p-value can be directly derived. The simplified runtime complexity for fitting this model is $\mathcal{O}(NI)$ for each test, i.e. for each pair of SNPs. A commonly used implementation of this model without the correction for covariates can be found in the popular bioinformatics toolset *PLINK* [2, 6].

Wienbrandt et al. published a more efficient implementation of the fitting of Equation 1 without covariates [14]. They benefit from the limited number of combinations for the genotypes $g_{A/B} \in G$ and create pre-calculated contingency tables $\mathcal{N}_{3\times3}$ by counting the occurrences of each genotype combination $(n_{j,k})$ separately for cases ($\mathcal{N}^{\text{case}}_{3\times3}$) and control samples ($\mathcal{N}^{\text{ctrl}}_{3\times3}$). This simplified the ML equation and reduced the runtime complexity to $\mathcal{O}(N + I)$.

2.2 Logistic Regression with Contingency Tables and Covariates

For a classification by genotype and the additional covariate vector, we introduce n_{v,g_A,g_B} which indicates the number of samples with the same covariate vector v and genotypes $g_{A/B} \in G$ at SNP positions A and B. In correspondence to [14], a contingency table that takes a covariate vector into account is defined as:

$$\mathcal{N}_{v,3\times3} = (n_{v,j,k})_{3\times3} \quad \text{with} \quad n_{v,j,k} = \sum_{i=0}^{N-1} [g_{i,A} = j \wedge g_{i,B} = k \wedge v = v_i] \quad (2)$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_26

Building on the extended contingency table, we update the ML function from Wienbrandt et. al. [14] by an additional covariate vector component to:

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

$$\text{with} \quad p_{v,g_A,g_B} = \frac{1}{1 + e^{-z_{v,g_A,g_B}}} \quad (3b)$$

$$\text{and} \quad z_{v,g_A,g_B} = \beta_0 + \beta_1 g_A + \beta_2 g_B + \beta_3 g_A g_B + \sum_{l=0}^{|\overrightarrow{v}|-1} \beta_{4+l} v_l \quad (3c)$$

Obviously, n_{v,g_A,g_B} is expected to be 1 for only one combination of v, g_A and g_B, and 0 otherwise. As w.l.o.g. there are N different covariate vectors in $\mathcal{V}$, the runtime complexity can directly be determined as $\mathcal{O}(NI)$, which is equal to the calculation without contingency tables.

2.3 Proxy Covariates and Clustering

To reduce the runtime complexity, our approach is to classify all samples by a fixed number of $K \ll N$ covariate vector prototypes, which we introduce as *proxy covariates*. In the logistic regression calculation, we substitute the real covariate vector v_i for a sample by its proxy covariate $\hat{v}_i$. We define the set of all proxy-covariates as $\hat{V}$, and Equation 3a evaluates to Equation 4.

$$\hat{\mathbf{L}}_{\ln}(\beta) = \sum_{\hat{v} \in \hat{\mathcal{V}}} \sum_{g_A=0}^{2} \sum_{g_B=0}^{2} \left[n^{\text{case}}_{\hat{v},g_A,g_B} \ln(p_{\hat{v},g_A,g_B}) + n^{\text{ctrl}}_{\hat{v},g_A,g_B} \ln(1 - p_{\hat{v},g_A,g_B}) \right] \quad (4)$$

As the number of distinct proxy-covariates is fixed ($|\hat{V}| = K$), we can evaluate the runtime complexity to $\mathcal{O}(N + IK)$. However, by replacing the real covariates with a prototype, we introduce an error ε which is preferably minimized: $\hat{\mathbf{L}}_{\ln}(\beta) = \mathbf{L}_{\ln}(\beta) + \varepsilon$.

In order to minimize the error ε, we choose *k-means clustering* of the covariate vectors and use the cluster centers as appropriate proxy covariates $\hat{v}_k$ that substitute the real covariate vectors of each sample as a prototype. In k-means clustering, it is guaranteed that for each sample the distance to its cluster center is minimized, i.e. the distance is shorter than to any other cluster center. Further, the cluster centers are well-distanced by partitioning the covariate vectors into nonoverlapping cells, which is essential for successfully fitting the regression equation. A common problem for k-means clustering can be noise, which should be eliminated by a proper quality control of the input data beforehand. Another problem is that the number of clusters K must be chosen in advance. In general, the best choice of K differs for each input dataset. In our evaluation, we demonstrate the influence of different K on our benchmark dataset.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_26

3 Results

To evaluate our new algorithm, we used a quality-controlled dataset comprising 7,808 samples divided into 3,520 German patients with inflammatory bowel disease (IBD) and 4,288 healthy controls, genotyped at 141,621 SNPs. From PCA we used the first 10 principal components (PCs) as covariates. We evaluated the dataset in several different configurations: (a) we determined the ground truth (GT) by running the implementation of Equation 3, i.e. the logistic regression with individual covariate correction without clustering, (b) we conducted logistic regression without covariate correction (LogReg), and (c) we evaluated logistic regression with k-means clustering of covariates with different choices of $K \in \{\, 2, 4, 8, 12, 16 \,\}$ (Equation 4).

