Sparse latent factor regression models for genome-wide and epigenome-wide association studies

Basile Jumentier[1,3] Kevin Caye[1] Barbara Heude[2]
Johanna Lepeule[3,⋆] Olivier François[1,⋆]

Authors' affiliations:

[1] Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Grenoble INP, TIMC-IMAG CNRS UMR 5525, 38000 Grenoble, France.

[2] Université de Paris, Institut National de la Santé et de la Recherche Médicale, Centre of Research in Epidemiology and Statistics, INSERM UMR 1153, F75004 Paris, France.

[3] Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale, Institute for Advanced Biosciences, INSERM U 1209, CNRS UMR 5309, 38000 Grenoble, France.

⋆ Corresponding authors: `johanna.lepeule@inserm.fr` , `olivier.francois@imag.fr`

Abstract

Association of phenotypes or exposures with genomic and epigenomic data faces important statistical challenges. One of these challenges is to account for variation due to unobserved confounding factors, such as individual ancestry or cell-type composition in tissues. This issue can be addressed with penalized latent factor regression models, where penalties are introduced to cope with high dimension in the data. If a relatively small proportion of genomic or epigenomic markers correlate with the variable of interest, sparsity penalties may help to capture the relevant associations, but the improvement over non-sparse approaches has not been fully evaluated yet. Here, we present least-squares algorithms that jointly estimate effect sizes and confounding factors in sparse latent factor regression models. In simulated data, sparse latent factor regression models generally achieved higher statistical performance than other sparse methods, including the least absolute shrinkage and selection operator (LASSO) and a Bayesian sparse linear mixed model (BSLMM). In generative model simulations, statistical performance was slightly lower (while being comparable) to non-sparse methods, but in simulations based on empirical data, sparse latent factor regression models were more robust to departure from the model than the non-sparse approaches. We applied sparse latent factor regression models to a genome-wide association study of a flowering trait for the plant *Arabidopsis thaliana* and to an epigenome-wide association study of smoking status in pregnant women. For both applications, sparse latent factor regression models facilitated the estimation of non-null effect sizes while overcoming multiple testing issues. The results were not only consistent with previous discoveries, but they also pinpointed new genes with functional annotations relevant to each application.

1 Introduction

Association studies represent one of the most powerful tool to identify genomic variation correlating with disease states, exposure levels or phenotypes. Those studies are divided into several categories according to the nature of the genomic markers evaluated. For example, genome-wide association studies (GWAS) focus on single-nucleotide polymorphisms in different individuals to estimate disease allele effects (Balding, 2006), while epigenome-wide association studies (EWAS) measure epigenetic marks, such as DNA methylation levels to derive associations between epigenetic variation and exposure levels or to assess effects on phenotypic traits (Rakyan et al., 2011). Despite their success in identifying the genetic architecture of phenotypic traits or genomic targets of exposure, association studies are plagued with the problem of confounding, which arises when unobserved variables correlate with the variable of interest and with the genomic markers simultaneously (Wang et al., 2017).

Historical approaches to the confounding issue account for hidden confounders by considering corrections for inflation (Devlin and Roeder, 1999) and empirical null-hypothesis testing methods (Efron, 2004). Alternative approaches evaluate hidden confounders by using linear combinations of observed variables, often called factors. In GWAS, a frequently-used factor approach consists of computing the largest principal components of the genotype matrix, and includes them as covariates in linear regression models (Price et al., 2006). The variable of interest may, however, be collinear to the largest principal components, and removing their effects can result in loss of statistical power. To overcome this issue and increase power, methods based on latent factor regression models have been proposed (Leek and Storey, 2007; Carvalho et al., 2008). Latent factor regression models employ deconvolution methods

in which unobserved variables, including batch effects, individual ancestry or tissue cell-type composition are integrated in the regression model by using latent factors. In these models, effect sizes and latent factors are estimated jointly. The latent factor regression framework encompasses several methods which include surrogate variable analysis (SVA, (Leek and Storey, 2007)), latent factor mixed models (LFMM, (Frichot et al., 2013)), residual principal component analysis (Kalaitzis and Lawrence, 2012), direct surrogate variable analysis (Lee et al., 2017), robust reduced-rank regression (She and Chen, 2017) and confounder adjusted testing and estimation (CATE, Wang et al. (2017)). Each method has specific merits relative to some category of association study, and the performances of the methods have been extensively debated in recent surveys (for example, see Kaushal et al. (2017)).

While GWAS and EWAS primarily report significance values, locus-specific effect sizes are summary statistics of crucial importance in interpretation and application of the results (Buniello et al., 2019; Battram et al., 2021). A common property of latent factor regression models is to use regularization parameters inducing constraints on effect size estimates. Among those methods, sparse regression models suppose that a relatively small proportion of all genomic variables correlate with the variable of interest or affect the phenotype, and evaluate associations while avoiding multiple testing problems (Tibshirani, 1996; Hoggart et al., 2008; Wu et al., 2009). Sparse regression models have also been coupled with linear mixed models to combine the benefits of both for polygenic trait studies with Bayesian sparse linear mixed model (BSLMM) (Zhou and Stephens, 2012; Zhou et al., 2013).

In this study, we introduce least-squares algorithms that jointly estimate effect sizes and confounding factors in sparse latent factor regression models. Effect sizes

are estimated based on regularized least-squares methods with L^1 and nuclear norm penalties. Thanks to the inclusion of sparsity constraints, the proposed algorithm allows identifying non-null effect sizes without the use of multiple statistical tests. We refer to our approach as *sparse latent factor mixed models* or *sparse* LFMM. We present estimation algorithms for sparse LFMM and theoretical results in the next section. Then we compare the performances of sparse LFMM with other sparse regression models (LASSO, BSLMM) and with non-sparse regression models (SVA, CATE, LFMM with ridge penalty). To illustrate our approach, we used sparse LFMM to perform a GWAS of flowering time for the plant *Arabidopsis thaliana* and to perform an EWAS of smoking status in pregnant women.

2 Latent factor regression models

2.1 Models

Latent factor regression models evaluate associations between the elements of a response matrix, $\mathbf{Y}$, and variables of interest, called *primary* variables, $\mathbf{X}$, measured for n individuals. The response matrix records p markers, which can represent any type of omic data (genotypes, DNA methylation, etc), collected for the individuals. The $\mathbf{X}$ matrix can also incorporate observed confounders such as age, sex, etc, and its dimension is $n \times d$, where d represents the total number of variables. Note that latent factor regression models and latent factor mixed models are synonymous terms. Latent factor mixed models contain fixed and latent effects, and those models should not be confused with linear mixed models, which contain fixed and random effects.

Latent factor regression models combine fixed and latent effects as follows

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

Fixed effect sizes are recorded in the $\mathbf{B}$ matrix, which has dimension $p \times d$. The $\mathbf{E}$ matrix represents residual errors, and has the same dimension as the response matrix. The matrix $\mathbf{W}$ is a *latent matrix* of rank K, defined by K latent factors (Leek and Storey, 2007; Frichot et al., 2013; Wang et al., 2017; Lee et al., 2017). The value of K is unknown, and it is generally determined by model choice procedures. The K latent factors, $\mathbf{U}$, form a matrix of dimension $n \times K$, defined from the singular value decomposition of the latent matrix $\mathbf{W} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{V}$ is a semi-orthogonal matrix of dimension $p \times K$ (Eckart and Young, 1936). The matrices $\mathbf{U}$ and $\mathbf{V}$ are unique up to a change of sign. At the exception of the next paragraph motivating the use of penalized loss functions, this study will, however, not use the decomposition the latent matrix into factor and loading matrices.

Naive statistical estimates for the $\mathbf{B}$ and $\mathbf{W}$ matrices in equation (1) could be obtained through the minimization of a classical least-squares loss function

$$\mathcal{L}(\mathbf{B}, \mathbf{W}) = \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2 \,, \tag{2}$$

where $\|.\|_F$ is the Frobenius matrix norm. A minimum value of the loss function is attained when $\mathbf{W}$ is computed as the rank K singular value decomposition of $\mathbf{Y}$ or principal component analysis. In this case, the $\mathbf{B}$ matrix could be obtained as the estimates of a linear regression of the residual matrix $(\mathbf{Y} - \mathbf{W})$ on $\mathbf{X}$. To motivate the introduction of regularization terms in the loss function, we remark that the interpretation of the principal components as confounder estimates may be incorrect,

because they fail to include any information on the primary variable, and the above construction is problematic. To see it, consider any matrix $\mathbf{P}$ with dimensions $d \times p$ and check that

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

As a consequence, both $\mathbf{B}$ and $(\mathbf{B} - \mathbf{VP}^T)$ correspond to valid solutions of the minimization problem, showing there is an infinite space of solutions. To conclude, the loss function needs to be modified in order to warrant dependency of $\mathbf{W}$ on both $\mathbf{Y}$ and $\mathbf{X}$, and to enable the computation of well-defined solutions.

