LFMM 2.0: Latent factor models for confounder adjustment in genome and epigenome-wide association studies

Kevin Caye[1] Olivier François[123,⋆]

[1] Université Grenoble-Alpes, [2] Centre National de la Recherche Scientifique, [3] Grenoble INP

Address: TIMC-IMAG UMR 5525, 38000 Grenoble, France.

⋆ Corresponding author: olivier.francois@grenoble-inp.fr

Abstract

Genome-wide, epigenome-wide and gene-environment association studies are plagued with the problems of confounding and causality. Although those problems have received considerable attention in each application field, no consensus have emerged on best practices in this respect. Current methods use approximate heuristics for estimating confounders, and often ignore correlation between confounders and primary variables, resulting in suboptimal power and precision. In this study, we developed a least-squares estimation theory of confounder estimation using latent factor models, providing a unique framework for several categories of genomic data. Based on statistical learning methods, the proposed algorithms are fast and efficient, and they were proven to provide optimal solutions mathematically. In simulations, the algorithms outperformed commonly used methods based on principal components and surrogate variable analysis. In analysis of methylation profiles and genotypic data, they provided new insights on the molecular basis on diseases and adaptation of humans to their environment. Software is available in the R package `lfmm` at `https://bcm-uga.github.io/lfmm/`.

1 Introduction

Association studies have been extensively used to identify candidate genes or molecular markers associated with disease states, exposure levels or phenotypic traits. Given a large number of target variables, the objective of those studies is to test whether any of the variables exhibits significant correlation with a primary variable of interest. The most common association studies are genome-wide association studies (GWAS) that focus on single-nucleotide polymorphisms (SNPs) by examining genetic variants in different individuals [2]. In recent years, other categories of association studies have emerged and become important. Of specific interest, epigenome-wide association studies (EWAS) measure DNA methylation levels in different individuals to derive associations between epigenetic variation and exposure levels or phenotypes [35]. Gene-environment association studies (GEAS) test for correlation between genetic loci and ecological variables in order to detect signatures of environmental adaptation [36].

Although they could bring useful information on the causes of diseases or on biological functions, association studies suffer from the problem of confounding. This problem arises when there exist unobserved variables that correlate both with primary variables and genomic data [42]. Confounding inflates test statistics, and early approaches consisted of introducing inflation factors to correct for the bias [5]. Statistically, inflation factors represent an empirical null-hypothesis testing approach, which is frequently used in gene expression studies [10]. GWAS have addressed the confounding issue by including known or inferred confounding factors as covariates in regression models. A prominent GWAS correction method is principal component analysis (PCA) which adjusts for confounding by using the largest PCs of the genotypic data [32]. A drawback of the approach is that the largest PCs may also correlate with the primary variables, and removing their effects can result in loss of statistical power. In gene expression studies where batch effects are source of unwanted variation, alternative approaches to the confounder problem have been proposed. These methods are based on latent factor regression models, also termed surrogate variable analysis (SVA) [26, 7]. Latent factor models have also been considered in GEAS (LFMM, [13]) and in EWAS for dealing with cell-type composition without reference samples (RefFreeEWAS, [21, 39]). Latent factor models employ deconvolution methods in which unobserved batch effects, ancestry or cell-type composition are integrated in the regression model using hidden factors. Those models have been additionally applied to transcriptome analysis [22]. As they do not make specific hypotheses regarding the nature of the data, latent factor models could be applied to any category of association studies regardless of their application field. Method choices and best practices are, however, specific to each field, and have been extensively debated in recent surveys [43, 24].

Most inference methods for latent factor regression models are based on heuristic approaches, lacking theoretical guarantees for identifiability, numerical convergence or statistical efficiency [42]. In addition, existing methods do not always address the confounding problem correctly, building confounder estimates on genetic markers only while ignoring the primary variables. In this study, we propose confounder estimation algorithms that explicitly account for the correlation between confounders and primary variables. The algorithms are based on two distinct regularized least squares methods for *latent factor mixed models*. We present the methods and theoretical developments in the next section. Then we demonstrate that the new methods achieve increased power compared to standard methods in simulations, in an EWAS of patients with rhumatoid arthritis, in a GWAS of patients with celiac disease, and lead to new discoveries in a GEAS of individuals from the 1,000 Genomes Project.

2 LFMM algorithms

Consider an $n \times p$ response matrix, $\mathbf{Y}$, recording data for n individuals. The individual data can correspond to genotypes, methylation profiles or gene expression levels measured from p genetic markers or probes. In addition, consider an $n \times d$ matrix, $\mathbf{X}$, of individual observations, recording variables of primary interest such as phenotypes or exposure levels. Additional covariates including age and gender of individuals as well as observed confounders could be included in the $\mathbf{X}$ matrix. Association methods evaluate correlation between the response matrix and the primary variables, and commonly rely on regression models. Latent factor mixed models (LFMMs) are particular regression models defined by a combination of fixed and latent effects [26, 13, 42] as follows

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^T + \mathbf{W} + \mathbf{E}, \tag{1}$$

where $\mathbf{W}$ is a $n \times p$ *latent matrix* of rank K. We defined $\mathbf{U}$ and $\mathbf{V}$ as the unique factor and loading matrices obtained from a (rank K) PCA of $\mathbf{W}$,

$$\mathbf{W} = \mathbf{U}\mathbf{V}^T .$$

Unobserved confounders are modeled through the $n \times K$ matrix of latent factors, $\mathbf{U}$, where the number of confounders, K, is determined by model choice procedures (see below). Loadings corresponding to each latent variable are recorded in the $\mathbf{V}$ matrix, which has dimension $p \times K$. Fixed effect sizes are recorded in the $\mathbf{B}$ matrix, which has dimension $p \times d$. The $\mathbf{E}$ matrix represents residual errors, and it has the same dimensions as the response matrix. In this section, we present two statistical learning algorithms for confounder estimation based on L^2 and L^1-regularized least-squares problems.

L^2-regularized least-squares problem. Statistical estimates of the parameter matrices $\mathbf{U}$, $\mathbf{V}$, $\mathbf{B}$ in equation (1) were computed after minimizing the following penalized loss function

$$\mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, \mathbf{B}) = \|\mathbf{Y} - \mathbf{U}\mathbf{V}^T - \mathbf{X}\mathbf{B}^T\|_F^2 + \lambda\|\mathbf{B}\|_2^2 , \quad \lambda > 0, \tag{2}$$

where $\|.\|_F$ is the Frobenius norm, $\|.\|_2$ is the L^2 norm, and λ is a regularization parameter. A positive value of the regularization parameter is necessary for identifying the parameter matrices $\mathbf{W} = \mathbf{U}\mathbf{V}^T$ and $\mathbf{B}$. To see this, note that for any matrix $\mathbf{P}$ with dimensions $d \times p$, we have

$$\|\mathbf{Y} - (\mathbf{U} - \mathbf{X}\mathbf{P})\mathbf{V}^T + \mathbf{X}(\mathbf{B}^T - \mathbf{P}\mathbf{V}^T))\|_F^2 = \|\mathbf{Y} - \mathbf{U}\mathbf{V}^T + \mathbf{X}\mathbf{B}^T\|_F^2.$$

This result entails that the minima of the unregularized ($\lambda = 0$) least-squares problem are not defined unequivocally, and infinitely many solutions of the least squares problem could exist unless a positive value is considered. As a consequence, any algorithm computing a low rank approximation of a response matrix using their first K principal components, and performing a linear regression of the residuals on $\mathbf{X}$ does not identify the regression and factor coefficients in equation (2) properly.