For the ground truth (GT) run, we kept the best 100,000 results after checking for plausibility, i.e. all results with a reported odds-ratio of zero or infinity were removed in advance. The results are sorted ascending by the reported p-value. From our test runs we keep the best 10 million results in expectation to be able to reproduce the best outcomes from the GT within these results. The tool was implemented in C++ and compiled with *g++* v13.3.0. Runtimes were measured as wall-clock execution times on our benchmark server equipped with two Intel Xeon Gold 6538Y+ processors. In total, 64 physical cores with 128 threads at a base clock frequency of 2.2 GHz were used.

We measure accuracy by firstly locating all exact SNP pairs from the best GT results (S=100,000) in the results of the test run. The difference in the scores is then calculated as the *relative error (RE)* in their logarithmized p-values. To summarize the errors, we calculate the *mean relative error (MRE)*. For a sample i and its p-value from the test run p_i and from GT p_i^{GT}, MRE is defined as follows:

$$\mathrm{MRE} = \frac{1}{S} \sum_{i=0}^{S-1} \mathrm{RE}_i \quad \text{with} \quad \mathrm{RE}_i = \frac{|\log p_i - \log p_i^{\mathrm{GT}}|}{|\log p_i^{\mathrm{GT}}|} \tag{5}$$

As it is possible that results from the GT cannot be found in the best results of a test run, we calculate two versions of the MRE: $\mathrm{MRE}^{\mathrm{best}}$ and $\mathrm{MRE}^{\mathrm{worst}}$ where we replace the corresponding p-value, that was not found in the test run, with either $p_i = \max\{\, p \,\}$ (best-case estimation, as we know that the p-value cannot be lower than the least best recorded result) or $p_i = 1$ (worst-case estimation, as this is the maximum for any p-value). Further, we determine the number of missed results from the GT in each test run.

Table 1 presents the runtimes as well as MRE with both estimations and the number of misses for all our test runs. For the run without covariates and the run with K=2 clusters, we generated quantile-quantile (Q-Q) plots in Figure 1 representing the best 100,000 results from the GT and where they were found in the corrsponding results from the test runs.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_26

Table 1. Runtimes and mean relative error (MRE) for logistic regression of our benchmark dataset without covariate correction (LogReg) and for a different number of clusters (K) in comparison to the ground truth (GT), which was calculated with covariate correction for each sample individually (without clustering). MRE^{best} and $\text{MRE}^{\text{worst}}$ represent best-case and worst-case estimations of the real MRE in the presence of missing results. The number of results missing in the GT run is listed in the last row.

	GT	LogReg	K=2	K=4	K=8	K=12	K=16
Runtime (HH:MM)	92:38	00:33	00:57	01:07	01:50	01:49	02:11
Speedup vs. GT	1.00	168.42	97.51	82.96	50.53	50.99	42.43
MRE^{best}	-	0.2335	0.1037	0.1139	0.1235	0.1123	0.1117
$\text{MRE}^{\text{worst}}$	-	0.3010	0.1039	0.1143	0.1245	0.1130	0.1121
Misses	-	1,437	31	54	125	83	58

4 Discussion

In this paper, we showed that proxy covariates for logistic regression in epistasis tests can reduce the computational complexity and, consequently, the execution time while preserving the accuracy compared to an ideal calculation with individual covariate correction for each sample. The runtime of the ground truth (GT) run could therefore be reduced from almost 4 days for our test dataset to only 57 minutes (for K=2) or up to 131 minutes (for K=16). This is a speedup between 42 and more than 97. Only calculating without covariate correction is still faster with a speedup of 168. To quantify this improvement, we measured the mean relative error (MRE) by comparing the scores generated by the best 100,000 results from the GT run with the scores from the same SNP pairs in the test runs. We demonstrated that using proxy covariates generated by k-means clustering with only two clusters (K=2) results in a worst-case estimation of $\text{MRE}^{\text{worst}}_{K=2} = 0.1039$, which is almost the same as the best-case estimation $\text{MRE}^{\text{best}}_{K=2} = 0.1037$. Thus, the worst case estimation for K=2 is more than two times better than the best case estimation without covariate correction ($\text{MRE}^{\text{best}}_{\text{LogReg}} = 0.2335$). We also highlight that for two clusters only 31 scores out of 100,000 from the GT were not found in the results of this test run, while, in contrast, 1,437 results from the GT were not among the best 10 million results of the *LogReg* run.

