RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY
www.geneticepi.org

WILEY

# Block coordinate descent algorithm improves variable selection and estimation in error-in-variables regression

Célia Escribe[1,2,†]    |    Tianyuan Lu[1,3,†]    |    Julyan Keller-Baruch[1,4]    |
Vincenzo Forgetta[1]    |    Bowei Xiao[1,3]    |    J. Brent Richards[1,4,5,6]    |
Sahir Bhatnagar[5,7]    |    Karim Oualkacha[8]    |    Celia M. T. Greenwood[1,4,5,9]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Québec, Canada

[2]Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, United States

[3]Quantitative Life Sciences Program, McGill University, Montreal, Québec, Canada

[4]Department of Human Genetics, McGill University, Montreal, Québec, Canada

[5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Québec, Canada

[6]Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom

[7]Department of Diagnostic Radiology, McGill University, Montreal, Québec, Canada

[8]Département de Mathématiques, Université du Québec à Montréal, Montreal, Québec, Canada

[9]Gerald Bronfman Department of Oncology, McGill University, Montreal, Québec, Canada

**Correspondence**
Celia M. T. Greenwood, Lady Davis
Institute for Medical Research, Jewish
General Hospital, Montreal, QC, Canada.
Email: celia.greenwood@mcgill.ca

## Abstract

Medical research increasingly includes high-dimensional regression modeling with a need for error-in-variables methods. The *Convex Conditioned Lasso (CoCoLasso)* utilizes a reformulated *Lasso* objective function and an error-corrected cross-validation to enable error-in-variables regression, but requires heavy computations. Here, we develop a *Block coordinate Descent Convex Conditioned Lasso (BDCoCoLasso)* algorithm for modeling high-dimensional data that are only partially corrupted by measurement error. This algorithm separately optimizes the estimation of the uncorrupted and corrupted features in an iterative manner to reduce computational cost, with a specially calibrated formulation of cross-validation error. Through simulations, we show that the *BDCoCoLasso* algorithm successfully copes with much larger feature sets than CoCoLasso, and as expected, outperforms the naïve *Lasso* with enhanced estimation accuracy and consistency, as the intensity and complexity of measurement errors increase. Also, a new smoothly clipped absolute deviation penalization option is added that may be appropriate for some data sets. We apply the *BDCoCoLasso* algorithm to data selected from the UK

Biobank. We develop and showcase the utility of covariate-adjusted genetic risk scores for body mass index, bone mineral density, and lifespan. We demonstrate that by leveraging more information than the naïve *Lasso* in partially corrupted data, the *BDCoCoLasso* may achieve higher prediction accuracy. These innovations, together with an R package, *BDCoCoLasso*, make error-in-variables adjustments more accessible for high-dimensional data sets. We posit the *BDCoCoLasso* algorithm has the potential to be widely applied in various fields, including genomics-facilitated personalized medicine research.

## 1 | INTRODUCTION

Modern medical research is increasingly built on modeling of high-dimensional data. Sparse regression methods, such as the *Lasso* (Tibshirani, 1996), *Generalized Lasso* (Tibshirani et al., 2011), *Grouped Lasso* (Yuan & Lin, 2006), adaptive *Lasso* (Zou, 2006), and *Elastic Net* (Zou & Hastie, 2005), have been widely applied to perform estimation and variable selection at the same time. However, high-dimensional data sets often contain less precise measurements of phenotypes than those that might be available in smaller studies. For example, large biobanks often use billing codes from electronic health care records as proxy measures for a physician-made diagnosis. It is well known that applying naïve regression methods to predictor variables that are measured with error can lead to attenuation of effect estimates (Chesher, 1991; Rosenbaum et al., 2010). Analogously, questionnaire data from large cohorts often contain many missing values (Obermeyer & Emanuel, 2016). Removing subjects who are missing at least one measurement can easily lead to removal of most subjects when data are high dimensional.

Many error-in-variables solutions have been proposed. In addition to simple complete case analysis and pairwise deletion, more rigorous methods, such as expectation-maximization algorithms (Dempster, 1977; Schafer, 1997), multiple imputation methods (Buuren, 2011), and full information maximum likelihood estimation (Enders, 2001; Friedman et al., 2010), have been developed, but these computationally expensive methods cannot be easily extended to high-dimensional settings. In contrast, Loh and Wainwright (2011) developed a penalized method for error-in-variables regression. Within a properly chosen constraint radius, a projected gradient descent algorithm will converge to a small neighborhood of the set of all global minimizers, and is promising for variable selection in a high-dimensional setting (Loh & Wainwright, 2011). Nevertheless, proper

choice of this constraint radius depends on knowledge of the parameters yet to be estimated (Datta et al., 2017). Hence, Datta and Zou (2017) developed the *Convex Conditioned Lasso* (*CoCoLasso*) that does not require prior knowledge of the unknown parameters. The *CoCoLasso* algorithm is able to correct for both additive measurement error and missing data, and showed a substantial increase in estimation accuracy and stability compared with the naïve *Lasso*.