2.2 Sparse estimation algorithms

L^1-regularized least-square problem. To solve the problems outlined in the above section, a sparse regularization approach is considered. This approach introduces penalties based on the L^1 norm of the regression coefficients and on the nuclear norm of the latent matrix

$$\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B}) = \left\|\mathbf{Y} - \mathbf{W} - \mathbf{XB}^T\right\|_F^2 + \mu\|\mathbf{B}\|_1 + \gamma\|\mathbf{W}\|_*, \quad \mu, \gamma > 0, \tag{3}$$

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 penalty induces sparsity on the fixed effects (Tibshirani, 1996), and 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 is introduced to penalize

large numbers of latent factors. With these penalty terms, $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$ is a convex function, and convex minimization algorithms can be applied to obtain estimates of $\mathbf{B}$ and $\mathbf{W}$ (Mishra et al., 2013).

Sparse latent factor mixed model algorithm. Like Tibshirani (1996), let us assume that the primary variables, $\mathbf{X}$, are scaled so that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ ($\mathbf{I}$ is the $d \times d$ identity matrix). Note that the extension to any set of primary variables is not difficult, and that our program implementation is general. We developed a block-coordinate descent method for minimizing the convex loss function $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$ with respect to $\mathbf{B}$ and $\mathbf{W}$. The algorithm is initialized from the 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

$$\mathcal{L}^{(1)}_{\text{sparse}}(\mathbf{B}) = \|(\mathbf{Y} - \hat{\mathbf{W}}_{t-1}) - \mathbf{X}\mathbf{B}^T\|_F^2 + \mu\|\mathbf{B}\|_1 \,, \tag{4}$$

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

$$\mathcal{L}^{(2)}_{\text{sparse}}(\mathbf{W}) = \|(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T) - \mathbf{W}\|_F^2 + \gamma\|\mathbf{W}\|_* . \tag{5}$$

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 the Friedman block-coordinate descent method (Friedman et al., 2007). According to (Tibshirani, 1996), we obtained

$$\hat{\mathbf{B}}_t = \text{sign}(\bar{\mathbf{B}}_t)(\bar{\mathbf{B}}_t - \mu)_+ \,, \tag{6}$$

where $s_+ = \max(0, s)$, $\text{sign}(s)$ is the sign of s, and $\bar{\mathbf{B}}_t$ is the linear 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 $\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T$ (Cai et al., 2008). This approximation starts with a singular value decomposition (SVD) 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 $(s_j)_{j=1,\ldots,n}$. Then, we obtain

$$\hat{\mathbf{W}}_t = \mathbf{M}\bar{\mathbf{S}}\mathbf{N}^T \tag{7}$$

where $\bar{\mathbf{S}}$ is the diagonal matrix with diagonal terms $\bar{s}_j = (s_j - \gamma)_+$, $j = 1, \ldots, n$. Building on results from (Tseng, 2001), the following statement holds.

Theorem 1. *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 minimum of the penalized loss function $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$.*

Note that 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 sparse LFMM estimates is generally longer than for the CATE algorithm (Wang et al., 2017) or the ridge LFMM algorithm detailed below (Caye et al., 2019). Sparse LFMM needs to perform SVD and projections several times until convergence while CATE and ridge LFMM require a single iteration.

2.3 Ridge regression algorithms

Latent factor regression algorithms with ridge penalties were first considered in (Caye et al., 2019). This approach was referred to as ridge LFMM. Since sparse LFMM and

ridge LFMM are implemented in the same computer package, we recall how estimates of the parameter matrices $\mathbf{B}$ and $\mathbf{W}$ are computed for ridge LFMM. This approach minimizes the following loss function

$$\mathcal{L}_{\text{ridge}}(\mathbf{B}, \mathbf{W}) = \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2 + \lambda\|\mathbf{B}\|_2^2\,, \quad \lambda > 0, \tag{8}$$

where $\|.\|_F$ is the Frobenius norm, $\|.\|_2$ is the L^2 norm, and λ is a regularization parameter. The minimization algorithm starts with an 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 an $d \times d$ unitary matrix and $\mathbf{\Sigma}$ is an $n \times d$ matrix containing the singular values of $\mathbf{X}$, denoted by $(\sigma_j)_{j=1,\ldots,d}$. The ridge LFMM estimates are computed as follows

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\text{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y}) \tag{9}$$

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{W}}), \tag{10}$$

where $\text{svd}_K(\mathbf{A})$ is the SVD of rank K of $\mathbf{A}$, $\mathbf{I}$ is the $d \times d$ identity matrix, and $\mathbf{D}_\lambda$ is the $n \times n$ diagonal matrix with coefficients defined as

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

For $\lambda > 0$, the solution of the regularized least-squares problem is unique (Caye et al., 2019), and the corresponding matrices are called the *ridge estimates*. For completeness, we provide a short proof for this result, stated in (Caye et al., 2019), in the appendix. Using random projections to compute low rank approximations, the complexity of the ridge LFMM algorithm is of order $O(n^2p + np\log K)$ (Halko et al., 2011). For studies in which the number of samples, n, is much smaller than

the number of response variables, p, computing times of ridge estimates are therefore faster than those of sparse LFMM.

2.4 Choice of hyperparameters

Number of factors. In order to choose the number of latent factors, K, we considered the matrix $\mathbf{D}_\lambda$, defined for the LFMM ridge algorithm, and the unitary matrix $\mathbf{Q}$, obtained from an SVD of $\mathbf{X}$. The number of latent factors, K, was estimated by using spectral analysis of the matrix $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$. In real data analyses, we used the scree plot of eigenvalues of the matrix $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$, Onalski's methods based on difference in eigenvalues available in the `R` package `cate` (Onatski, 2010), and Leek's method available in the `R` package `sva` (Leek, 2011).

Regularization parameters. The L^1-regularization parameter, μ, was determined by estimating the proportion of null effect sizes. This proportion was estimated as the proportion of null p-values from statistical tests performed with ridge LFMM by using the function `pi0est` from the package `qvalue` (Storey et al., 2021). In ridge LFMM, the regularization parameter was set to the default value of $\lambda = 10^{-5}$. Having set the proportion of non-null effect sizes, μ was computed by using the *regularization path* approach proposed by Friedman et al. (2010) as follows. The regularization path algorithm was initialized with the smallest values of μ such that

$$\text{sign}(\bar{\mathbf{B}}_1)(\bar{\mathbf{B}}_1 - \mu)_+ = 0, \tag{11}$$

where $\bar{\mathbf{B}}_1$ is the linear regression estimate. Then we built a sequence of μ values that decreased from the inferred value of the parameter $\mu^{\max}$ to $\mu^{\min} = \epsilon\mu^{\max}$. We eventually measured the number of non-null 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. Based on the singular values $(\lambda_1, \ldots, \lambda_n)$ of the matrix $\mathbf{D}_0\mathbf{Q}^T\mathbf{Y}$, we set

$$\gamma = \frac{(\lambda_K + \lambda_{K+1})}{2}. \tag{12}$$

With this value of γ, sparse LFMM converged to latent matrix estimates having rank K in our experiments.

3 Simulation study

In simulation studies, sparse LFMM was compared to two other sparse methods and three non-sparse algorithms, all based on the generative model defined in equation (1). In a first series of model-based experiments, we compared sparse LFMM with three non-sparse approaches, where the performance of each algorithm was measured in scenarios with higher or lower effect sizes and confounding intensities. In a second series with more realistic experiments, we simulated phenotypes based on association with empirical genotypes (Francois and Caye, 2018).

Estimation algorithms. As a baseline, we used Least Absolute Shrinkage and Selection Operator (LASSO) regression models (Tibshirani, 1996; Friedman et al., 2010). LASSO regression did not include any correction for confounding, and strong biases were expected in effect size estimates. The models were implemented in the `R` package `glmnet`, and their regularization parameter was selected by using a 5-fold cross validation approach (Zeng et al., 2017). We also used the Bayesian Sparse Linear Mixed Models (BSLMM) implemented in the `GEMMA` software (Zhou et al., 2013).