The results from Table 1 and Figure 1 also show that our new approach with k-means clustered covariates performs better over the full range of tested K compared to an implementation without covariate correction. However, it is noticeable that runs with more clusters (K>2) do not necessarily perform better (in terms of MRE and missed p-values) than for K=2. The $\text{MRE}^{\text{worst}}$ for K=8 is the worst MRE in our test runs with clustering and is around 20 % lower than the MRE^{best} for K=2. The effect is even higher for missing results. In the run with K=8 four times more results from the GT than in the version with K=2 k-mean clusters were not found. We explain this behaviour by the uniformity of our test dataset. Our GWAS quality control ensured that outliers were removed in

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_26

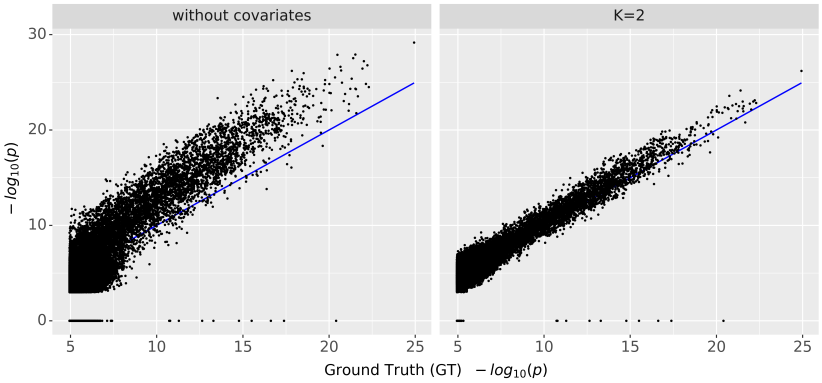

Fig. 1. Q-Q plots that visualize the different $-log_{10}(p)$-values calculated without any covariates (left subfigure), and clustered covariates from Equation 4 with K=2 (right subfigure) in comparison to those calculated by Equation 3 (referred to as ground truth (GT)). The plots show the best 100,000 GT results (x-axis) and their corresponding results in the test runs (y-axis). If a SNP pair was not found in the output file, the $-log_{10}(p)$-value was set to 0, and the corresponding data points are located on the x-axis. The blue main diagonal indicates the ideal correlation.

advance and the study was originally focused on a homogeneous study population such that the clustering of PCs did not lead to clearly distinct clusters for K>2. Nevertheless, setups with heterogeneous datasets are not uncommon, hence we will examine further non-uniform datasets to confirm this thesis in the future.

Finally, our goal is to use covariate correction for epistasis detection in datasets that exceed the size of our test data significantly. Based on the implementation of an exhaustive GWAIS in Wienbrandt et al. [14] that used hardware acceleration to handle datasets with millions of genetic markers and tens of thousands of samples, we target the implementation of our strategy on FPGA and GPU acceleration hardware as well. With a speedup in the same order of magnitude, the analysis of our test dataset would be finished in seconds, and large datasets can be processed in terms of hours and days instead of years.

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

Acknowledgements. 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 cohort was first described in Sazonovs et al. [7]. The project was funded by DFG Research Unit 5042: miTarget - miTarget - The Microbiome as a Therapeutic Target in Inflammatory Bowel Diseases (RU 5042; Project-ID 426660215, Project: INF (EL 831/5-1, EL 831/5-2)). The study received infrastructure support from the German Research Foundation (DFG) Cluster of Excellence 2167 “Precision Medicine in Chronic Inflammation (PMI)” (EXC 2167-390884018).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_26

Bibliography

[1] Carmelo, V.A.O., Kogelman, L.J.A., Madsen, M.B., Kadarmideen, H.N.: WISH-R- a fast and efficient tool for construction of epistatic networks for complex traits and diseases. BMC Bioinformatics **19**(1), 277 (Jul 2018)

[2] 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**, 1–16 (2015)

[3] Guo, X., Meng, Y., Yu, N., Pan, Y.: Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering. BMC Bioinformatics **15**(1), 102 (2014)

[4] Hivert, V., Sidorenko, J., Rohart, F., Goddard, M.E., Yang, J., et al.: Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. The American Journal of Human Genetics **108**(5), 786–798 (2021)

[5] 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), 10144–10160 (08 2024)

[6] Purcell, S., Neale, B., Todd-Brown, K., et al.: PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics **81**, 559–575 (2007)

[7] 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), 1275–1283 (2022)

[8] Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., Meyre, D.: Benefits and limitations of genome-wide association studies. Nature Reviews Genetics **20**(8), 467–484 (2019)

[9] Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., et al.: Genome-wide association studies. Nature Reviews Methods Primers **1**(1), 59 (2021)

[10] Upton, A., Trelles, O., Cornejo-García, J.A., Perkins, J.R.: Review: High-performance computing to detect epistasis in genome scale data sets. Brief. Bioinform. **17**(3), 368–379 (2016)

[11] 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), 325–340 (2010)

[12] Wang, K.: Genetic association tests in the presence of epistasis or gene-environment interaction. Genetic Epidemiology **32**(7), 606–614 (2008)

[13] Wei, W.H., Hemani, G., Haley, C.S.: Detecting epistasis in human complex traits. Nature Reviews Genetics **15**(11), 722–733 (2014)

[14] 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**, 183–193 (2019)

[15] Ziegler, A., König, I.R.: A Statistical Approach to Genetic Epidemiology. Wiley-VCH, 2nd edn. (2010)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_26