However, when the data are only partially corrupted (i.e., some features are free of measurement error), the *CoCoLasso* still performs estimation for all features in an undifferentiated manner, limiting the implementation of the approach for large feature sets due to the intensive matrix computations required. Such circumstances of partial corruption are common for genetic epidemiology studies based on large genotyped cohorts, where the genotypes are accurately measured by highly reliable high-throughput sequencing or microarrays, but lifestyle or clinical risk factors (except for age and sex) are measured with various types of error. For instance, in the UK Biobank, one of the largest health registries to date, participants had accurately measured hundreds of thousands of single nucleotide polymorphisms (SNPs) with little missing data, but most covariates based on questionnaires or health care records contained missing data (Bycroft et al., 2018). Samples with such corrupted covariates are usually discarded, potentially leading to underuse of information. Therefore, inspired by the *CoCoLasso*, we propose here a *Block coordinate Descent Convex Conditioned Lasso* (*BDCoCoLasso*) algorithm that makes it possible to perform higher-dimensional error-in-variables regressions by separately optimizing estimation of the parameter estimates for uncorrupted and corrupted features in an iterative manner. Our proposal requires the implementation of a carefully calibrated cross-validation strategy. Furthermore, we build in the smoothly clipped absolute deviation (SCAD) penalty

(Fan & Li, 2001) in the new algorithm. In simulations, we confirm that our algorithm provides equivalent results to the *CoCoLasso*, and demonstrates better performance than the naïve *Lasso*, with increasing benefit as the dimension increases. Although this approach will still encounter computational limitations for many corrupted features, we substantially enlarge the magnitude of problems that can be analyzed with an error-in-variables approach. We demonstrate the potential practical utility of the *BDCoCoLasso* by deriving covariate-adjusted genetic risk scores to predict body mass index, bone mineral density, and lifespan in a subset of the UK Biobank (Bycroft et al., 2018).

The rest of the manuscript is organized as follows. In Section 2, we briefly review the *CoCoLasso* method, and then we describe our new version that allows blocks of features with different corruption states—*BDCoCoLasso*. We describe simulation settings and results in Section 3. Section 4 illustrates the performance of our algorithm on the UK Biobank data.

## 2 | METHODS

In this section, we first review the principles of the *CoCoLasso*. We then seek to improve its computational efficiency and stability when the covariate matrix is partially corrupted or when different types of measurement error exist simultaneously, by implementing a block coordinate descent algorithm (Rosenbaum et al., 2013). We also implement a SCAD penalty (Fan & Li, 2001) to avoid overshrinkage when some features have strong effects.

### 2.1 | The *CoCoLasso*

Suppose a true covariate matrix $X_{n \times p}$, with $n$ observations and $p$ features, is measured as a corrupted covariate matrix $Z_{n \times p}$, where measurement error can be:

1. *Additive error*: $Z_{ij} = X_{ij} + A_{ij}$, where $A_{ij}$ represents additive error;
2. *Missing data*: $Z_{ij} = X_{ij}M_{ij}$, where $M_{ij} = 1$ or $M_{ij} = 0$.

It has been shown that using a classical *Lasso* with an objective function taking the form

$$\frac{1}{2n}\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1 \qquad (1)$$

can lead to biased estimation of $\beta$ (Datta et al., 2017; Loh & Wainwright, 2011), where $y_{n \times 1}$ is the continuous response.

Alternatively, this objective function can be reformulated as

$$\frac{1}{2}\beta'\Sigma\beta - \rho'\beta + \lambda\|\beta\|_1, \qquad (2)$$

where $\Sigma = \frac{1}{n}Z'Z$ and $\rho = \frac{1}{n}Z'y$. Loh and Wainwright (2011) proposed that $\Sigma$ and $\rho$ could be replaced by their unbiased estimators $\hat{\Sigma}$ and $\hat{\rho}$ such that $E(\hat{\Sigma}) = \Sigma$ and $E(\hat{\rho}) = \rho$. However, since the new covariance matrix $\hat{\Sigma}$ can have negative eigenvalues, particularly when the covariate matrix is high dimensional ($p \geq n$), the new optimization problem with the objective function

$$\frac{1}{2}\beta'\hat{\Sigma}\beta - \hat{\rho}'\beta + \lambda\|\beta\|_1 \qquad (3)$$

is not necessarily convex. Loh and Wainwright (2011) showed that by setting certain constraints on $\beta$, the problem could become convex, yet it is necessary to have prior knowledge of $\beta$ to find a suitable constraint.