Ridge estimates (LFMM2). To computer the least-squares estimates of the latent factors, minimization of the $\mathcal{L}_{\text{ridge}}$ function started with a singular value decomposition (SVD) of the explanatory matrix, $\mathbf{X} = \mathbf{Q}\mathbf{\Sigma}\mathbf{R}^T$, where $\mathbf{Q}$ is an $n \times n$ unitary matrix, $\mathbf{R}$ is a $d \times d$ unitary matrix and $\mathbf{\Sigma}$ is an $n \times d$ matrix containing the singular values of $\mathbf{X}$, denoted by $(\sigma_j)_{j=1..d}$. The ridge estimates are described as follows

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\text{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y}) \tag{3}$$

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{W}}), \tag{4}$$

where $\text{svd}_K(\mathbf{A})$ is the rank K singular value decomposition of the matrix $\mathbf{A}$, $\mathbf{Id}_d$ is the $d \times d$ identity matrix, and $\mathbf{D}_\lambda$ is the $n \times n$ diagonal matrix with coefficients defined by

$$\mathbf{d}_\lambda = \left(\sqrt{\frac{\lambda}{\lambda + \sigma_1^2}}, \ldots, \sqrt{\frac{\lambda}{\lambda + \sigma_d^2}}, 1, \ldots, 1\right).$$

Theorem 1. *The estimates $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$ obtained from the principal component analysis of the matrix $\hat{\mathbf{W}}$, and the estimate $\hat{\mathbf{B}}^T$ define a global mimimum of the penalized loss function $\mathcal{L}_{\text{ridge}}$.*

The proof of Theorem 1 was based on mathematical properties of the SVD, and it can be found in appendix. The result describes a simple algorithm for computing the matrice of confounder estimates, $\hat{\mathbf{U}}$, with a computing cost determined by the algorithmic complexity of low rank approximation. According to [18], computing $\hat{\mathbf{U}}$ requires $O(npK)$ operations. This complexity reduces to $O(np\ln(K))$ operations when random projections are used (our implementation). Accounting for the computational cost of $\mathbf{Q}^T\mathbf{Y}$, the complexity of the LFMM2 algorithm of order $O(n^2p+np\ln(K))$. For studies in which the number of samples, n, is much smaller than the number of response variables, p, the computing time of ridge estimates is approximately the same as running the low rank approximation SVD algorithm on the response matrix twice.

L^1-regularized least-squares problem. In addition to the ridge estimates, a sparse regularization approach was considered by introducing penalties on the loss function based on the L^1 and nuclear norms

$$\mathcal{L}_{\text{lasso}}(\mathbf{W},\mathbf{B}) = \frac{1}{2}\left\|\mathbf{Y}-\mathbf{W}-\mathbf{X}\mathbf{B}^T\right\|_F^2 + \mu\|\mathbf{B}\|_1 + \gamma\|\mathbf{W}\|_*\,, \quad \mu,\gamma>0, \tag{5}$$

where $\|\mathbf{B}\|_1$ denotes the L^1 norm of $\mathbf{B}$, μ is an L^1 regularization parameter, $\mathbf{W}$ is the latent matrix, $\|\mathbf{W}\|_*$ denotes its nuclear norm, and γ is a regularization parameter for the nuclear norm. The L^1 norm was introduced for inducing sparsity on the fixed effects [40]. The L^1 penalty corresponds to the prior information that not all response variables may be associated with the primary variables. More specifically, the prior implies that a restricted number of rows of the effect size matrix $\mathbf{B}$ are non-zero. The second regularization term is based on the nuclear norm, and it was introduced to penalize large numbers of latent factors. In addition, it defined a convex function, and convex mimimization algorithms could be applied in order to minimize the $\mathcal{L}_{\text{lasso}}$ function [31].

Lasso estimation algorithm (LFMM1). Let us assume that the explanatory variables, $\mathbf{X}$, were scaled so that $\mathbf{X}^T\mathbf{X} = \mathbf{Id}_d$. Under this assumption, we developed a convergent block-coordinate descent method for minimizing the convex loss function $\mathcal{L}_{\text{lasso}}$ with respect to $\mathbf{B}$ and $\mathbf{W}$. The algorithm is initialized from a null-matrix, $\hat{\mathbf{W}}_0 = 0$, and iterates the following steps

1. Find $\hat{\mathbf{B}}_t$ a minimum of the penalized loss function

$$\mathcal{L}^{(1)}_{\text{lasso}}(\mathbf{B}) = \frac{1}{2}\|(\mathbf{Y}-\hat{\mathbf{W}}_{t-1})-\mathbf{X}\mathbf{B}^T\|_F^2 + \mu\|\mathbf{B}\|_1\,, \tag{6}$$

2. Find $\hat{\mathbf{W}}_t$ a minimum of the penalized loss function

$$\mathcal{L}^{(2)}_{\text{lasso}}(\mathbf{W}) = \frac{1}{2}\|(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}_t^T)-\mathbf{W}\|_F^2 + \gamma\|\mathbf{W}\|_*. \tag{7}$$

The algorithm cycles through the two steps until a convergence criterion is met or the allocated computing resource is depleted. Each minimization step has a well-defined and unique solution. To see it, note that Step 1 corresponds to an L^1-regularized regression of the residual matrix $\mathbf{Y}-\hat{\mathbf{W}}_{t-1}$ on the explanatory variables. To compute the regression coefficients, we used Friedman's block-coordinate descent method [14]. According to [40], we obtained

$$\hat{\mathbf{B}}_t = \text{sign}(\bar{\mathbf{B}}_t)(\bar{\mathbf{B}}_t-\mu)_+ \tag{8}$$

where $s_+ = \max(0,s)$, $\text{sign}(s)$ is the sign of s and $\bar{\mathbf{B}}_t$ is the classical regression estimate $\bar{\mathbf{B}}_t = \mathbf{X}^T\mathbf{Y}-\hat{\mathbf{W}}_{t-1}$. Step 2 consists of finding a low rank approximation of the residual matrix $\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}_t^T$ [6]. This approximation starts with a singular value decomposition of the residual matrix $\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}_t^T = \mathbf{M}\mathbf{S}\mathbf{N}^T$, with $\mathbf{M}$ a unitary matrix

of dimension $n \times n$, $\mathbf{N}$ a unitary matrix of dimension $p \times p$, and $\mathbf{S}$ the matrix of singular values $(s_j)_{j=1..n}$. Then, we obtained

$$\hat{\mathbf{W}}_t = \mathbf{M}\bar{\mathbf{S}}\mathbf{N}^T \tag{9}$$

where $\bar{\mathbf{S}}$ is the diagonal matrix with diagonal terms $\bar{s}_j = (s_j - \gamma)_+$, $j = 1, ..., n$. Building on results from [41],the following statement holds.

Theorem 2. *Let $\mu > 0$ and $\gamma > 0$. Then the block-coordinate descent algorithm cycling through Step 1 and Step 2 converges to estimates of $\mathbf{W}$ and $\mathbf{B}$ defining a global mimimum of the penalized loss function $\mathcal{L}_{\text{lasso}}$.*

The proof of Theorem 2 can be found in the appendix. The algorithmic complexities of Step 1 and Step 2 are bounded by a term of order $O(pn + K(p + n))$. The computing time of lasso estimates is generally longer than for the ridge estimates, because the LFMM1 algorithm needs to run the SVD and projection steps several times until convergence while the ridge method (LFMM2) requires a single iteration.