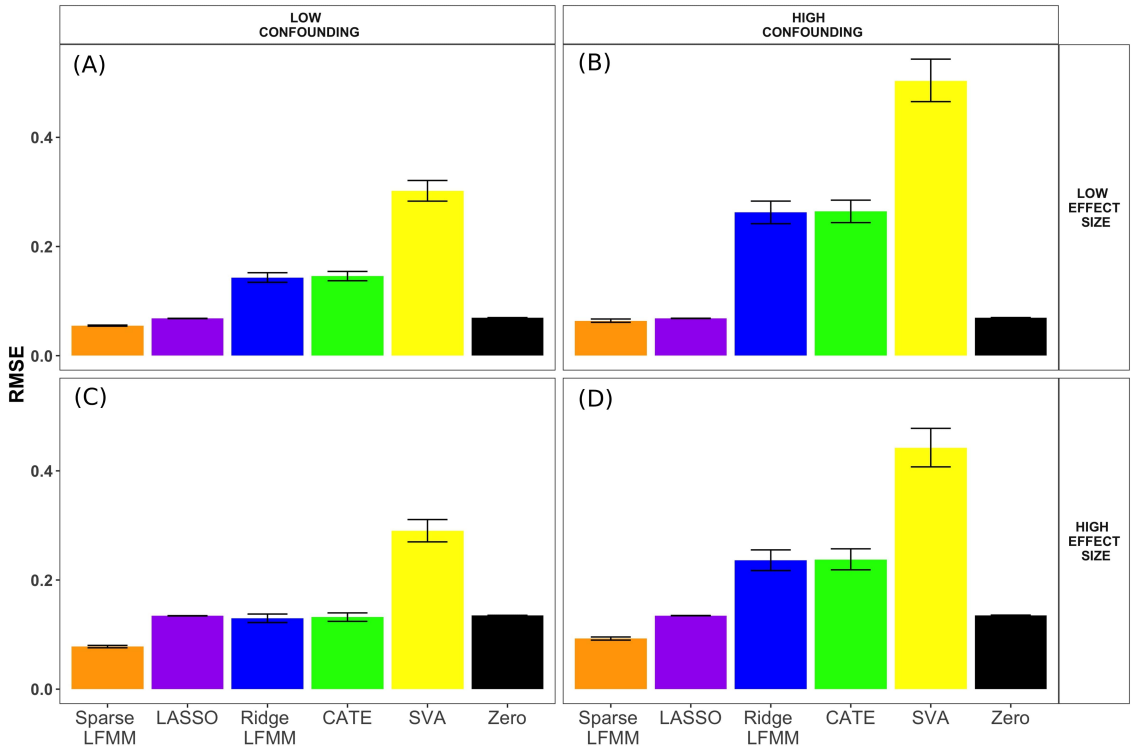

Figure 1. Root Mean Square Error (RMSE) as a function of the effect size of causal markers and confounding intensity. Two sparse methods (sparse LFMM, LASSO) and three non-sparse methods (ridge LFMM, CATE and SVA) were compared. The **zero** value corresponds to an RMSE obtained with all effect sizes set to zero (null-model error). Generative model simulation parameters: (A) Lower effect sizes and confounding intensities (B) Lower effect sizes and higher confounding intensities. (C) Higher effect sizes and lower confounding intensities. (D) Higher effect sizes and confounding intensities.

BSLMM is a hybrid method that combines sparse regression models with linear mixed models. BSLMM uses a Markov chain Monte Carlo (MCMC) method to estimate effect sizes. The MCMC burn-in and sampling periods were set to 10,000 (Zeng et al., 2017). To determine the proportion of non-zero effect sizes, two parameters were tuned ($p_{\min}$ and $p_{\max}$). Those parameters correspond to the logarithm of the maximum and minimum expected proportions of non-zero effect size. To compare sparse LFMM to non-sparse algorithms, we first used Surrogate Variable Analysis (SVA, (Leek and Storey, 2007)). The algorithm starts with estimating the loading values of a principal component analysis for the residuals of the regression of the response matrix $\mathbf{Y}$ on $\mathbf{X}$. In a second step, SVA determines a subset of response variables exhibiting low correlation with $\mathbf{X}$, and uses this subset of variables to estimate the latent factors. SVA was implemented in the `R` package `sva`. Next, we implemented the Confounder Adjusted Testing and Estimation (CATE) method (Wang et al., 2017). CATE uses a linear transformation of the response matrix such that the first axis of this transformation is collinear to $\mathbf{X}$ and the other axes are orthogonal to $\mathbf{X}$. CATE was used without negative controls, and it was implemented in the `R` package `cate`. We eventually used the ridge LFMM algorithm from the package `lfmm` (Caye et al., 2019).

Generative model simulations. In a first series of experiments, we compared sparse LFMM with LASSO and three non-sparse approaches (ridge LFMM, CATE, SVA), and the performance of each algorithm was measured in four scenarios showing high, medium or low effect sizes and confounding intensities. Artificial data were simulated from the generative model defined in equation (1), and the *confounding in-*

tensity was defined as the percentage of variance of the primary variable $\mathbf{X}$ explained by the latent factors $\mathbf{U}$.

More specifically, we simulated a single primary variable ($d = 1$), $K = 6$ latent factors, and a response matrix according to a multivariate Gaussian model (Caye et al., 2019). The joint distribution of $(\mathbf{X}, \mathbf{U})$ was $\mathrm{N}(\mathbf{0}, \mathbf{S})$, where $\mathbf{S}$ had diagonal terms $(s_1^2, \ldots, s_K^2, 1)$, and non-diagonal terms set to zero, except for the covariance between $\mathbf{u}_k$ and $\mathbf{X}$, which was set to $c_k \rho$. The c_k coefficients were sampled from a uniform distribution, and ρ was proportional to $1/\sqrt{\sum_k c_k^2/s_k^2}$. The standard deviations, (s_k), were randomly sampled from the interval (2, 6) for $k \leq 4$, and from the interval (0, 1) for $k = 5, 6$. The coefficient of proportionality was chosen so that the confounding intensity could take relatively low ($R^2 = 0.1$) or high ($R^2 = 0.5$) values.

To create sparse models, only a small proportion of effect sizes, equal to 0.8%, were allowed to be different from zero. Non-null effect sizes were sampled according to a Gaussian distribution, $\mathrm{N}(B, 0.2)$, where B could take two values, $B = 0.75$ (low value), $B = 1.5$ (medium value) and $B = 3.0$ (high value). Residual errors and loadings, $\mathbf{V}$, were sampled according to a standard Gaussian distribution. The dimensions of the response matrix were set to $n = 400$ individuals and $p = 10,000$ variables. The response matrix was eventually created by simulating from the generative model (equation (1)), and by using a probit link in order to mimic DNA methylation data. Two hundred simulations were performed for each combination of parameters (1200 simulations).

Results from generative model simulations. Estimation algorithms were used with their default parameters. The number of latent factors was set to $K = 6$ in

latent factor models, and the proportion of null-hypothesis was set to 1% in sparse models. Statistical errors (RMSE) of effect size estimates were computed for each method (Figure 1). To provide a reference value for the RMSE, we measured the error made when all effect sizes were estimated as being null ("zero" value or null-model error). The null-model error was equal to 0.069 in lower effect size scenarios and equal to 0.135 in higher effect size scenarios. The RMSEs of sparse LFMM ranged from 0.055 to 0.092, less than those of the null-model. The RMSEs of LASSO were close to the ones of sparse LFMM in low effect size scenarios. In contrast, non-sparse methods led to RMSEs higher than the null-model error, ranging between 0.13 and 0.26 for ridge LFMM and CATE, and rising up to 0.50 for SVA. For the effect sizes associated with causal markers, non-sparse methods reached lower RMSE values than those of sparse methods, ranging between 0.12 and 0.26 for ridge LFMM and CATE, and between 0.60 and 1.03 for sparse LFMM (Figure S1).

To evaluate the capabilities of methods to identify true positives, we used precision, which corresponds to the proportion of true positives in a list of positive markers, recall, which is the number of true positives divided by the number of causal markers, and F-score, which is the harmonic mean of precision and recall. To compute precision and F-score in generative model experiments, a list of $N = 100$ markers with the largest absolute estimated effect sizes was considered for each data set and method. The size of the list, N, was chosen as being representative of the number of markers subject to experimental validation. For this value, precision and F-score could not be larger than 80% and 89% respectively. To alleviate the choice of a particular value of N, we computed precision in (ranked) top hit lists of length n for $1 \leq n \leq N$, and reported the area under the curve (AUC) normalized by its maximum value. Top hit

lists with large AUC values contain more true positives than those with small AUC values.

For all methods, precision and F-score were higher in scenarios with higher effect size and lower confounding intensity. In all scenarios, sparse LFMM obtained better scores than SVA. Sparse LFMM reached higher performances than the LASSO when the sizes of the causal effects were medium or high (Figure 2). The situation was reversed only in the case of small effect sizes and for high confounding intensity. In those simulations, sparse LFMM obtained slightly lower scores than ridge LFMM and CATE. The difference was important when the size of causal effects was small (F-score ≈ 0.68 against F-score ≈ 0.85), but the differences were small for medium and large effect sizes. For large effect sizes, sparse LFMM, ridge LFMM and CATE were close to the maximal F-score (0.89) and precision (0.80). AUC scores were strongly correlated to F-scores, and the methods were ranked in the same way as with the other performance measures (Figure S2).

In summary, sparse LFMM was generally preferable to LASSO and SVA in generative model simulations. Sparse LFMM was associated with the smallest overall statistical error, but the estimates were biased more severely with this method than with non-sparse methods.