Datta and Zou (2017) therefore proposed the *CoCoLasso* that adopts the adapted objective function but finds a nearest positive semidefinite matrix for $\hat{\Sigma}$:

$$\frac{1}{2}\beta'\tilde{\Sigma}\beta - \hat{\rho}'\beta + \lambda\|\beta\|_1, \qquad (4)$$

where $\tilde{\Sigma} = \text{argmin}_{\Sigma_1 \geq 0} \|\hat{\Sigma} - \Sigma_1\|_{\max}$. Here, the elementwise maximum norm for matrix $\Gamma$ is defined as $\|\Gamma\|_{\max} = \max|\Gamma_{ij}|$. This nearest positive semidefinite matrix can then be solved by an alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011).

### 2.2 | Two-block coordinate descent for partially corrupted covariate matrix

The *CoCoLasso* enables error-in-variables regression in general, but when the feature set is large, the required matrix calculations are demanding. Implementing a block coordinate descent could substantially improve the computational efficiency when the covariate matrix is only partially corrupted. Specifically, projection of the covariance matrix onto a positive semidefinite subspace, that is, $\tilde{\Sigma}$, within the *CoCoLasso*, requires multiple operations on matrices of dimension $p \times p$, which are order $o(p^3)$. In contrast, our *BDCoCoLasso* requires these operations only on the corrupted subblocks of the covariance matrix, which are anticipated to be much smaller. Suppose

the true covariate matrix $X_{n \times p}$ is now measured as $[X_{1_{n \times p_1}}, Z_{2_{n \times p_2}}]$, where $p_1 + p_2 = p$, $X_1$ is measured without error, and $Z_2$ is measured with error. We then need to estimate $\beta = (\beta_1, \beta_2)$ where $\beta_1$ is a coefficient vector for the noncorrupted covariates, and $\beta_2$ is a coefficient vector for the corrupted covariates. We derive the objective function as

$$\frac{1}{2n}\|y - X_1\beta_1 - Z_2\beta_2\|_2^2 + \lambda\|\beta_1\|_1 + \lambda\|\beta_2\|_1. \quad (5)$$

We conceive a two-step block coordinate descent algorithm based on (2)–(4):

1. We first consider $\beta_2$ fixed, and we solve

$$\operatorname*{argmin}_{\beta_1} \frac{1}{2}\beta_1'\Sigma_1\beta_1 - \tilde{\rho}_1'\beta_1 + \lambda\|\beta_1\|_1, \quad (6)$$

where $\tilde{\rho}_1 = \frac{1}{n}X_1'(y - \tilde{Z}_2\beta_2)$ and $\Sigma_1 = \frac{1}{n}X_1'X_1$. $\tilde{Z}_2$ is defined as

(a) in the additive error setting, $\tilde{Z}_2 = Z_2$;
(b) in the missing-error setting, specifically, we define a ratio matrix $R_{p_2 \times p_2}$ indicating the presence or absence of data as

$$R_{jl} = \frac{n_{jl}}{n},$$

where $n_{jl}$ is the number of samples for which both the $j$th and the $l$th features are measured and $n_{jj}$ is the number of samples for which the $j$th feature is measured. Note that $\tilde{Z}_2\beta_2$ is used to correct for measurement error in the corrupted covariates. We then have $\tilde{Z}_2 = Z_2\operatorname{diag}(1/R)$, that is, $\tilde{Z}_{2_{ij}} = \frac{Z_{2_{ij}}}{R_{jj}}$ for $i = 1, ..., n$ and $j = 1, ..., p_2$.

2. We next consider $\beta_1$ fixed, with a value optimized in the previous step, and we solve

$$\operatorname*{argmin}_{\beta_2} \frac{1}{2}\beta_2'\tilde{\Sigma}_2\beta_2 - \tilde{\rho}_2'\beta_2 + \lambda\|\beta_2\|_1, \quad (7)$$

where $\tilde{\rho}_2$ is an unbiased surrogate of $\frac{1}{n}Z_2'(y - X_1\beta_1)$ and $\tilde{\Sigma}_2$ is the nearest positive semidefinite matrix of $\hat{\Sigma}_2$. For $\tilde{\rho}_2$ and $\hat{\Sigma}_2$,

(a) in the additive error setting, $\tilde{\rho}_2 = \frac{1}{n}Z_2'(y - X_1\beta_1)$ and $\hat{\Sigma}_2 = \frac{1}{n}Z_2'Z_2 - \Sigma_A$, where $\Sigma_{A_{p_2 \times p_2}}$ is a known

variance–covariance matrix for features measured with additive error;

(b) in the missing error setting, $\tilde{\rho}_2 = \frac{1}{n}Z_2'(y - X_1\beta_1)$ $\operatorname{diag}(1/R)$ and $\hat{\Sigma}_2 = \frac{1}{n}Z_2'Z_2/R$. Here, $/$ represents elementwise division.

We then alternate between the two steps until convergence. Following similar arguments as in Datta et al. (2017), we can ensure that both problems are equivalent to a *Lasso* problem. The complete optimization procedure is described in Algorithm 1.