Statistical tests. Suppose we test a single primary variable ($d = 1$, the extension to $d > 1$ variables is straightforward). To test association between $\mathbf{X}$ and the response variables Y_j, we used the latent score estimates obtained from the LFMM1 or LFMM2 methods as covariates in multiple linear regression models. Our approach is similar to other methods for confounder adjustment in association studies [32, 37, 27, 34, 16]. It differs from other approaches through the latent scores estimates, $\hat{\mathbf{U}}$, that capture the part of response variation not explained by the primary variable. To test for correlation with the response variable Y_j, we estimated the regression coefficients in a linear regression model

$$\mathbf{Y}_j = \mathbf{X}\beta_j + \hat{\mathbf{U}}\alpha_j^T + \mathbf{E}_j\,, \quad j = 1, \ldots, p. \tag{10}$$

To test the null hypothesis $H_0 : \beta_j = 0$, we used a Student distribution with $n - K - 1$ degrees of freedom [19]. To improve test calibration and false discovery rate estimation, we eventually applied an empirical-null testing approach to the test statistics [10].

Remark that in the above equation, causality is modeled when $\mathbf{X}$ is an exposure variable and $\mathbf{Y}$ represents a biological measure such as gene expression or DNA methylation levels. When $\mathbf{X}$ is a phenotypic trait and $\mathbf{Y}$ represents a biological measure such as a genotype, direct effect sizes can be estimated by switching the response and explanatory variables in the regression model ($\mathbf{X} = \mathbf{Y}_j\beta_j$). In addition, tests based on generalized linear models or mixed linear models could be implemented according to similar principles. In the case of mixed linear models, the covariance matrix for random effects can be computed from the K estimated factors as $C = \mathbf{U}\mathbf{U}^T/n$. The methods presented in this study and their extensions were implemented in the `R` package `lfmm`.

R package availability. The `R` package `lfmm` is available from the following URL: `https://github.com/bcm-uga/lfmm`.

3 Results and Discussion

Simulation study. In a series computer experiments, we simulated quantitative trait variables for a world-wide sample of 1,758 human individuals from the 1000 Genomes Project database [1]. Our simulations considered various levels of confounding and numbers of causal variables in the data. For each simulation setting, five data sets were created, representing a total number of 125 data sets. As a baseline, we fitted simple linear regression models (LRM) without adjusting the data for potential confounding effects. This procedure was expected to result in a severe inflation of the test statistic. Six additional association methods were applied to the simulated data: PCA, two variants of SVA, CATE[42] and two variants of LFMM. Additionnally, we implemented an *oracle* method that performed association tests aware of the generating mechanism and confounders. By using the (true) confounders as covariates in its regression model, the oracle method was expected to provide an upper bound on the power of association method that estimates confounding effects (Fig. 1A, Fig. S1). In most simulations, the power of LFMM1, LFMM2 and CATE was identical to the power of the oracle method. PCA (EIGENSTRAT) had genomic inflation factors close to one (Fig. S1), but the power of this method decreased with increasing numbers of causal loci and correlation between confounders and the primary variable. SVA methods had genomic inflation factors close to one for lower levels of confounding (Fig. S1), but their power remained lower than the oracle method for all numbers of causal loci. LFMM1, LFMM2 and CATE acheived substantially more power than SVA1, SVA2 and PCA for higher levels of confounding.

Rheumatoid arthritis (RA) EWAS. We performed an EWAS using whole blood methylation data from a study of patients with RA [28]. The cell composition of blood in RA patients is a known source of confounding, and unaccounting for cell type heterogeneity leads to an increased rate of false discoveries [23, 34]. Cross-validation identified hyperparameter values for the LFMM2 algorithm, and ten confounders were selected (Fig. S2). We implemented five methods for confounder adjustement in EWAS (Fig. 1B, Fig. S3). The resulting discoveries were compared with CpG sites detected by using a reference-based method [20, 45]. PCA, SVA1 and LFMM1 recovered 80% of the reference-based candidates within their eleven top hits, in agreement with the results of a previous analysis with REFACTOR [34]. PCA and SVA1 provided almost identical lists of candidate sites for an expected FDR of 1%. LFMM1 had higher power than PCA and SVA1, and ranked new discoveries above previously discovered candidates (Fig. 1B, Fig. S4). New discoveries included CpG sites in the genes *SPEC* and *LYN* playing an important role in the regulation of innate and adaptive immune responses, and in *HLA-DRB1* having known association with RA [25] (Fig. 1B, Table S1).

Celiac disease (CD) GWAS. Next we performed a GWAS using SNPs from a study of patients with CD [8]. In GWAS, systematic differences in allele frequencies between patients, known as population structure, are assumed to result in spurious associations and in an increased number of false positive tests [2]. Cross-validation selected high nine axes of variation in the data (Fig. S5). We implemented four methods for confounder adjustement in GWAS (Fig. 1C, Fig. S5), and we compared the regions found by those methods with the GWAS Catalog for CD. LFMM2 and PCA (EIGENSTRAT[32]) had the smallest false positive rate overall. LFMM1 and CATE had the highest power to detect regions with SNPs included in the GWAS Catalog. Pooling discoveries from all factor methods for an expected FDR level of 1% identified 282 genomic regions, containing 28% of all loci referenced in the GWAS catalog for CD (Fig. 1C, Table 1). For the most powerful method (LFMM1), six genomic regions or loci among the twenty top hits were not referenced in the GWAS