Empirical simulation experiments. In a second series of experiments, we used more realistic simulations to compare sparse LFMM to other sparse and non-sparse methods. Those simulations departed from generative model simulations, and were introduced to evaluate the robustness of effect size estimates in each approach. Phenotypes were simulated for $n = 162$ publicly available Single Nucleotide Polymor-

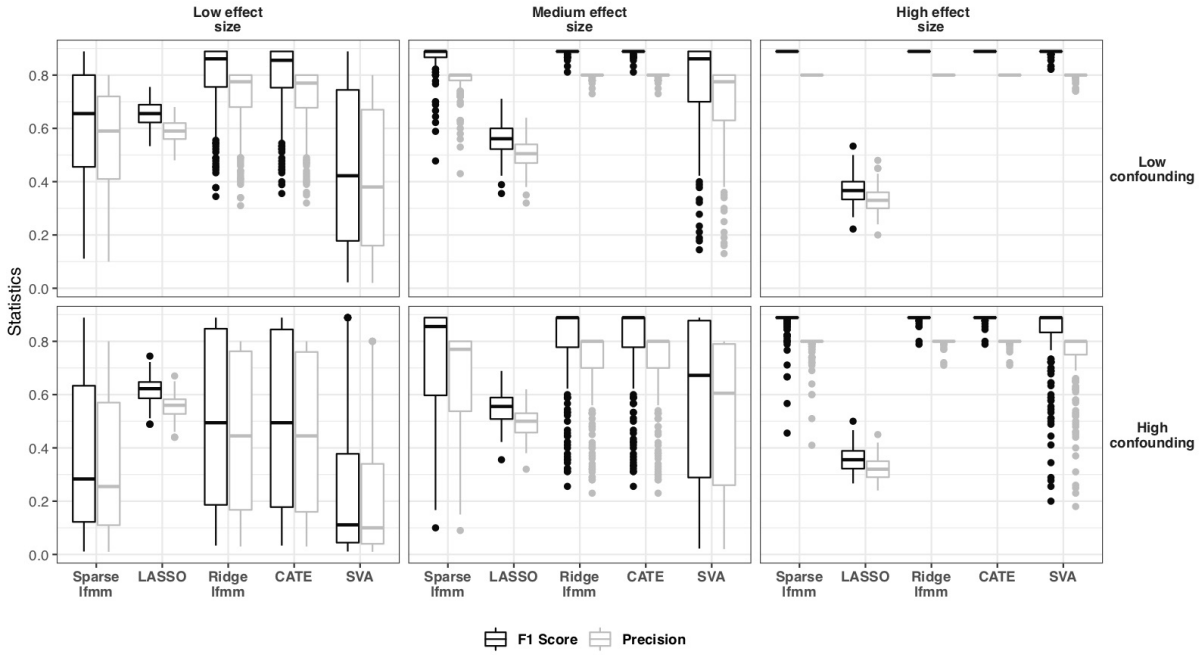

Figure 2. F-score and precision for sparse and non-sparse algorithms in generative model simulations. Two sparse methods (sparse LFMM, LASSO) and three non-sparse methods (ridge LFMM, CATE and SVA) were compared for various levels of effect size and confounding intensity in generative model simulations. F-score was defined as the harmonic mean of precision and recall.

phisms (SNPs) genotyped from the fifth chromosome of the model plant *Arabidopsis thaliana* (Atwell et al., 2010). The response matrix contained $p = 53,859$ SNPs, with minor allele frequency greater than 5%. Phenotypic simulations incorporated realistic features such as gene-by-environment ($G \times E$) interactions using environmental variables extracted from a bioclimatic database.

Simulations were performed as follows. Considering a predefined subset of causal markers (J), a phenotypic trait, x_i, was created for each individual from the following model

$$x_i = \sum_{j \in J} \beta_j y_{ij} + \sum_{j \in J} \delta_j x_{ij} e_i + \sum_{k=1}^{K} u_{ik} + \epsilon_i \,,$$

where y_{ij} represents the genotype of individual i at locus j, e_i corresponds to the first principal component of nineteen variables extracted from the bioclimatic database, $\mathbf{u}_k$ represents a confounding factor, and ϵ_i is a residual noise. The β parameters represent causal variant effect sizes. The number of confounding factors was $K = 6$, and the phenotypes were generated from a combination of five causal SNPs with identical effect sizes. Two values of effect size were used, $\beta = 6$ (lower effect size) and $\beta = 9$ (higher effect size). $G \times E$ interactions were represented by the δ parameters. Two values of gene-by-environment interaction were implemented, $G \times E = 0.1$ (lower $G \times E$) and $G \times E = 0.9$ (higher $G \times E$). For each parameter combination, two hundred simulations were performed.

Results from empirical simulation experiments. In empirical simulations, the F-score was modified to account for linkage disequilibrium (LD) in the data, and candidate markers within a window of size 10kb around a causal marker were considered to be true discoveries (LD-$r^2 < 0.2$, Francois and Caye (2018)). In lower $G \times E$ sce-

narios, sparse LFMM obtained the highest scores (F-score in the interval (0.57,0.60), precision in (0.81,0.82), Figure 3AC) compared to BSLMM (F-score in (0.36,0.44)), and to non-sparse methods (F-score in the interval (0.25, 0.28)). In higher $G \times E$ scenarios, all methods had poor performances in the lower effect size scenario, but sparse LFMM obtained among the highest F-score and precision. When the effect size was higher, sparse LFMM reached higher performances (F-score ≈ 0.28 and accuracy ≈ 0.33) than the other methods (Figure 3D). In those realistic simulations, sparse LFMM demonstrated greater robustness to departure from the generative model assumptions than the sparse methods BSLMM and LASSO, and compared favorably to non-sparse methods (ridge LFMM, CATE and SVA).

Runtimes. Next, we evaluated runtimes for sparse LFMM, and compared those runtimes with BSLMM and ridge LFMM (Figure S3, runtimes measured on an Intel Xeon w-2145 CPU 3.7 GHz). Whatever the number of individuals or markers, ridge LFMM was the fastest method, and sparse LFMM was the slowest method. Higher computation times for sparse LFMM were not surprising because the method iterates many cycles before convergence, whereas ridge LFMM is an exact approach. It took around 2,000 seconds ($\approx$ 33 minutes) for sparse LFMM to complete runs with $n = 1,000$ individuals and $p = 100,000$ markers. While ridge LFMM took a few seconds to compute estimates, around thirty minutes remained an acceptable duration in practical applications. With default values of its MCMC parameters, BSLMM runtimes were of the same order as those of sparse LFMM.

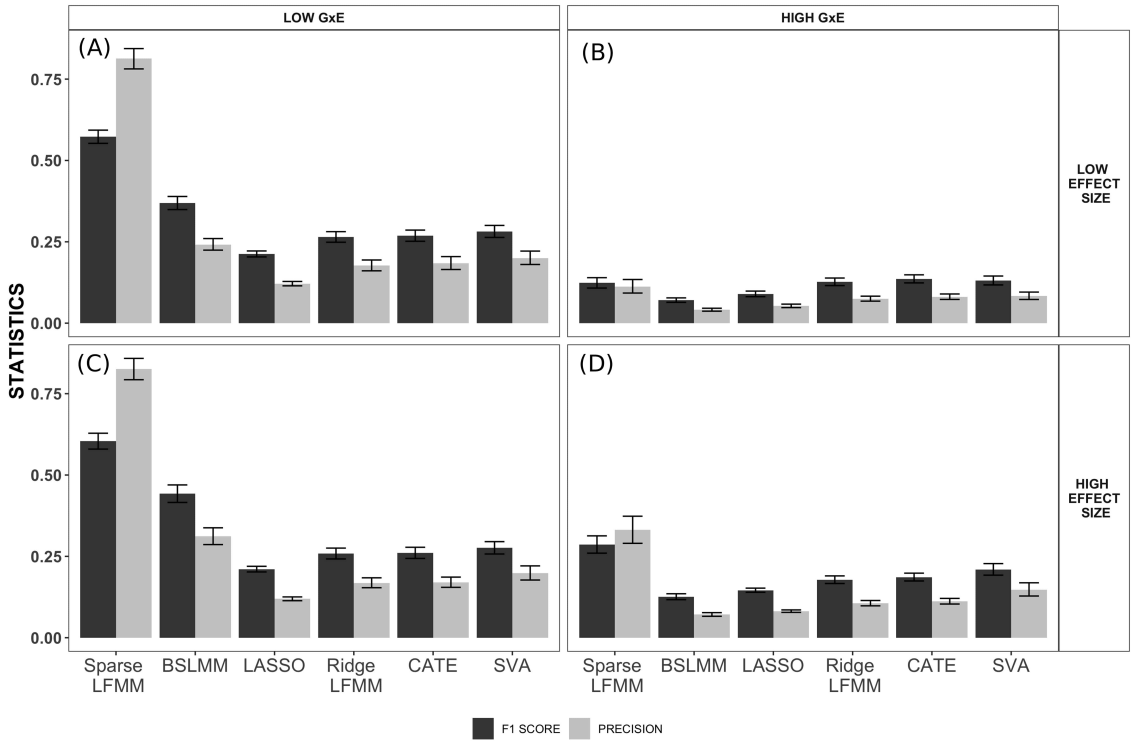

Figure 3. Empirical simulation data (F-score and precision). F-score and precision as a function of the effect size of the causal markers and of the strength of the interaction between genotype and environment ($G \times E$). Three sparse methods (sparse LFMM, BSLMM and LASSO) and three non-sparse methods (ridge LFMM, CATE and SVA) were compared. F-score is the harmonic mean of precision and recall. Simulation parameters: (A) Lower effect sizes and lower $G \times E$ (B) Lower effect sizes and higher $G \times E$. (C) Higher effect sizes and lower $G \times E$. (D) Higher effect sizes and higher $G \times E$.