Of note, the estimation problem can be defined as finding the global solution for $(\beta_1, \beta_2)$, and our two-step procedure can be seen as equivalent to replacing $\Sigma_2$ by its nearest positive definite matrix, $\hat{\Sigma}_2$, in (5). Use of this substitution might not lead to a jointly convex problem. However, since both marginal problems (6) and (7) are convex, and both have suitable properties (i.e., both are strongly convex and smooth), our generalized alternating minimization algorithm can guarantee global minimization (Jain & Kar, 2017; Kelley, 1999).

---

**Algorithm 1** Two-block coordinate descent

**Input** $\Sigma_1, \tilde{\Sigma}_2, R, y, \lambda, X_1, Z_2$, error
**Initialize** $\beta_{01} \leftarrow \mathbf{0}$; $\beta_{02} \leftarrow \mathbf{0}$
**while** until convergence **do**
　**if** error = missing **then**
　　$\tilde{Z}_2 = Z_2\operatorname{diag}(1/R)$
　**end if**
　**if** error = additive **then**
　　$\tilde{Z}_2 = Z_2$
　**end if**
　$\tilde{\rho}_1 \leftarrow \frac{1}{n}X_1'(y - \tilde{Z}_2\beta_{02})$
　$\beta_1 \leftarrow \operatorname{argmin}_{\beta_1} \frac{1}{2}\beta_1'\Sigma_1\beta_1 - \tilde{\rho}_1'\beta_1 + \lambda\|\beta_1\|_1$
　**if** error = missing **then**
　　$\tilde{\rho}_2 \leftarrow \frac{1}{n}Z_2'(y - X_1\beta_1)\operatorname{diag}(1/R)$
　**end if**
　**if** error = additive **then**
　　$\tilde{\rho}_2 \leftarrow \frac{1}{n}Z_2'(y - X_1\beta_1)$
　**end if**
　$\beta_2 \leftarrow \operatorname{argmin}_{\beta_2} \frac{1}{2}\beta_2'\tilde{\Sigma}_2\beta_2 - \tilde{\rho}_2'\beta_2 + \lambda\|\beta_2\|_1$
　**Update** $\beta_{01} \leftarrow \beta_1$; $\beta_{02} \leftarrow \beta_2$
**end while**
**Output** $\beta_1, \beta_2$

---

Cross-validation to choose the penalization parameter, $\lambda$, must be appropriately implemented for the block implementation. Therefore, extending the concept in *CoCo-Lasso* (Datta et al., 2017), a $K$-fold cross-validated $\lambda$ can be obtained by minimizing the total cross-validation error while correcting for the two blocks separately,

$$\hat{\lambda} = \underset{\lambda}{\text{argmin}} \frac{1}{K} \sum_{k=1}^{k=K} \hat{\beta}_{k,1}(\lambda)' \Sigma_{k,1} \hat{\beta}_{k,1}(\lambda)$$
$$+ \hat{\beta}_{k,2}(\lambda)' \tilde{\Sigma}_{k,2} \hat{\beta}_{k,2}(\lambda) - 2\tilde{\rho}'_{k,1} \hat{\beta}_{k,1} - 2\tilde{\rho}'_{k,2} \hat{\beta}_{k,2}$$
$$+ 2\hat{\beta}'_{k,2} \hat{\Sigma}_{k,21} \hat{\beta}_{k,1}. \tag{8}$$

Here, $\hat{\beta}_{k,1}$ and $\hat{\beta}_{k,2}$ are estimated as described above for $\hat{\beta}_1$ and $\hat{\beta}_2$ based on data not in the $k$th-fold; $\Sigma_{k,1}$ and $\tilde{\Sigma}_{k,2}$ are derived as described above for $\Sigma_1$ and $\tilde{\Sigma}_2$ based on data in the $k$th-fold. $\hat{\Sigma}_{k,21}$ is an unbiased surrogate of $\Sigma_{k,21} = \frac{1}{n} Z'_{k,2} X_{k,1}$, where $Z_{k,2}$ and $X_{k,1}$ are in the $k$th-fold. More specifically,

1. in the additive error setting, where the additive error is centered to have zero mean, $\hat{\Sigma}_{k,21} = \frac{1}{n} Z'_{k,2} X_{k,1}$;
2. in the missing error setting, $\hat{\Sigma}_{k,21} = \frac{1}{n} \tilde{Z}'_{k,2} X_{k,1}$ where $\tilde{Z}_{k,2} = Z_{k,2} \text{diag}(1/R_k)$.

Although either an additive error setting or a missing error setting can be approached in the aforementioned two-step manner, data often contain variables subject to both types of errors. Therefore, we further propose a generalized algorithm that copes with a mixed error setting, described in Supporting Information.