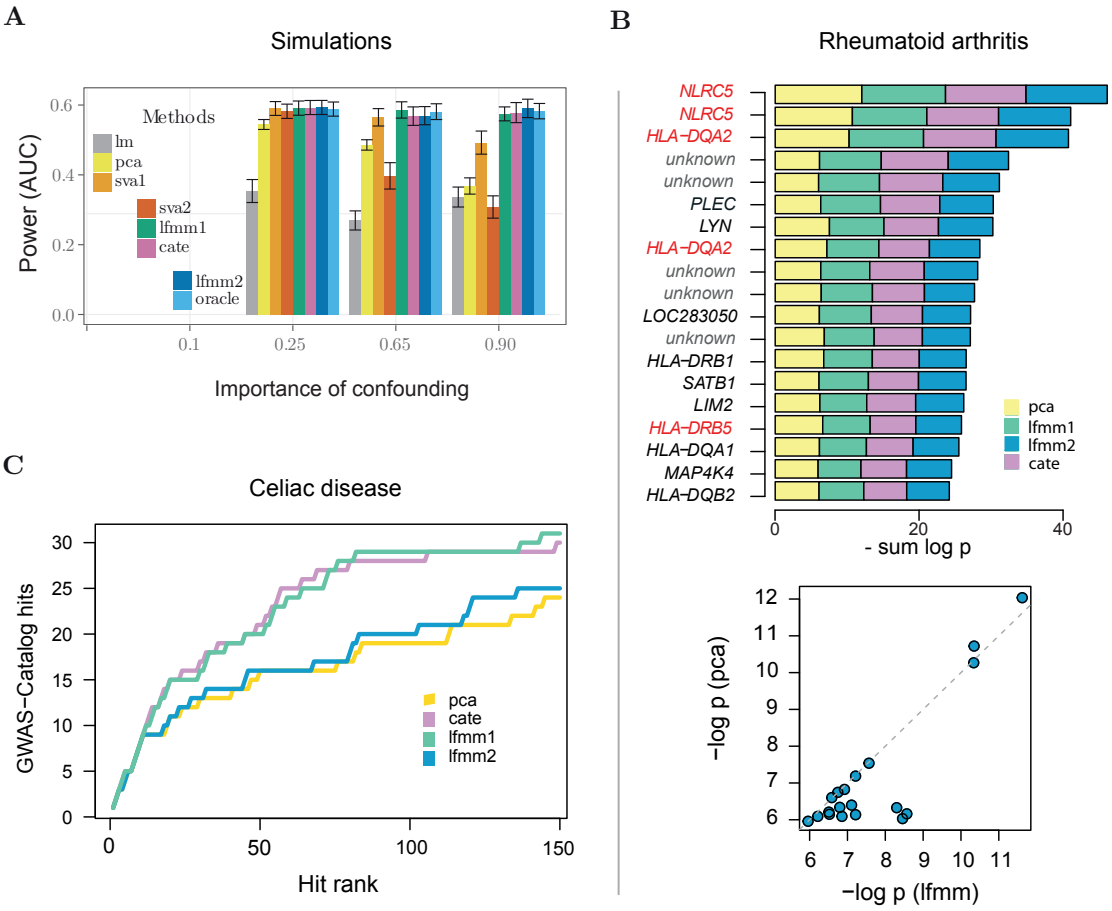

Figure 1. Power of tests. A) Simulations. Power measured by AUC for three levels of confounding. The importance of confounding corresponds to the squared correlation between confounders and primary variables. B) Rheumatoid Arthritis EWAS. Top: List of 19 methylation probes corresponding to the shared top hits of four methods: PCA, CATE, LFMM1 and LFMM2. The list was controlled for a false discovery rate of 1%. Highlighted genes correspond to previously reported discoveries. Bottom: Quantiles plot indicating that PCA correction lead to more conservative tests than methods estimating latent factors. C) Celiac disease GWAS. Ability of four methods to recover genomic regions with known associations with CD. PCA correction led to more conservative tests than methods estimating latent factors.

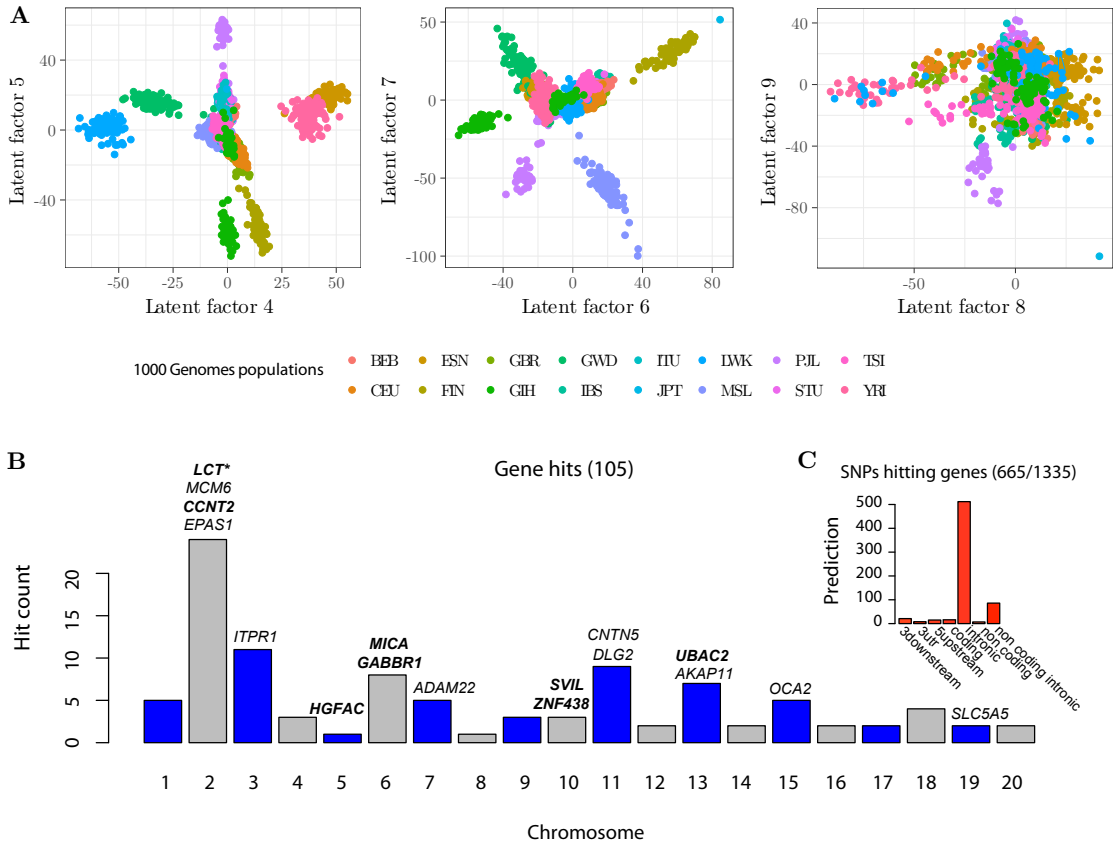

Figure 2. Gene environment association study. Association study based on genomic data from the 1000 Genomes Project database and climatic data from the Worlclim database. A) Latent factors estimated by LFMM2. B) Target genes corresponding to top hits of the GEAS analysis (expected FDR level of 5%). The highlighted genes correspond to functional variants. Predictions were obtained from the variant effect predictor program.

catalog [17] (Table 1).

Human GEAS. To detect genomic signatures of adaptation to climate in humans, we performed a GEAS using 5,397,214 SNPs for 1,409 individus from the 1,000 Genomes Project [1], and bioclimatic data from the WorldClim database [12] (Fig. S6). Nine confounders were estimated by LFMM2, mainly describing correlation between population structure and climate in the sample (Fig. 2A, Fig. S7, Fig. S8). Four methods for confounder adjustement led to a list of 836 (1335) SNPs after pooling the list of candidates from the four methods (expected FDR = 1%-5%). A variant prediction analysis reported a large number of SNPs in intergenic and intronic regions, with an over-representation of genic regions (Fig. 2B). Top hits represented genomic regions important for adaptation of humans to environmental conditions. The hits included functional variants in the *LCT* gene, and SNPs in the *EPAS1* and *OCA2* genes previously reported for their role in adaptation to diet, altitude or in eye color [11] (Fig. 2B, Fig. S9, Table S2).

Conclusions. In this study we introduced two statistical learning algorithms for confounder estimation based on L^2 and L^1-regularized least-squares problems for latent factor regression models. We used those algorithms

for testing associations between a response matrix $\mathbf{Y}$ and a primary variable matrix $\mathbf{X}$ in EWAS, GWAS and GEAS. In those applications, standard association methods have mainly focused on corrections for specific confounding effects such as individual relatedness or cell type composition. In contrast, LFMMs do not put prior information on any particular source of confounding, but account for correlation between confounders and primary variables. Compared to PCA approaches, LFMMs gained power by removing the part of genetic variation that could not be explained by the primary variables [13]. In GWAS, LFMM extends tests performed by the EIGENSTRAT program by improving estimates of principal components [32]. For EWAS, LFMMs extended surrogate variable analysis (SVA) [26] also achieving an increased power. In comparison with other algorithms, LFMM1 and LFMM2 have mathematical guarantees to provide globally optimal solutions of least-squares estimation problems, and the proposed estimates could be shown to reach oracle asymptotical efficiency for large sample sizes [42].