4 Data analysis

Sparse LFMM and other methods were used to perform association studies for two distinct types of genomic data including genotypic and epigenetic markers. To illustrate the use of latent factor models in a context where confounding is difficult to control for, we performed a GWAS of flowering time using $p = 53,859$ SNPs genotyped in chromosome 5 for $n = 162$ European accessions of the model plant *Arabidopsis thaliana*. Our second application evaluated association between smoking during pregnancy and placental DNA methylation for women from a mother-cohort (Heude et al., 2016).

GWAS of flowering time in *Arabidopsis thaliana*. We considered $n = 162$ European accessions and $p = 53,859$ SNPs from the fifth chromosome of the *Arabidopsis thaliana* genome to investigate associations with a flowering time phenotype (FT16-TO: 0000344, (Atwell et al., 2010)). FT16 corresponds to the number of days required for an individual plant to reach the flowering stage. In sparse LFMM, the percentage of non-null effect size was set to 1%, in agreement with the proportion of null-hypothesis estimated from non-sparse approaches. The parameters $p_{\min}$ and $p_{\max}$ defining sparsity in the BSLMM algorithm were fixed to $p_{\min} = -5$ and $p_{\max} = -4$ respectively. These values correspond to the logarithm of expected proportions of non-null effect sizes in BSLMM. For all factor methods, the number of latent factors was set to $K = 10$, determined by the eigenvalue methods (see Figure S4).

The sparse methods (sparse LFMM, LASSO, BSLMM) differed in their estimate of the number of null effect sizes (Figure 4ABC, Figure S5). The LASSO approach estimated 99.85% null effect sizes while the proportions were equal to 99.24% and

98.18% for BSLMM and sparse LFMM respectively. The LASSO was the most conservative approach, and sparse LFMM the most liberal one. Sparse LFMM shared 3.9% of hits with LASSO, and 5.5% with BSLMM (Figure S5). Less than 1% of all hits were common to the three approaches. The (non-null) effect sizes for hits varied on distinct scales, with LASSO exhibiting the strongest biases. All sparse methods detected the same top hit at around 4 Mb, corresponding to a SNP located within the *FLC* gene, consistent with the results of (Atwell et al., 2010). The second hit in (Atwell et al., 2010), located in the gene *DOG1*, was also identified by sparse LFMM. BSLMM had more difficulties in identifying previously discovered genes. Given the high correlation – greater than 94 % – between effect sizes obtained with non-sparse methods, we grouped their results by averaging their estimates. Non-sparse methods exhibited effect sizes in a range of values closer to sparse LFMM than to LASSO and BSLMM, but higher statistical errors were observed for those approaches (Figure 4D). Overall, we found a significant correlation between the non-null effect sizes estimated by sparse LFMM and the corresponding effect sizes found by non-sparse methods ($\rho = 0.8065$, $P < 10^{-16}$). In addition, sparse LFMM and the non-sparse methods detected new hits around 13.9 Mb and 6.5 Mb of chr 5, corresponding to the *SAP* and *ACL5* genes respectively.

EWAS of exposure to smoking during pregnancy. Considering beta-normalized methylation levels at $p = 425,878$ probed CpG sites for $n = 668$ women, we tested association between smoking status (219 current smoker women and 449 non-current smoker women) and DNA methylation (mDNA) levels in placental tissues. Detailed information on the study population and protocols for placental DNA methylation

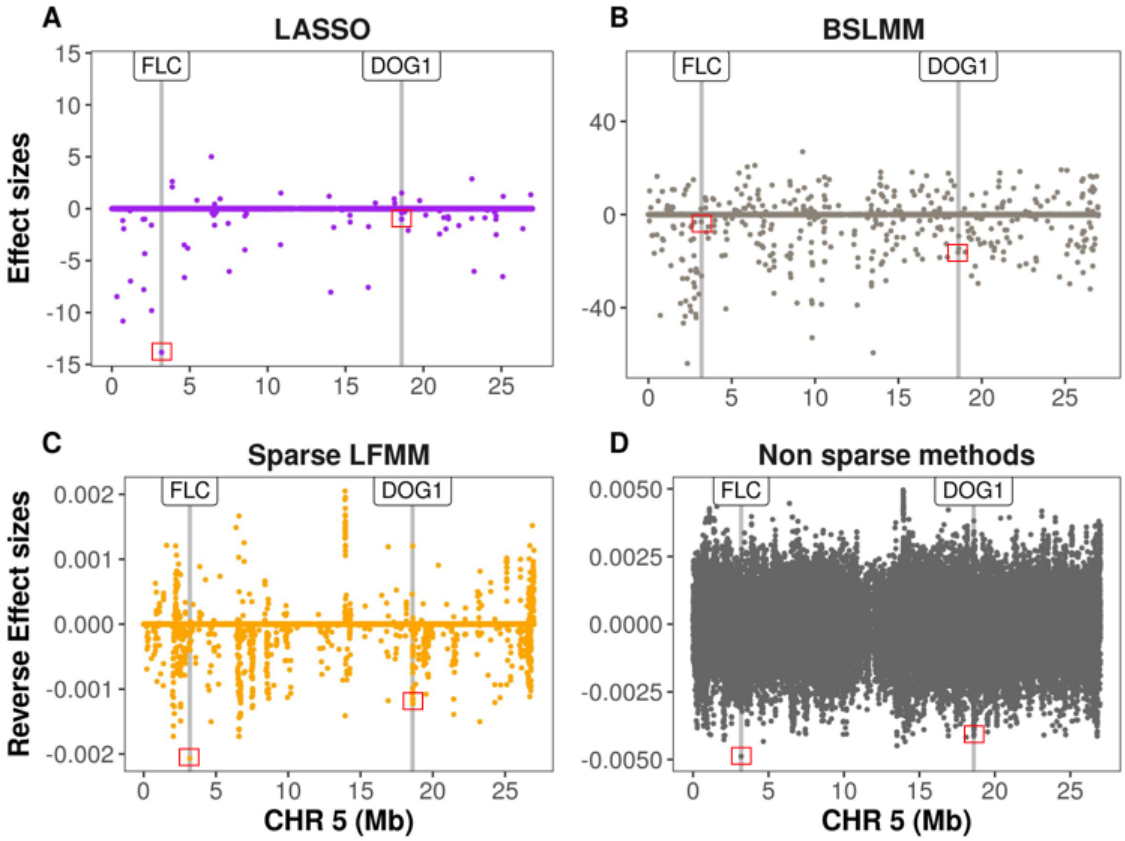

Figure 4. GWAS of flowering time in *A. thaliana* **(chromosome 5).** A) Effect size estimates for LASSO. B) Effect size estimates for BSLMM. C) Reverse effect size estimates for sparse LFMM. D) Average of reverse effect size estimates for non-sparse methods (ridge LFMM, CATE and SVA). Grey bars represent *Arabidopsis* SNPs associated with the FT16 phenotype in (Atwell et al., 2010), and correspond to the *FLC* and *DOG1* genes.

assessment processing could be found in (Abraham et al., 2018; Rousseaux et al., 2019). For the proportion of non-null effect sizes in sparse LFMM, a conservative value (0.1%) was used (Rousseaux et al., 2019). For latent factor models, the number of latent factors was set to $K = 7$ (see Figure S6).

Using sparse LFMM, the estimated proportion of null effect sizes was equal to 99.698%, for a total number of 1,287 hits (Figure S7). To characterize those 1,287 CpGs, we evaluated whether there was an enrichment of enhancer and promoter regions compared to the methylome (Figure S8 and Figure S9). Promotor and enhancer regions were identified according to the Illumina's Infinium HumanMethylation450 BeadChip annotations. A fraction of 25.48% were found in enhancer regions, compared to 22.73% for the whole methylome, and 6.83% were found in promoter regions, compared to 19.94% for the whole methylome. We compared the CpGs having the highest effect sizes identified by each method (Figure S10). Sparse LFMM shared 45.3 % of hits with non-sparse models (represented by ridge LFMM), and 2.8 % of hits with LASSO (Table S1). Among the 51 top hits shared by sparse LFMM and ridge LFMM, 25 were found in the body of a gene, 11 were not associated with a gene, 20 were in enhancer regions and 2 in promoter regions. We averaged the effect sizes of non-sparse methods ridge LFMM, CATE and SVA because their correlation was greater than 99%. The results of sparse LFMM agreed with the results of non-sparse methods better than with those of LASSO. The Pearson correlation between the non-null effect sizes estimated by sparse LFMM and the corresponding effect sizes estimated non-sparse methods was equal to $\rho = 80.38\%$ ($P < 10^{-16}$), whereas the Pearson correlation between non-null effect sizes of sparse LFMM and LASSO was equal to $\rho = 61.86\%$ ($P < 10^{-16}$). To detail results for a specific genomic region,

we focused on chromosome 3, which contained the epigenome-wide top hit for sparse LFMM and for non-sparse methods (cg27402634, located in an enhancer region, Figure 5), and also detected with LASSO (Figure S11). Sparse LFMM hits shared three additional CpGs with non-sparse methods: cg09627057, cg18557837 and cg12662091. Overall, sparse LFMM detected 61 CpGs with non-null effect sizes: 43 were located in genes, 22 in enhancer regions and 6 in promoter regions.