## 2.3 | Implementation of a SCAD penalty

For potential application in scenarios where the causal variables are few but have large effect sizes, using the *Lasso* penalty may lead to overshrinkage (Fan & Li, 2001). To resolve this potential issue, we have also implemented a nonconcave SCAD penalty (Fan & Li, 2001). The SCAD penalty is given by

$$p_\lambda^{\text{SCAD}}(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \le \lambda, \\ -\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)} & \text{if } \lambda \le |\beta_j| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{otherwise} \end{cases} \tag{9}$$

and its first derivative with respect to $\beta_j$ is given by

$$p_\lambda^{'\text{SCAD}}(\beta_j) = \begin{cases} \lambda\text{sign}(\beta_j) & \text{if } |\beta_j| \le \lambda, \\ \frac{-\beta_j + a\lambda\text{sign}(\beta_j)}{a-1} & \text{if } \lambda \le |\beta_j| \le a\lambda, \\ 0 & \text{otherwise}. \end{cases} \tag{10}$$

Substituting the regular $L1$ penalty used in the *Lasso* by the SCAD penalty can retain large coefficients while shrinking smaller coefficients to zero. Thus, the SCAD penalty is able to produce a sparse solution and more accurate estimation for large coefficients.

Following Zou and Li (2008), we implement a local linear approximation of the penalization function:

$$p_\lambda^{\text{SCAD}}(|\beta_j|) \approx p_\lambda^{\text{SCAD}}(|\tilde{\beta}_j|) + p_\lambda^{'\text{SCAD}}(|\tilde{\beta}_j|)(|\beta_j| - |\tilde{\beta}_j|), \tag{11}$$

where $p_\lambda$ and $p_\lambda'$ are given by Equations (9) and (10), respectively, and $\tilde{\beta}_j$ is the estimate obtained from the previous iteration.

Equivalently,

$$p_\lambda^{\text{SCAD}}(|\beta_j|) = \lambda \frac{p_\lambda^{'\text{SCAD}}(|\tilde{\beta}_j|)}{\lambda} |\beta_j| = \lambda w_j |\beta_j|, \tag{12}$$

where a weight $w_j$ specific to the $j$th feature is introduced to the regular $L1$ penalty and is updated after each iteration. This implementation enables an adaptive *BDCoCoLasso*.

In principle, the hyperparameter $a$ in the SCAD penalty should be estimated through cross-validation. However, the resulting two-dimensional cross-validation would be computationally expensive. Fan and Li (2001) proposed that $a = 3.7$ should be suitable for many problems, and that the algorithm performance does not improve significantly with $a$ selected by data-driven approaches. We therefore set $a = 3.7$ in all simulations described below.

In addition to the SCAD penalty, other weighting schemes, such as the minimax concave penalty (Zhang, 2010), could be implemented in the future for improved generalizability.

## 3 | SIMULATION STUDY

Simulations were designed to explore the performance of *BDCoCoLasso* as a function of the number and proportion of corrupted features. Furthermore, we wanted to ensure that our results matched *CoCoLasso* when both methods could be implemented, that is, for fairly modest $p$, and a single type of error.

## 3.1 | Simulation design

We first simulated an uncorrupted covariate matrix $X_{n \times p}$ from a multivariate normal distribution with $n$ observations, zero mean, and a predefined correlation structure across $p$ features. We explored a

lower-dimensional setting ($n = 10,000$ and $p = 200$) and a higher-dimensional setting ($n = 1000$ and $p = 2000$) in combination with two common covariance matrix designs to introduce correlation between features ($\Sigma_X$):

1. *An autoregressive setting*: $\Sigma_{X_{ij}} = 0.5^{|i-j|}$.

2. *A symmetric setting*: $\Sigma_{X_{ij}} = 0.5 + 0.5I_{i=j}$.

We then generated the response as

$$y = X\beta_0 + \epsilon = (X_1, Z_2)\beta_0 + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right). \tag{13}$$

To ensure a realistic signal-to-noise ratio, we set $\sigma_\epsilon = 2$. When assessing the performance of the *CoCoLasso* algorithm, Datta and Zou used $\beta_0 = (3, 1.5, 0, 0, 2, 0, ..., 0)'$ to generate strong signals from only a few features. Likewise, to start with a simulation that was similar to theirs, we set

$$\beta_0 = (\underbrace{3, 1.5, 0, 0, 2}_{\text{without error}} 0, ..., 0, \underbrace{2, 0, 0, 1.5, 3}_{\text{with error}})',$$

where three of the features measured without error and three of the features measured with error were assigned to be causal with relatively large effect sizes.

Since we anticipate that this algorithm will be useful in large cohorts where $n > p$, and anticipating multiple associated features with small effect sizes, we simulated more scenarios with $n = 10,000$ and $p = 200$. We assigned different fractions of features to be causal (5% or 20%), and created higher dimensionality ($p = 200, 500,$ or $1000$) while sampling $\beta_0$ from a standardized normal distribution $\mathcal{N}(0, 1)$.