Like several factor methods, the computational speed of LFMM methods is mainly influenced by the algorithmic complexity of low rank approximation of large matrices. The algorithmic complexity of LFMM methods in similar to PCA or SVA, of order $O(np\ln(K))$ for LFMM1 and $O(n^2p + np\ln(K))$ for LFMM2. LFMM2 is generally faster than LFMM1 because all computations are based on a unique round of SVDs. These approaches are faster than algorithms based on mixed linear models [44] and faster than Bayesian methods currently used in GEAS [13]. Although potential improvements such as random effects, logistic regressions and stepwise conditional tests were not included in our results, those options are available with the `lfmm` program, and may provide additional power to detect true associations.

4 Materials and Methods

Multiple linear regression with principal components. We implemented a standard approach that estimates confounders from the PCA of the response matrix $\mathbf{Y}$. Scaling the response matrix, this approach is similar to the EIGENSTRAT method [32]. As a baseline, we also implemented linear regression models (LRM) that did not include correction for confounding. In the presence of confounding, LRM are expected to lead to inflation of false positive tests, whereas the PCA approach is expected to lead to overly conservative tests.

Surrogate variable analysis. Surrogate variable analysis (SVA) was introduced to overcome the problems caused by heterogeneity in gene expression studies. SVA is based on the latent factor regression model presented in this study, and is potentially useful for confounder adjustment in any type of genome-wide association studies. Two distinct SVA algorithms were implemented SVA1 [26] and SVA2 [27]. In a first step, the SVA1 algorithm estimates loadings of a PCA of the residuals of the regression of the response matrix $\mathbf{Y}$ on $\mathbf{X}$. The second step of the SVA1 algorithm determines a subset of response variables exhibiting low correlation with $\mathbf{X}$, and uses this subset of variables to estimate the score matrix. SVA1 is similar to LFMM2 with regularization term set to $\lambda = 0$, a degenerate case for the least-squares problem. The SVA2 method is an iterative approach. In SVA2, the second step of SVA1 is modified so that a weight is given to each response variable. Weights are used to compute a weighted PCA of the regression residuals, and the cycle is iterated until a convergence criterion is met. The SVA methods were implemented by using the `R` package `sva` [27].

Confounder adjusted testing and estimation. Confounder adjusted testing and estimation (CATE) [42] is a recent estimation method based on latent factor regression models. CATE uses a linear transformation

of the response matrix such that the first axis of this transformation is colinear to $\mathbf{X}$. CATE and LFMM2 apply different transformations to the response matrix, but the CATE estimates are comparable to LFMM2 estimates (although CATE estimates do not solve a least-squares problem). Asymptotic results obtained for CATE estimates were also valid for LFMM2 estimates. The CATE method was implemented in the `R` package `cate` [42].

Simulated data. We used empirical data from a world-wide sample of 1,758 human genotypes from the 1,000 Genomes Project [1] for simulating quantitative phenotypes, $\mathbf{X}$, latent factors $\mathbf{U}$ and a response matrix $\mathbf{Y}$. For the simulation, an initial data matrix, $\mathbf{Y}^\star$, including 52,211 single nucleotide polymorphisms (SNPs) from chromosomes 1 and 2 was considered. To create K articifial confounders, we first performed a PCA of the $\mathbf{Y}^\star$ matrix, and retained $K = 5$ principal components. The eigenvalues, s_k^2, were computed for each retained component. Then a primary variable $\mathbf{X}$ and five latent variables latentes, $\mathbf{U}$, were simulated by using a multivariate Gaussian distribution

$$(\mathbf{U}, \mathbf{X}) \sim \mathcal{N}(0, \mathbf{S}),$$

where $\mathbf{S}$ was the covariance matrix defined by

$$\mathbf{S} = \begin{bmatrix} s_1^2 & 0 & \cdots & \rho c_1 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & s_K^2 & \rho c_K \\ \rho c_1 & \cdots & \rho c_K & 1 \end{bmatrix}.$$

The c_k coefficients were sampled from a uniform distribution taking values in the range $(-1, 1)$, and ρ was inversely proportional to the square root of $\sum_k c_k^2/s_k^2$ (which was less than one). The coefficient of proportionality was chosen so that the percentage of variance of $\mathbf{X}$ explained by the latent factors ranged between $(0.1, 1)$. The effect size matrix, $\mathbf{B}$, was generated by setting a proportion of effect sizes to zero. Non-zero effect sizes were sampled according to a standard Gaussian distribution $N(0, 1)$. The proportion of null effect sizes ranged between 80% and 99%. We eventually created a response matrix, $\mathbf{Y}$, by simulating from the generative model of the latent factor model as follows

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^T + \mathbf{U}\mathbf{V}^T + \mathbf{E}. \tag{11}$$

In those simulations, K latent variables had the same variance as original PCs from the 1,000 Genomes Project data set, and we controlled the correlation between simulated phenotypes (primary variables) and confounders.

To evaluate the capabilities of methods to identify true positives, we used the area under the precision-power curve (AUC) as a global estimate of power. Precision is the proportion of true positives in a candidate list of positive tests. Power is the number of true positives divided by the number of true associations. To evaluate whether the methods have inflated number of false positives, we computed a genomic inflation factor using the median of squared z-scores divided by the median of the chi-squared distribution with one degree of freedom [5].

Rheumatoid arthritis (RA) data set. We performed an EWAS using whole blood methylation data from a study of patients with rheumatoid arthritis [28, 34, 45]. The RA data are publicly available and were downloaded from the GEO database (accession number GSE42861). For this study, beta-normalized methylation levels at 485,577 probed CpG sites were measured for 354 cases and 335 controls [28]. Following [45], probed CpG sites having a methylation level lower than 0.2 or greater than 0.8 were filtered out. Then, the data were centered

and scaled for a standard deviation of one. Since the cell composition of blood in RA patients typically differs from that in the general population, there is a risk for false discoveries that stem from unaccounted-for cell type heterogeneity [34]. Age, gender and covariates such as tobacco consumption may also have significant effects on DNA methylation. To evaluate whether the methods presented here can correct confounding due to those factors, we did not include them as covariates in regression analyses. Seven EWAS methods were applied to the RA data set, including LRM, PCA, two variants of SVA, CATE and two variants of LFMM. Candidate lists of CpG sites were controlled for a false discovey rate of 1% after recalibration of the test significance values, and compared to the candidates obtained with a reference-based method and controlling for age, gender and smocking status. FDR control was implemented through the `qvalue` function of the `R` program [38].