5 Discussion

We introduced sparse latent factor regression methods for the joint estimation of effect sizes and latent factors in genomic and epigenomic association studies. These methods were implemented in the `R` package `lfmm`, and are available from the CRAN repository. In generative and in empirical simulations, sparse LFMM obtained higher F-score and precision than previously introduced sparse methods, BSLMM and LASSO. Compared to three non-sparse methods (ridge LFMM, CATE and SVA), statistical errors of effect size estimates were overall reduced. In generative model simulations, ridge LFMM and CATE reached higher precision and F-score than sparse LFMM. An explanation for this result may be that generative model simulations favor methods for which the estimation of latent factors is unconstrained, and the nuclear norm penalty put on the latent matrix may increase bias in effect size estimates. In contrast, sparse LFMM reached the highest performances in simulations based on empirical data. By being more robust to departure from model assumptions, simulations based on empirical data showed that sparse LFMM can outperform non-sparse methods (see Lotterhos (2019) for a similar conclusion based on population genetic simulations).

A limitation of the sparse LFMM algorithm was its relatively slow execution

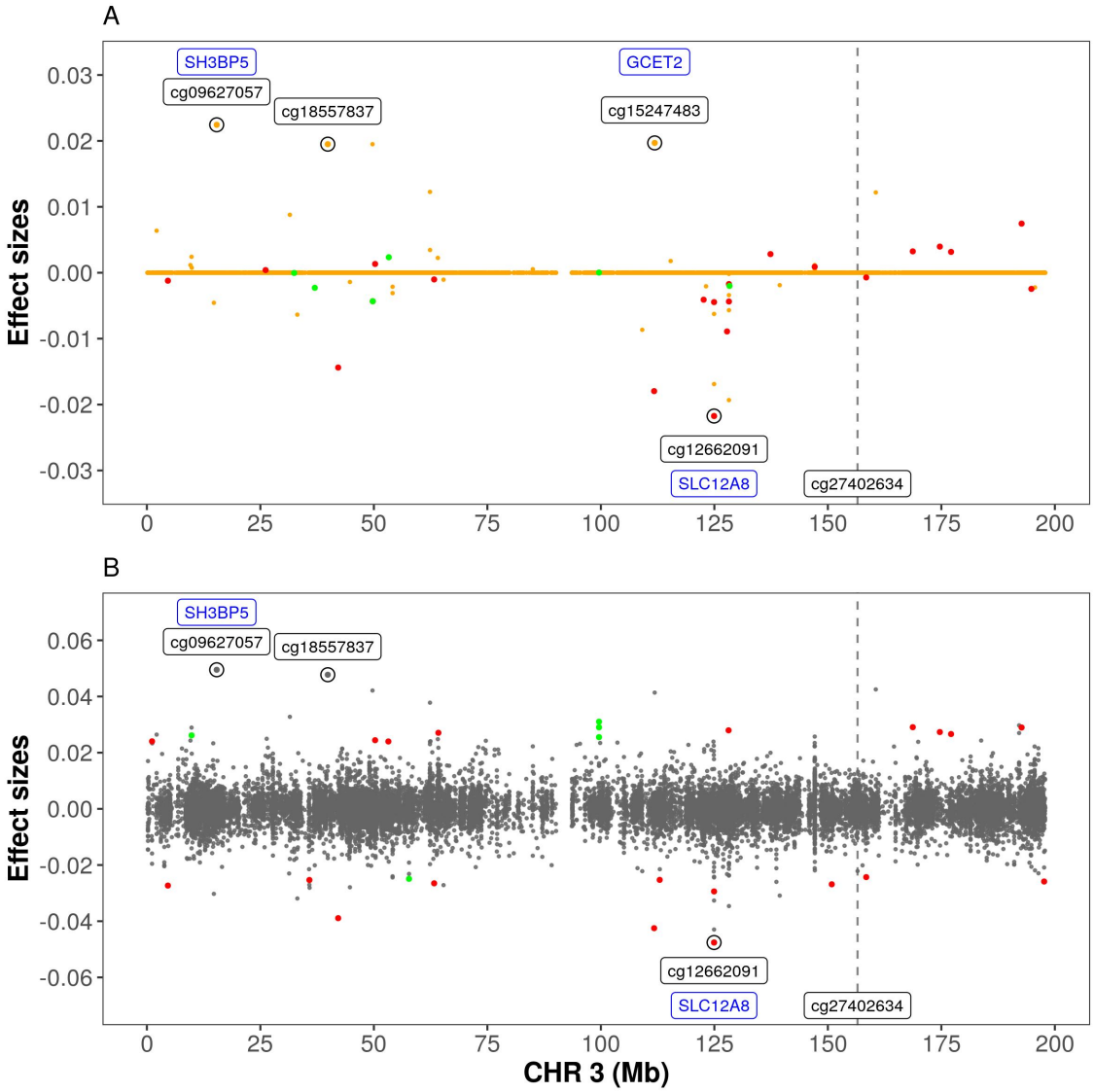

Figure 5. DNA methylation EWAS of smoking status in pregnant women (chromosome 3). A) Estimated effect size for sparse LFMM. The effect size at cg27402634 is equal to $\beta = -0.117$ (out of range). B) Estimated effect size for non-sparse methods (ridge LFMM, CATE and SVA). The effect size at cg27402634 is equal to $\beta = -0.141$ (out of range). CpGs with the highest effect sizes are circled (genes are in blue color). Red dots represent CpGs located in enhancer regions. Green dots represent CpGs located in promoter regions.

time, preventing deep exploration of hyperparameter space in real applications. Our strategy was to propose simple guidelines to evaluate the proportion of non-null effect size through an empirical null hypothesis approach, for the number of factors and for the nuclear norm parameter. For causal markers, the effect sizes estimated by sparse LFMM and the corresponding effect sizes estimated by non-sparse methods were strongly correlated. Because effect size estimates had a lower bias in non-sparse methods compared to sparse methods, the results suggest to combine sparse LFMM with a non-sparse method in the following way. At a first stage, sparse LFMM can be used to estimate the support of causal markers having non-null effect sizes. Then ridge LFMM or CATE can be used to estimate the effect sizes of the selected markers. By using sparse LFMM together with a hypothesis testing method, the combined approach alleviates the difficult problem of defining corrections for multiple correlated tests.

In a GWAS of flowering time using 53,859 SNPs in the fifth chromosome of 162 European accessions of the plant *Arabidopsis thaliana*, sparse LFMM identified the *FLC* and *DOG1* genes associated with the FT16 phenotype. The two genes were previously reported as being associated with FT16 in (Atwell et al., 2010). The *FLC* gene plays a central role in flowering induced by vernalization (Sheldon et al., 2000), and *DOG1* is involved in the control of dormancy and seed germination (Nishimura et al., 2018). The second hit of sparse LFMM corresponded to SNPs linked to *SAP*, which is a transcriptional regulator involved in the specification of floral identity (Byzova et al., 1999). This association was also significant for non-sparse methods. In addition, the new method detected SNPs located in the *ACL5* gene, which plays a role in internodal growth and organ size (Hanzawa et al., 1997). In summary,

sparse methods facilitated the selection of non-null effect sizes. The results for sparse LFMM were not only consistent with previous discoveries, but they also identified new candidate genes with interesting functional annotations.

Next, we applied sparse LFMM in an EWAS of placental DNA methylation in women exposed to smoking during pregnancy, which is considered an important risk factor for child health (Lumley et al., 2009). The CpG marker cg27402634, which had the highest effect size both in sparse LFMM and in non-sparse methods, was located in an enhancer region, close to the *LEKR1* gene which was associated with birth weight in a GWAS from the Early Growth Genetics (EGG) consortium (`http://egg-consortium.org/birth-weight.html`). This marker was also detected as a top hit in an independent study of placental methylation and smoking (Morales et al., 2016), and with slightly different study populations and measures of tobacco consumption Rousseaux et al. (2019). In addition to cg27402634, the list of top hits contained several CpGs reported in similar studies, including cg21992501 in the gene *TTC27* (Cardenas et al., 2019), cg25585967 and cg17823829, respectively in the *TRIO* and *KDM5B* genes (Morales et al., 2016; Everson et al., 2019). To characterize the full list of hits, we found an enrichment of enhancer regions and a depletion of promoter regions for CpGs, in agreement with the findings of (Rousseaux et al., 2019). Overall sparse LFMM allowed us to confirm previously discovered associations, and to detect new associations including genes for which methylation changes have detrimental effects on the health of the child.