Next, we introduced different types of error to the covariate matrix:

1. For the additive error setting, the corrupted design matrix was generated as $Z_2 = X_2 + A$ where $A \sim \mathcal{N}(0, \tau I)$. We explored different $\tau$ parameters in combination with different fractions (at least 10%) of features measured with additive error.
2. For the missing error setting, the corrupted design matrix was generated as $Z_2 = X_2 \odot M$ where each element of $M$ follows a Bernoulli distribution: $m_{ij} \sim \mathcal{B}(1 - r)$ where $r$ is the missing rate. We explored different values for the missing rate $r$ in combination with different fractions (at least 10%) of features measured with missing data.
3. For the mixed error setting, we generated $y = (X_1, Z_2, Z_3)\beta_0 + \epsilon$ where $Z_2$ and $Z_3$ were generated as the additive error setting and the missing error setting, respectively. We explored different combinations of $\tau$ for $Z_2$ and $r$ for $Z_3$.

All parameters used in the simulations are summarized in Table 1. In all simulations, simulation of $Z$ and $y$ was repeated for the same $\beta_0$ twice to create a training data set for model fitting and a test data set of equal size ($n = 10,000$ or $1000$) for assessing prediction accuracy. We used fivefold cross-validation in the training data to optimize the $\lambda$ parameter. We repeated each simulation scenario 100 times. Data were then analyzed with *BDCoCoLasso*, naïve *Lasso*, and for the simulation scenarios with strong signals in Table 1,

**TABLE 1** Summary of simulation design

| Err. | No. Obs. | No. Fts. | No. Causal Fts. | % Fts. with Additive Err. | % Fts. Missing | $\beta$ | $\tau$ | $r$ |
|---|---|---|---|---|---|---|---|---|
| Additive | 10,000 | 200 | 6 | 10 | | $\beta_0$ | $\tau_0$ | |
| Missing | 10,000 | 200 | 6 | | 10 | $\beta_0$ | | $r_0$ |
| Additive | 1000 | 2000 | 6 | 10 | | $\beta_0$ | $\tau_0$ | |
| Missing | 1000 | 2000 | 6 | | 10 | $\beta_0$ | | $r_{0,\text{high-dim.}}$ |
| Additive | 10,000 | 200 | 5%, 20% | 10, 20, 50 | | $\sim \mathcal{N}(0,1)$ | 0.2 | |
| Missing | 10,000 | 200 | 5%, 20% | | 10, 20, 50 | $\sim \mathcal{N}(0,1)$ | | 0.2 |
| Additive | 10,000 | 500, 1000 | 5% | 10, 20, 50 | | $\sim \mathcal{N}(0,1)$ | 0.2 | |
| Missing | 10,000 | 500, 1000 | 5% | | 10, 20, 50 | $\sim \mathcal{N}(0,1)$ | | 0.2 |
| Mixed | 10,000 | 200 | 5% | 10, 20, 50 | 10, 20, 50 | $\sim \mathcal{N}(0,1)$ | 0.2, 0.5, 0.8 | 0.2, 0.5 |

*Note:* All simulations were replicated 100 times in each of the autoregressive covariance setting and the symmetric covariance setting, respectively.
$\beta_0 = (3, 1.5, 0, 0, 2, 0, ..., 0, 2, 0, 0, 1.5, 3)$, $\tau_0 \in \{0, 0.05, 0.10, ..., 0.70, 0.75, 0.80\}$, and $r_{0,\text{high-dim.}} \in \{0, 0.05, 0.10, ..., 0.30, 0.35, 0.40\}$ as high missing rates lead to completely missing data in some features with a small number of observations.
Abbreviations: Err., errors; Fts., features; Obs., observations.

also with *BDCoCoLasso-SCAD*, using the variant of SCAD penalty, as well as the adaptive *Lasso*. All methods were implemented using a 2.6-GHz quad-core processor with 32 GB of random access memory. The data sets were also analyzed with *CoCoLasso* for comparison of computational cost. The four following criteria were used to compare the performance of different methods:

1. Computational time (for some scenarios).
2. Total-mean-square error in the training data set: $\|\beta_0 - \hat{\beta}\|_2$.
3. False-positive rate (FPR), that is, the number of truly zero coefficients estimated to be nonzero.
4. *Sparsity*: The fraction of features correctly estimated to be zero or nonzero.
5. Variance explained ($R^2$) in the test data set: $(\frac{\text{Cov}(X_{\text{test}}\hat{\beta}, y_{\text{test}})}{\sqrt{\text{Var}(X_{\text{test}}\hat{\beta})\text{Var}(y_{\text{test}})}})^2$.

When the naïve *Lasso* and the adaptive *Lasso* were applied to corrupted data in the additive error setting, estimates could be directly obtained, without taking the measurement error into account. However, in the missing error setting, since removing all observations with missing data would occasionally lead to insufficient numbers of samples, we used the classical mean imputation method to impute missing data. The adaptive weight $\hat{w}_{j,\text{adaptive}}$ for the $j$th feature in the adaptive *Lasso* was obtained by Ridge regression with fivefold cross-validation: $\hat{w}_{j,\text{adaptive}} = \frac{1}{|\hat{\beta}_{j,\text{Ridge}}|}$. We did not apply more sophisticated imputation methods, such as the Multivariate Imputation by Chained Equations (Buuren, 2011), since they would have prohibitive computational costs in a high-dimensional setting.