Celiac disease (CD) data set. We performed a GWAS using SNPs from a study of patients with celiac disease [8]. The CD data were downloaded from the Wellcome Trust Case Control Consortium `https://www.wtccc.org.uk/`. For this study, SNP genotypes were recorded at 485,577 loci for 4,496 cases and 10,659 controls. The genotype matrix was preprocessed so that SNPs with minor allele frequency lower than 5% and individuals with relatedness greater than 8% were removed from the matrix. We used the program BEAGLE to impute missing data in the genotype matrix [4]. We performed LD pruning to retain SNPs with the highest frequency in windows of one hundred SNPs. The filtering steps were implemented in the PLINK software [33], and resulted in a subset of 80,275 SNPs. Five GWAS methods were applied to the CD data set: LRM, PCA (EIGENSTRAT), CATE, and two LFMM estimation algorithms. For the last four methods, the confounders were identified based on the 80,275 pruned genotypes. The tests were performed on the full set of imputed genotypes, and the SNP positions were grouped into clumps of correlated SNPs, using the clumping algorithm implemented in PLINK. The significance value for a clump of SNPs was considered to be the lowest value among all positions. FDR control was applied on the clumped significance values using `qvalue`. Candidates resulting from the five analyses were compared to the GWAS catalog for known association with CD [29]. Chromosome 6, which contained the strongest association signals with CD, was treated separately. For this chromosome, all methods performed equally well at detecting six SNPs from the HLA locus referenced in the GWAS catalog.

Gene-environment association study. We performed a GEAS using whole genome sequencing data and bioclimatic variables to detect genomic signatures of adaptation to climate in humans. The data are publicly available, and they were downloaded from the 1,000 Genomes Project phase 3 [1] and from the WorldClim database [12]. The genomic data included 84.4 millions of genetic variants genotyped for 2,506 individuals from 26 world-wide human populations. Nineteen bioclimatic data were downloaded for each individual geographic location, considering capital cities of their country of origin. The bioclimatic data were summarized by projection on their first principal component axis. The genotype matrix was preprocessed so that SNPs with minor allele frequency lower than 5% and individuals with relatedness greater than 8% were removed from the matrix. Admixed individuals from Afro-american and Afro-Caribbean populations were also removed from the data set. After those filtering steps, the response matrix contained 1,409 individus and 5,397,214 SNPs. We performed LD pruning to retain SNPs with the highest frequency in windows of one hundred SNPs, and identified a subset of 296,948 informative SNPs. Four GEAS methods were applied to the 1,000 Genomes Project data set: PCA (EIGENSTRAT), CATE, and two LFMM estimation algorithms. For all methods the latent factors were estimated from the pruned genotypes, and association tests were performed for all 5,397,214 loci. Candidates obtained from clumps with an expected FDR level of 1% were analyzed using the Variant Effect Predictor

(VEP) program [30].

Cross-validation and model choice. Choosing regularization parameters and the number of latent factors can be achieved by using cross-validation methods. We developed a cross-validation approach appropriate to latent factor regression models. Cross-validation partitions the data into a training set and a test set. The training set is used to fit model parameters, and prediction errors can be measured on the test set. In our approach, the reponse and explanatory variables were partitioned according to their rows (individuals). We denote by I the subset of individual labels on which prediction errors are computed. Estimates of effect sizes, $\hat{\mathbf{B}}_{-I}$, and factor loadings, $\hat{\mathbf{V}}_{-I}$, were obtained from the training set. Next, the set of columns of the response matrix were partitioned. Denoting by J the subset of columns on which the prediction errors were computed, a score matrix was estimated from the complementary subset as follows

$$\hat{\mathbf{U}}_{-J} = (\mathbf{Y}[I, -J] - \mathbf{X}[I,](\hat{\mathbf{B}}_{-I}[-J,])^T)\hat{\mathbf{V}}_{-I}[-J,]^T. \tag{12}$$

In this notation, the brackets indicate which subsets of rows and columns of a matrix were selected. A prediction error was computed as follows

$$\text{Error} = \frac{1}{\#I\#J}\left\|\mathbf{Y}[I,J] - \hat{\mathbf{U}}_{-J}\hat{\mathbf{V}}_{-I}[J,]^T - \mathbf{X}[I,]\hat{\mathbf{B}}_{-I}[J,]^T\right\|_F. \tag{13}$$

Parameters leading to the lowest prediction errors were retained for data analysis.

Additional heuristics were used to determine the number of latent factors and the nuclear norm parameter for LFMM1 latent matrix estimates. For choosing the number of latent factors, K, we considered the matrix $\mathbf{D}_\lambda$ defined in the statement of Theorem 1, and the $\mathbf{Q}$ unitary matrix obtained from the SVD of $\mathbf{X}$. The number of latent factors, K, was estimated after a spectral analysis of the matrix $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$. We determined it by estimating the number of components in a PCA of the matrix $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$. In our experiments, we used the "elbow" method based on the scree-plot of PC eigenvalues. Estimated values for K were confirmed based on prediction errors computed by cross-validation. The L^1-regularization parameter, μ, was determined after the proportion of non-zero effect sizes in the $\mathbf{B}$ matrix, which was estimated by cross-validation. Having set the proportion of non-zero effect sizes, μ was computed by using the regularization path approach proposed in [15]. The regularization path algorithm was initialized with the smallest values of μ such that

$$\hat{\mathbf{B}}_1 = \text{sign}(\bar{\mathbf{B}}_1)(\bar{\mathbf{B}}_1 - \mu)_+ = 0, \tag{14}$$

where $\hat{\mathbf{B}}_1$ resulted from Step 1 in the lasso (LFMM1) estimation algorithm. Then, we built a sequence of μ values that decreases from the inferred value of the parameter, $\mu^{\max}$, to $\mu^{\min} = \epsilon\mu^{\max}$. We eventually computed the number of non-zero elements in $\hat{\mathbf{B}}_t$, and stopped when the target proportion was reached. The nuclear norm parameter (γ) determines the rank of the latent matrix $\mathbf{W}$. We used a heuristic approach to evaluate γ from the number of latent factors K. The singular values $(\lambda_1, \ldots, \lambda_n)$ of the response matrix $\mathbf{Y}$ were computed, and we set

$$\gamma = \frac{(\lambda_K + \lambda_{K+1})}{2}. \tag{15}$$

In our experiments, the lasso estimation algorithm always converged to a latent matrix estimate having rank K.

Table 1. CD GWAS. Genomic regions corresponding to the first twenty hits of the LFMM1 algorithm (SNP Ids). Rows in bold style correspond to SNPs referenced in the GWAS Catalog for a previously reported association with CD. Chromosome 6 was not included in the analysis.