Conclusion. Accounting for variation due to unobserved confounding factors is extremely difficult in any type of association study. Assuming that a small proportion

of all markers correlate with the exposure or phenotype, we addressed the confounding issue by using sparse latent factor regression models, providing mathematical guarantees that global solutions of least squares estimation problems are proposed. Sparsity constraints in our algorithm allowed the selection of markers alleviating the issue of multiple tests. The application of our method to real data sets highlighted new associations with relevant biological meaning.

Program Availability. All codes are publicly available, implemented in the `R` package `lfmm` available from GitHub for its development version (`https://bcm-uga.github.io/lfmm/`) and released in the Comprehensive `R` Archive Network (`https://cran.r-project.org/`). The scripts used in simulation experiments and in data analyses are available from the following link: `https://bcm-uga.github.io/scripts_sagmb`.

Data Availability and Ethics Statement. The *Arabidopsis thaliana* data are publicly available from the 1,001 Genomes database (`https://1001genomes.org/`). The EDEN individual-level data have restricted access owing to ethical and legal conditions in France. They are available upon request from the EDEN steering committee at `etude.eden@inserm.fr` and through collaborations with the principal investigators of EDEN. The EDEN cohort received approval from the ethics committee (CCPPRB) of Kremlin Bicêtre and from the French data privacy institution "Commission Nationale de l'Informatique et des Libertés (CNIL)". Written consent was obtained from the mother for herself and for the offspring.

Mathematical proofs

This section provides mathematical proofs for the results stated in section 2.

Theorem 1. *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 minimum of the penalized loss function $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$.*

Proof. The proof arguments are based on a result of (Tseng, 2001). Consider the Cartesian product of closed convex sets $A = A_1 \times A_2 \times \cdots \times A_m$, and let $f(\mathbf{z})$ be a continuous convex function defined on A and such that

$$f(z_1, \cdots, z_m) = g(z_1, \cdots, z_m) + \sum_{i=1}^{m} f_i(z_i)\,,$$

where $g(\mathbf{z})$ is a differentiable convex function, and for each $i = 1, \ldots, m$, $f_i(z_i)$ is a continuous convex function. Let $(\mathbf{z}^t)$ be the sequence of values defined by the following block-coordinate descent algorithm

$$z_i^{t+1} \in \arg\min_{\zeta \in A_i} f(z_1^t, \ldots, z_{i-1}^t, \zeta, z_{i+1}^t, \ldots, z_m^t)\,, \quad i = 1, \ldots, m. \tag{13}$$

Then a limit point of the sequence $(\mathbf{z}^t)$ defines a global minimum of the function $f(\mathbf{z})$. The theorem's proof is a consequence of the convexity of the penalized loss function $\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B})$, and the fact that we can write

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

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. Tseng's result can be applied with the function $f(\mathbf{B}, \mathbf{W}) = \mathcal{L}_{\text{sparse}}(\mathbf{B}, \mathbf{W})$ to conclude the proof (see also (Bertsekas, 1995)).

□

Theorem 2. *Let $\lambda > 0$ and assume $\sigma_i^2 > 0$ for all $i = 1, \ldots, d$. The estimates $\hat{\mathbf{W}}$ and $\hat{\mathbf{B}}$ computed as follows*

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\text{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y}) \tag{14}$$

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{W}}), \tag{15}$$

where $\text{svd}_K(\mathbf{A})$ is the rank K SVD 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 as

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

define a global minimum of the penalized loss function $\mathcal{L}_{\text{ridge}}(\mathbf{B}, \mathbf{W})$.

Proof. Given $\mathbf{W}$, a global minimum for $\mathcal{L}_{\text{ridge}}(\mathbf{B}, \mathbf{W})$ is obtained with the ridge estimates for a linear regression of the response matrix $\mathbf{Y} - \mathbf{W}$ on $\mathbf{X}$.

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \mathbf{W}). \tag{16}$$

Thus, the problem amounts to minimizing the function $\mathcal{L}(\mathbf{W}) = \mathcal{L}_{\text{ridge}}(\hat{\mathbf{B}}, \mathbf{W})$ with respect to $\mathbf{W}$. By definition of the $\mathbf{D}_\lambda$ and $\mathbf{Q}$ matrices, the loss function rewrites as

$$\mathcal{L}(\mathbf{W}) = \left\|\mathbf{D}_\lambda\mathbf{Q}^T(\mathbf{Y} - \mathbf{W})\right\|_F^2. \tag{17}$$

Minimizing the above loss function is equivalent to finding the best approximation of rank K for the matrix $\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}$. According to (Eckart and Young, 1936), this approximation is given by the rank K singular value decomposition of $\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}$. Eventually we obtain that

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1} \mathrm{svd}_K(\mathbf{D}_\lambda \mathbf{Q}^T \mathbf{Y}) \tag{18}$$

defines the unique global minimum of the $\mathcal{L}(\mathbf{W})$ function. □

Acknowledgments

This article was developed in the framework of the Grenoble Alpes Data Institute, supported by the French National Research Agency under the Investissements d'Avenir program (ANR-15-IDEX-02). It received support from LabEx PERSYVAL Lab, ANR-11-LABX-0025-01, and from the French National Research Agency (Agence Nationale pour la Recherche) ETAPE, ANR-18-CE36-0005. The EDEN mother-child study was supported by Foundation for medical research (FRM), National Agency for Research (ANR), National Institute for Research in Public health (IRESP), French Ministry of Health (DGS), French Ministry of Research, INSERM Bone and Joint Diseases National Research (PRO-A), and Human Nutrition National Research Programs, Nestlé, French National Institute for Population Health Surveillance (InVS), French National Institute for Health Education (INPES), the European Union FP7 programmes (FP7/2007-2013, HELIX, ESCAPE, ENRIECO, Medall projects), Diabetes National Research Program, French Agency for Environmental Health Safety (ANSES), Mutuelle Générale de l'Education Nationale (MGEN), French national agency for food security, French-speaking association for the study of

diabetes and metabolism (ALFEDIAM). We thank all the participants and members of the EDEN mother-child cohort study group.

References

E. Abraham, S. Rousseaux, L. Agier, L. Giorgis-Allemand, J. Tost, J. Galineau, A. Hulin, V. Siroux, D. Vaiman, M.-A. Charles, B. Heude, A. Forhan, J. Schwartz, F. Chuffart, E. Bourova-Flin, S. Khochbin, R. Slama, and J. Lepeule. Pregnancy exposure to atmospheric pollution and meteorological conditions and placental DNA methylation. *Environment International*, 118:334–347, Sept. 2018. ISSN 01604120. doi: 10.1016/j.envint.2018.05.007. URL `https://linkinghub.elsevier.com/retrieve/pii/S0160412017320433`.

S. Atwell, Y. S. Huang, B. J. Vilhjalmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, R. Jiang, N. W. Muliyati, X. Zhang, M. A. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J. R. Ecker, N. Faure, J. M. Kniskern, J. D. G. Jones, T. Michael, A. Nemri, F. Roux, D. E. Salt, C. Tang, M. Todesco, M. B. Traw, D. Weigel, P. Marjoram, J. O. Borevitz, J. Bergelson, and M. Nordborg. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, June 2010. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature08800. URL `http://www.nature.com/articles/nature08800`.

D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews. Genetics*, 7(10):781–791, Oct. 2006. ISSN 1471-0056. doi: 10.1038/nrg1916.

T. Battram, P. Yousefi, G. Crawford, C. Prince, M. S. Babei, G. Sharp, C. Hatcher, M. J. Vega-Salas, S. Khodabakhsh, O. Whitehurst, R. Langdon, L. Mahoney, H. R. Elliott, G. Mancano, M. Lee, S. H. Watkins, A. C. Lay, G. Hemani, T. R. Gaunt, C. L. Relton, J. R. Staley, and M. Suderman. The EWAS Catalog: a database of epigenome-wide association studies. Technical report, OSF Preprints, Feb. 2021. URL `https://osf.io/837wn/`. type: article.

D. Bertsekas. Nonlinear Programming. *Athena Scientific*, 48, Jan. 1995. doi: 10.1057/palgrave.jors.2600425.

A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1): D1005–D1012, Jan. 2019. ISSN 1362-4962. doi: 10.1093/nar/gky1120.

M. V. Byzova, J. Franken, M. G. Aarts, J. de Almeida-Engler, G. Engler, C. Mariani, M. M. Van Lookeren Campagne, and G. C. Angenent. Arabidopsis STERILE APETALA, a multifunctional gene regulating inflorescence, flower, and ovule development. *Genes & Development*, 13(8):1002–1014, Apr. 1999. ISSN 0890-9369. doi: 10.1101/gad.13.8.1002. URL `http://www.genesdev.org/cgi/doi/10.1101/gad.13.8.1002`.

J.-F. Cai, E. J. Candes, and Z. Shen. A Singular Value Thresholding Algorithm for

Matrix Completion. *arXiv:0810.3286 [math]*, Oct. 2008. URL http://arxiv.org/abs/0810.3286. arXiv: 0810.3286.

A. Cardenas, S. M. Lutz, T. M. Everson, P. Perron, L. Bouchard, and M.-F. Hivert. Mediation by Placental DNA Methylation of the Association of Prenatal Maternal Smoking and Birth Weight. *American Journal of Epidemiology*, 188(11):1878–1886, Nov. 2019. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwz184. URL https://academic.oup.com/aje/article/188/11/1878/5566169.