## 3.2 | Simulation results

### 3.2.1 | *BDCoCoLasso* outperforms *Lasso* when covariate matrix is partially corrupted, and can cope with much larger data sets than the *CoCoLasso*

To ensure validity of our implementation, we analyzed the same data with *BDCoCoLasso* as well as the *CoCoLasso* algorithm without the block coordinate descent procedures. As expected, we found that all the estimates obtained by the *BDCoCoLasso* were numerically the same as those obtained by the *CoCoLasso* (numerical discrepancies were below the convergence tolerance), while the latter had a higher computational cost (Figure 1c,d). The computational efficiency of the

*BDCoCoLasso* was more prominent in higher-dimensional data with stronger correlations between features. For instance, on 1000 observations of 2000 features simulated with a symmetric covariance structure, it took the *BDCoCoLasso* approximately 10 min, on average, to construct the model, whereas the ordinary *CoCoLasso* had an average running time above 10 h.

Also as expected based on Datta et al. (2017), the *BDCoCoLasso* achieved better performance than the naïve *Lasso* in most scenarios. In both the lower-dimensional setting ($n = 10, 000$ and $p = 200$) and the higher-dimensional setting ($n = 1000$ and $p = 2000$), with 10% features measured with error and strong signals, the *BDCoCoLasso* yielded smaller total-mean-square error, lower FPR, and higher sparsity compared with the naïve *Lasso* (Figures 1, S1, and S2). The *BDCoCoLasso* was relatively insensitive to the increase in additive error rates or missing rates, while the naïve *Lasso* had considerably worse performance as corruption rates increased. Although the naïve *Lasso* achieved a slightly better prediction accuracy in the test data set with small values of $\tau$ or $r$, its predictive performance deteriorated more rapidly than the *BDCoCoLasso* (Figures S1 and S2).