Chr	SNP Id	LD block (Mb)	Odds ratio [95% CI]	P value	Q value	Genes
3	**rs1464510**	**189.56-189.61**	**1.30 [1.24-1.36]**	**3.8e-23**	**1.5e-20**	***LPP***
3	**rs17810546**	**160.99-161.32**	**1.35 [1.26-1.45]**	**1.8e-16**	**6.1e-14**	***IQCJ-SCHIP1, IL12A-AS1, IL12A***
4	**rs13151961**	**123.19-123.56**	**0.73 [0.68-0.78]**	**1.7e-14**	**5.3e-12**	***KIAA1109, ADAD1***
12	**rs653178**	**110.25-110.49**	**1.22 [1.16-1.28]**	**6.8e-13**	**2.1e-10**	***CUX2, LINC02356, SH2B3, ATXN2***
2	**rs917997**	**102.26-102.61**	**1.27 [1.20-1.35]**	**1.5e-12**	**4.6e-10**	***IL1RL1, IL18R1, IL18RAP, MIR4772, SLC9A4, SLC9A2***
4	rs6840978	123.73-123.77	0.77 [0.72-0.82]	1.2e-11	3.5e-09	*IL21-AS1*
3	rs9811792	161.12-161.18	1.18 [1.12-1.24]	6.6e-11	1.9e-08	*IL12A-AS1*
3	**rs13098911**	**45.98-46.21**	**1.32 [1.22-1.43]**	**2.1e-10**	**5.8e-08**	***FYCO1, FLT1P1, CCR3***
1	**rs2816316**	**190.77-190.80**	**0.78 [0.72-0.83]**	**2.2e-10**	**6.2e-08**	
3	**rs6441961**	**46.26-46.33**	**1.21 [1.15-1.27]**	**1.7e-08**	**4.6e-06**	***CCR3, UQCRC2P1***
2	**rs4675374**	**204.29-204.52**	**1.23 [1.16-1.31]**	**2.1e-07**	**5.4e-05**	***CD28, ICOS***
2	rs1018326	181.54-181.78	1.15 [1.10-1.21]	4.4e-07	1.1e-04	*UBE2E3, LINC01934*
3	rs7648827	46.56-46.56	1.22 [1.12-1.33]	4.6e-07	1.2e-04	*LRRC2*
2	**rs13003464**	**60.95-61.24**	**1.19 [1.13-1.25]**	**5.0e-07**	**1.3e-04**	***LINC01185, REL, PUS10, RNA5SP95, KIAA1841, C2orf74***
10	**rs1250552**	**80.71-80.74**	**0.84 [0.80-0.88]**	**5.2e-07**	**1.3e-04**	***ZMIZ1***
3	rs7629708	189.56-189.62	1.17 [1.11-1.24]	1.0e-06	2.5e-04	*LPP*
22	**rs2298428**	**20.13-20.31**	**1.17 [1.10-1.24]**	**1.3e-06**	**3.3e-04**	***HIC2, UBE2L3, YDJC, CCDC116***
18	rs1394466	48.93-49.30	1.14 [1.08-1.20]	1.5e-06	3.6e-04	*DCC*
18	**rs1893217**	**12.80-12.84**	**1.16 [1.08-1.23]**	**1.6e-06**	**4.0e-04**	***PTPN2***
1	**rs864537**	**165.66-165.70**	**0.87 [0.83-0.92]**	**1.7e-06**	**4.2e-04**	***POU2F1, CD247***

Appendix

In this section, we provide proofs for Theorems 1 and 2. For Theorem 1, we define the singular value decomposition (SVD) of the explanatory matrix, as $\mathbf{X} = \mathbf{Q}\mathbf{\Sigma}\mathbf{R}^T$, where $\mathbf{Q}$ is an $n \times n$ unitary matrix, $\mathbf{R}$ is a $d \times d$ unitary matrix and $\mathbf{\Sigma}$ is an $n \times d$ matrix containing the singular values of $\mathbf{X}$, $(\sigma_j)_{j=1..d}$. Let $\lambda > 0$, and the estimates given by

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\mathrm{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y}) \tag{16}$$

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T), \tag{17}$$

where $\mathbf{D}_\lambda$ is the $n \times n$ diagonal matrix with diagonal terms

$$(\mathbf{D}_{\lambda,i,i})_{i=1..n} = \left(\sqrt{\frac{\lambda}{\lambda+\sigma_1^2}}, \ldots, \sqrt{\frac{\lambda}{\lambda+\sigma_d^2}}, 1, \ldots, 1\right).$$

We prove that those estimates define a global mimimum of the function $\mathcal{L}_{\text{ridge}}$.

Proof. If we assume $\mathbf{U}$ are $\mathbf{V}$ to be known, then the $\mathcal{L}_{\text{ridge}}$ function is convex with respect to the variable $\mathbf{B}$. A global minimum for this variable can be found by computing the derivative of $\mathcal{L}_{\text{ridge}}$ with respect to $\mathbf{B}$ and setting it to zero. This leads to the following solution

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \mathbf{U}\mathbf{V}^T). \tag{18}$$

This solution is merely the ridge estimate for a linear regression of the response matrix $\mathbf{Y} - \mathbf{U}\mathbf{V}^T$ on $\mathbf{X}$. Thus, the problem amounts to minimizing the (implicit) function

$$\mathcal{L}'(\mathbf{U}, \mathbf{V}) = \mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, \hat{\mathbf{B}}), \tag{19}$$

where $\hat{\mathbf{B}}$ was defined above. Consider the SVD of $\mathbf{X}$

$$\mathbf{X} = \mathbf{Q}\mathbf{\Sigma}\mathbf{R}^T, \tag{20}$$

where $\mathbf{Q}$ is a unitary matrix of dimensions $n \times n$, $\mathbf{R}$ is a unitary matrix of dimensions $d \times d$, and $\mathbf{\Sigma}$ is a matrix of dimensions $n \times d$ containing the singular values $(\sigma_j)_{j=1..d}$. The $\mathcal{L}'$ function rewrites as follows

$$\mathcal{L}'(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \left\| \mathbf{D}_\lambda^2 \mathbf{Q}^T (\mathbf{Y} - \mathbf{U}\mathbf{V}^T) \right\|_F^2 + \frac{1}{2} \lambda \left\| \mathbf{C}_\lambda \mathbf{Q}^T (\mathbf{Y} - \mathbf{U}\mathbf{V}^T) \right\|_F^2 \tag{21}$$

where $\mathbf{C}_\lambda$ is a matrix of dimensions $d \times n$, with zero coefficients except for the diagonal terms

$$\{\mathbf{C}_{\lambda,i,i}\}_{i=1..d} = \left\{ \frac{\sigma_i}{\sigma_i^2 + \lambda} \right\}_{i=1..d}. \tag{22}$$

The $\mathbf{D}_\lambda$ is a diagonal matrix of dimensions $n \times n$ such that

$$\{\mathbf{D}_{\lambda,i,i}\}_{i=1..n} = \left\{ \sqrt{\frac{\lambda}{\lambda + \sigma_1^2}}, \dots, \sqrt{\frac{\lambda}{\lambda + \sigma_d^2}}, 1, \dots, 1 \right\}. \tag{23}$$

Direct calculus shows that we have

$$\begin{aligned} \mathcal{L}'(\mathbf{U}, \mathbf{V}) &= \frac{1}{2} \left\| \sqrt{(\mathbf{D}_\lambda^2 + \mathbf{C}_\lambda^2)} \mathbf{Q}^T (\mathbf{Y} - \mathbf{U}\mathbf{V}^T) \right\|_F^2 \\ &= \frac{1}{2} \left\| \mathbf{D}_\lambda \mathbf{Q}^T (\mathbf{Y} - \mathbf{U}\mathbf{V}^T) \right\|_F^2 . \end{aligned}$$

This equation shows that mimimizing the objective function $\mathcal{L}'$ is equivalent to finding the best approximation of rank K for $\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}$. According to [9], this solution is given by the rank K SVD of $\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}$. Eventually, this concludes the proof that

$$\begin{aligned} \hat{\mathbf{U}}\hat{\mathbf{V}}^T &= \mathbf{Q}\mathbf{D}_\lambda^{-1} \text{svd}_K(\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}) \\ \hat{\mathbf{B}}^T &= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{Id}_d)^{-1} \mathbf{X}^T (\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T) \end{aligned}$$

defined a global minimum for the $\mathcal{L}_{\text{ridge}}$ function. □

Now, turn to the proof of Theorem 2. Let $\mu > 0$ and $\gamma > 0$. Then Theorem2 states that the block-coordinate descent algorithm converges to estimates of $\mathbf{W}$ and $\mathbf{B}$ defining a global mimimum of the function $\mathcal{L}_{\text{lasso}}$.