C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008. ISSN 0162-1459. doi: 10.1198/016214508000000869.

K. Caye, B. Jumentier, J. Lepeule, and O. Francois. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology and Evolution*, 36(4):852–860, Apr. 2019. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msz008. URL https://academic.oup.com/mbe/article/36/4/852/5290100.

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55 (4):997–1004, Dec. 1999. ISSN 0006-341X. doi: 10.1111/j.0006-341x.1999.00997.x.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, Sept. 1936. ISSN 1860-0980. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

B. Efron. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004. ISSN 0162-1459. URL https://www.jstor.org/stable/27590356. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

T. M. Everson, M. Vives-Usano, E. Seyve, A. Cardenas, M. Lacasana, J. M. Craig, C. Lesseur, E. R. Baker, N. Fernandez-Jimenez, B. Heude, P. Perron, B. Gonzalez-Alzaga, J. Halliday, M. A. Deyssenroth, M. R. Karagas, C. Iniguez, L. Bouchard, P. Carmona-Saez, Y. J. Loke, K. Hao, T. Belmonte, M. A. Charles, J. Martorell-Marugan, E. Muggli, J. Chen, M. F. Fernandez, J. Tost, A. Gomez-Martin, S. J. London, J. Sunyer, C. J. Marsit, J. Lepeule, M.-F. Hivert, and M. Bustamante. Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth. preprint, Genomics, June 2019. URL http://biorxiv.org/lookup/doi/10.1101/663567.

O. Francois and K. Caye. Naturalgwas: An R package for evaluating genomewide association methods with empirical data. *Molecular Ecology Resources*, 18(4):789–797, July 2018. ISSN 1755098X. doi: 10.1111/1755-0998.12892. URL http://doi.wiley.com/10.1111/1755-0998.12892.

E. Frichot, S. D. Schoville, G. Bouchard, and O. Francois. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, 30(7):1687–1699, July 2013. ISSN 1537-1719, 0737-4038. doi: 10.1093/molbev/mst063. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst063.

J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate op-

timization. *The Annals of Applied Statistics*, 1(2):302–332, Dec. 2007. ISSN 1932-6157. doi: 10.1214/07-AOAS131. URL http://arxiv.org/abs/0708.1485. arXiv: 0708.1485.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL http://www.jstatsoft.org/v33/i01/.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, Jan. 2011. ISSN 0036-1445, 1095-7200. doi: 10.1137/090771806. URL http://epubs.siam.org/doi/10.1137/090771806.

Y. Hanzawa, T. Takahashi, and Y. Komeda. ACL5: an Arabidopsis gene required for internodal elongation after flowering. *The Plant Journal*, 12(4):863–874, Oct. 1997. ISSN 1365-313X. doi: 10.1046/j.1365-313X.1997.12040863.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-313X.1997.12040863.x.

B. Heude, A. Forhan, R. Slama, L. Douhaud, S. Bedel, M.-J. Saurel-Cubizolles, R. Hankard, O. Thiebaugeorges, M. De Agostini, I. Annesi-Maesano, M. Kaminski, and M.-A. Charles. Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. *International Journal of Epidemiology*, 45(2):353–363, Apr. 2016. ISSN 0300-5771, 1464-3685. doi: 10.1093/ije/dyv151. URL https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyv151.

C. J. Hoggart, T. G. Clark, M. De Iorio, J. C. Whittaker, and D. J. Balding. Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, 32 (2):179–185, Feb. 2008. ISSN 0741-0395. doi: 10.1002/gepi.20292.

A. Kalaitzis and N. Lawrence. Residual Component Analysis: Generalising PCA for more flexible inference in linear-Gaussian models. *arXiv:1206.4560 [cs, stat]*, June 2012. URL `http://arxiv.org/abs/1206.4560`. arXiv: 1206.4560.

A. Kaushal, H. Zhang, W. J. J. Karmaus, M. Ray, M. A. Torres, A. K. Smith, and S.-L. Wang. Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinformatics*, 18(1), Dec. 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1611-2. URL `http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1611-2`.

S. Lee, W. Sun, F. A. Wright, and F. Zou. An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika*, 104(2):303–316, June 2017. ISSN 0006-3444. doi: 10.1093/biomet/asx018.

J. T. Leek. Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics*, 67(2):344–352, June 2011. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2010.01455.x.

J. T. Leek and J. D. Storey. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, 3(9):12, 2007.

K. Lotterhos. The effect of neutral recombination variation on genome scans for selection. *Biometrika*, 9:1851–1867, 2019.

J. Lumley, C. Chamberlain, T. Dowswell, S. Oliver, L. Oakley, and L. Watson. Interventions for promoting smoking cessation during pregnancy. *The Cochrane Database of Systematic Reviews*, (3):CD001055, July 2009. ISSN 1469-493X. doi: 10.1002/14651858.CD001055.pub3.

B. Mishra, G. Meyer, F. Bach, and R. Sepulchre. Low-rank optimization with trace norm penalty. *arXiv:1112.2318 [cs, math]*, June 2013. URL `http://arxiv.org/abs/1112.2318`. arXiv: 1112.2318.

E. Morales, N. Vilahur, L. A. Salas, V. Motta, M. F. Fernandez, M. Murcia, S. Llop, A. Tardon, G. Fernandez-Tardon, L. Santa-Marina, M. Gallastegui, V. Bollati, X. Estivill, N. Olea, J. Sunyer, and M. Bustamante. Genome-wide DNA methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy. *International Journal of Epidemiology*, 45(5):1644–1655, Oct. 2016. ISSN 0300-5771, 1464-3685. doi: 10.1093/ije/dyw196. URL `https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyw196`.

N. Nishimura, W. Tsuchiya, J. J. Moresco, Y. Hayashi, K. Satoh, N. Kaiwa, T. Irisa, T. Kinoshita, J. I. Schroeder, J. R. Yates, T. Hirayama, and T. Yamazaki. Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme. *Nature Communications*, 9(1), Dec. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04437-9. URL `http://www.nature.com/articles/s41467-018-04437-9`.

A. Onatski. DETERMINING THE NUMBER OF FACTORS FROM EMPIRICAL DISTRIBUTION OF EIGENVALUES. *The Review of Economics and Statistics*,

92(4):1004–1016, 2010. ISSN 0034-6535. URL https://www.jstor.org/stable/40985808. Publisher: The MIT Press.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, Aug. 2006. ISSN 1061-4036. doi: 10.1038/ng1847.

V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews. Genetics*, 12(8):529–541, July 2011. ISSN 1471-0064. doi: 10.1038/nrg3000.

S. Rousseaux, E. Seyve, F. Chuffart, E. Bourova-Flin, M. Benmerad, M.-A. Charles, A. Forhan, B. Heude, V. Siroux, R. Slama, J. Tost, D. Vaiman, S. Khochbin, J. Lepeule, and the EDEN mother-child cohort study group. Maternal exposure to cigarette smoking induces immediate and durable changes in placental DNA methylation affecting enhancer and imprinting control regions. preprint, Genomics, Nov. 2019. URL http://biorxiv.org/lookup/doi/10.1101/852186.

Y. She and K. Chen. Robust reduced-rank regression. *Biometrika*, 104(3):633–647, Sept. 2017. ISSN 0006-3444. doi: 10.1093/biomet/asx032.

C. C. Sheldon, D. T. Rouse, E. J. Finnegan, W. J. Peacock, and E. S. Dennis. The molecular basis of vernalization: The central role of FLOWERING LOCUS C (FLC). *PLANT BIOLOGY*, 97(6):6, Jan. 2000.

J. D. Storey, A. J. Bass, A. Dabney, D. Robinson, and G. Warnes. qvalue: Q-value

estimation for false discovery rate control, 2021. URL https://bioconductor.org/packages/qvalue/.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246. URL https://www.jstor.org/stable/2346178. Publisher: [Royal Statistical Society, Wiley].

P. Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, June 2001. ISSN 1573-2878. doi: 10.1023/A:1017501703105. URL https://doi.org/10.1023/A:1017501703105.

J. Wang, Q. Zhao, T. Hastie, and A. B. Owen. Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics*, 45(5):1863–1894, Oct. 2017. ISSN 0090-5364. doi: 10.1214/16-AOS1511. URL https://projecteuclid.org/euclid.aos/1509436821.

T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics (Oxford, England)*, 25(6):714–721, Mar. 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp041.

P. Zeng, X. Zhou, and S. Huang. Prediction of gene expression with cis-SNPs using mixed models and regularization methods. *BMC Genomics*, 18(1):368, May 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-3759-6. URL https://doi.org/10.1186/s12864-017-3759-6.

X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, July 2012. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.2310. URL `http://www.nature.com/articles/ng.2310`.

X. Zhou, P. Carbonetto, and M. Stephens. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics*, 9(2):e1003264, Feb. 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003264. URL `https://dx.plos.org/10.1371/journal.pgen.1003264`.