Proof. The result is a consequence of the convexity of $\mathcal{L}_{\text{lasso}}$ function, and the fact that we can write

$$\mathcal{L}_{\text{lasso}}(\mathbf{B}, \mathbf{W}) = g(\mathbf{B}, \mathbf{W})/2 + f_1(\mathbf{B}) + f_2(\mathbf{W})$$

where $g(\mathbf{B}, \mathbf{W}) = \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2$ is a differentiable convex function, and $f_1(\mathbf{B}) = \|\mathbf{B}\|_1^2$, $f_2(\mathbf{W}) = \|\mathbf{W}\|_*^2$ are continuous convex functions. The proof of Theorem 2 relies on the following proposition adapted from [3] and [41].

Proposition. Let $A = A_1 \times A_2 \times ... \times A_m$ be a Cartesian product of closed convex sets. Consider a continuous convex function f defined on A as follows

$$f(x_1, ..., x_m) = g(x_1, ..., x_m) + \sum_{i=1}^{m} f_i(x_i), \tag{24}$$

where g is a differentiable convex function, and for all i, f_i is a continuous convex function. Let (x^{t+1}) be the sequence of values defined by the block-coordinate descent algorithm

$$x_i^{t+1} \in \arg\min_{\chi \in A_i} f(x_1^t, \ldots, x_{i-1}^t, \chi, x_{i+1}^t, \ldots, x_m^t), \quad i = 1, ..., m. \tag{25}$$

Then a limit point of (x^t) defines a global minimum of f. □

Accession codes. SNP genotypes used in our simulation analysis are publicly available and were downloaded from the 1000 Genome Project database. The RA data are publicly available and were downloaded from the GEO database (accession number GSE42861). The CD data are publicly available and were downloaded from the Welcome Trust Case Control Consortium database (agreement number 1248).

Acknowledgments. This work has been supported by a grant from LabEx PERSYVAL Lab, ANR-11-LABX-0025-01, funded by the French program Investissement d'Avenir.

Author contributions. K.C. performed research, contributed analytic tools, and analyzed data. O.F. designed research, contributed analytic tools, analyzed data, and wrote the paper.

References

[1] The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526** 68-74.

[2] Balding, D.J. (2006) A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **7** 781-781.

[3] Bertsekas, D. P. (1999) *Nonlinear Programming*. Belmont: Athena Scientific.

[4] Browning, B. L. and Browning, S. R. (2016) Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J Hum. Genet.*, **98** 116-126.

[5] Devlin, B. and Roeder K. (1999) Genomic control for association studies *Biometrics*, **55** 997-1004.

[6] Cai, J-F., Candès, E.J. and Shen, Z. (2010) A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, **20** 1956-1982.

[7] Carvalho, C. M. *et al.* (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Stat. Assoc.*, **103** 1438-1456.

[8] Dubois, P.C.A. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42** 295-302.

[9] Eckart, C. and Young, G. (1936) The approximation of one matrix by another of lower rank. *Psychometrika*, **1** 211-218.

[10] Efron, B. (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Am. Stat. Assoc.*, **99** 96-104.

[11] Fan, S., Hansen, M. E., Lo, Y. and Tishkoff, S. A. (2016) Going global by adapting local: A review of recent human adaptation. *Science*, **354** 54-59.

[12] Fick, S. E. and Hijmans, R. J. (2017) Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.*, **37** 4302-4315.

[13] Frichot, E., Schoville, S. D., Bouchard, G. and François, O. (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.*, **30** 1687-1699.

[14] Friedman, J. *et al.* (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1** 302-332.

[15] Friedman, J., Hastie, T., and Tibshirani, T. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**.

[16] Gerard, D. and Stephens, M. (2017) Empirical Bayes shrinkage and false discovery rate estimation allowing for unwanted variation. *arXiv* preprint arXiv:1709.10066.

[17] van Heel, D.A. *et al.* (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.*, **39** 827-829.

[18] Halko, N., Martinsson, P. G. and Tropp, J. A. (2011) Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, **53** 217-288.

[19] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics, Springer, NY, USA.

[20] Houseman, E. A. *et al.* (2012) Reference-free cell mixture adjustments in analysis of DNA methylation data. *BMC Bioinformatics*, **13** 86.

[21] Houseman, E. A., Molitor, J. and Marsit, C. J. (2014) *Bioinformatics*, **30** 1431-1439.

[22] van Iterson, M., van Zwet, E. W. and Heijmans, B. T. (2017) Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.*, **18**, 19.

[23] Jaffe, A. E. and Irizarry, R. A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, **15** R3.

[24] Kaushal, A. *et al.* (2017) Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinformatics*, **18** 216.

[25] Kurkó J. *et al.* (2013) Genetics of rheumatoid arthritis - a comprehensive review. *Clin. Rev. Allergy. Immunol.*, **45**170-179.

[26] Leek, J. T. and Storey, J. D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3** e161.

[27] Leek, J. T. and Storey, J. D. (2008) A general framework for multiple testing dependence. *Proc. Nat. Acad. Sci. USA*, **105** 18718-18723.

[28] Liu, Y. *et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31** 142-147.

[29] MacArthur, J. *et al.* (2016) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45** D896-D901.

[30] McLaren, W. *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17** 122.

[31] Mishra, B., Meyer, G., Bach, F. and Sepulchre, R. (2013) Low-rank optimization with trace norm penalty. *SIAM J. Optim.*, **23** 2124-2149.

[32] Price, A.L. *et al.* (2006) Principal component analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38** 904-909.

[33] Purcell, S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J Hum. Genet.*, **81** 559-575.

[34] Rahmani, E. *et al.* (2016) Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* **13** 443-445.

[35] Rakyan, V. K., Down, T. A., Balding, D. J. and Beck, S. (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, **12** 529-541.

[36] Savolainen, O., Lascoux, M. and Merilä, J. (2013) Ecological genomics of local adaptation. *Nat. Rev. Genet.*, **14** 807-820.

[37] Song, M., Hao, W. and Storey, J.D. (2015) Testing for genetic associations in arbitrarily structured populations. *Nat. Genet.*, **47** 550-554.

[38] Storey, J.D., Bass, A.J., Dabney A. and Robinson D. (2015) qvalue: Q-value estimation for false discovery rate control. R package version 2.6.0.`http://github.com/jdstorey/qvalue`

[39] Teschendorff, A.E. and Relton, C.L. (2017) Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* doi:10.1038/nrg.2017.86

[40] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58** 267-288.

[41] Tseng, P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theor. Appl.*, **109** 475-494.

[42] Wang, J., Zhao, Q., Hastie, T. and Owen, A.B. (2017) Confounder adjustment in multiple hypothesis testing. *Ann. Statist.*, **45** 1863-1894.

[43] Zheng, S. C. *et al.* (2017) Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. *Nat. Methods*, **14** 216.

[44] Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **11** 407-409.

[45] Zou, J., Lippert, C., Heckerman, D., Aryee, M. and Listgarten, J. (2014) Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods*, **11** 309-311.