Moreover, implementing a SCAD penalty in the lower-dimensional setting ($n = 10, 000$ and $p = 200$) with strong signals further improved the estimation accuracy of the *BDCoCoLasso*. As indicated in Figures 1a and S1, the *BDCoCoLasso* with SCAD penalty yielded smaller total-mean-square error with a 100% sparsity when no measurement error occurred ($\tau = 0$ or $r = 0$). Further, it consistently outperformed the *BDCoCoLasso* implementing an $L1$ penalty, the naïve *Lasso* as well as the adaptive *Lasso* with increasing $\tau$ and $r$. Notably, while the adaptive *Lasso* had comparable performance to the *BDCoCoLasso* with SCAD penalty, and was slightly better than *BDCoCoLasso* with an $L1$ penalty when the intensity of measurement error was considerably weak, its accuracy could attenuate substantially with a higher $\tau$ or $r$. However, the SCAD penalty implementation had a low prediction accuracy despite consistently achieving an FPR close to 0 and almost 100% sparsity (Figures S1 and S2). This situation can arise when there are many highly correlated predictor variables. Since SCAD has good performance in variable selection, it does not retain many noncausal variables. In contrast, prediction models created by some other methods may retain several noncausal variables that are highly correlated with the true causal predictors; this obviously leads to worse metrics for sensitivity and sparsity, but can in fact lead to better $R^2$ even in test data.
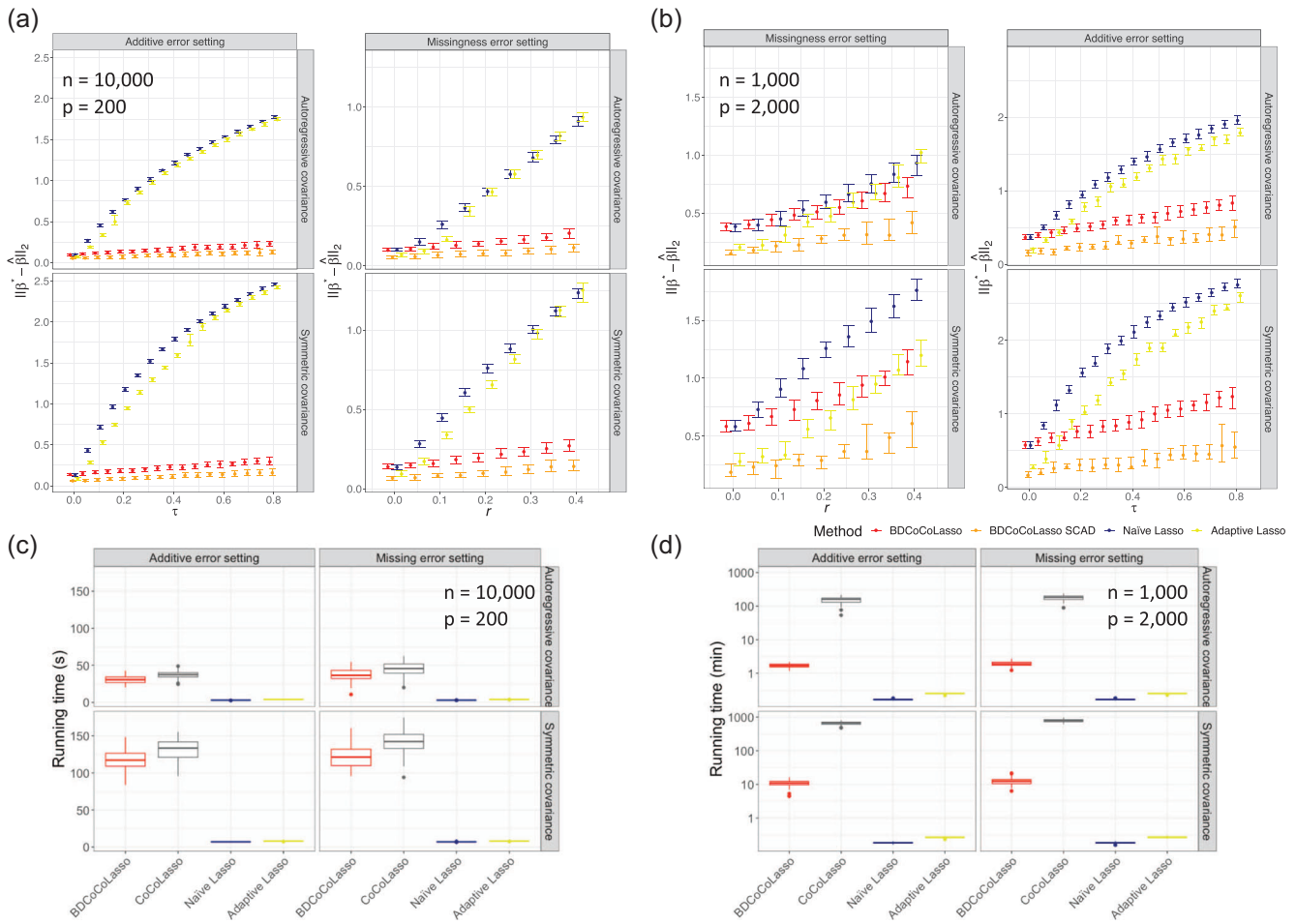
**FIGURE 1** Performance of *BDCoCoLasso*, *BDCoCoLasso-SCAD*, and *Lasso* with increasing additive error ($\tau$) and missing rates ($r$) for the simulation scenarios in the first four rows of Table 1, where six features were assigned to be causal with large effect sizes. Panels (a) and (b) show squared bias. Dots denote median total-mean-square error and error bars show the interquartile range based on 100 replicates in each simulation setting. When $\tau = 0$ or $r = 0$, no measurement error exists. All simulations were based on (a) 10,000 observations of 200 features or (b) 1000 observations of 2000 features where 10% of the features were measured with error. In (b), as $r$ increased, frequently all observations of a feature were missing. Therefore, scenarios with $r > 0.4$ were not explored. Comparison of running time in (c) the lower-dimensional settings and (d) the higher-dimensional settings indicates the substantially improved computational efficiency of the *BDCoCoLasso* over *CoCoLasso*. Running time was summarized over all replicates in each simulation setting. All methods were implemented using a 2.6-GHz quad-core processor. BDCoCoLasso, Block coordinate Descent Convex Conditioned Lasso; CoCoLasso, Convex Conditioned Lasso; SCAD, smoothly clipped absolute deviation

## 3.2.2 | The *BDCoCoLasso* also outperforms naïve *Lasso* with weakened signals, increased error rate, and increased dimensionality

In the lower-dimensional setting ($n = 10,000$ and $p = 200$), when the magnitude of causal feature effect sizes was reduced, and more causal features were introduced, the estimation accuracy and stability for the naïve *Lasso* decreased substantially (Figures 2 and S3). In contrast, although an increase in the number of causal features and the correlation between features rendered the signals more elusive and resulted in an increase in FPR and a decrease in sparsity, the *BDCoCoLasso* always maintained better estimation accuracy than the naïve *Lasso* with better

consistency across replicates (Figures 2 and S3). Also as expected, the *BDCoCoLasso* was clearly less sensitive to changes in the proportion of features measured with error (Figure 2). Such an improved estimation accuracy persisted when the covariate matrix contained more features ($p = 500$ or 1000; Figures 3 and S4).

## 3.2.3 | The *BDCoCoLasso* handles measurement error with mixed types

The new three-block coordinate descent algorithm (Supporting Information) copes seamlessly with coexistence of both types of error (Figures 4 and S5). As demonstrated in
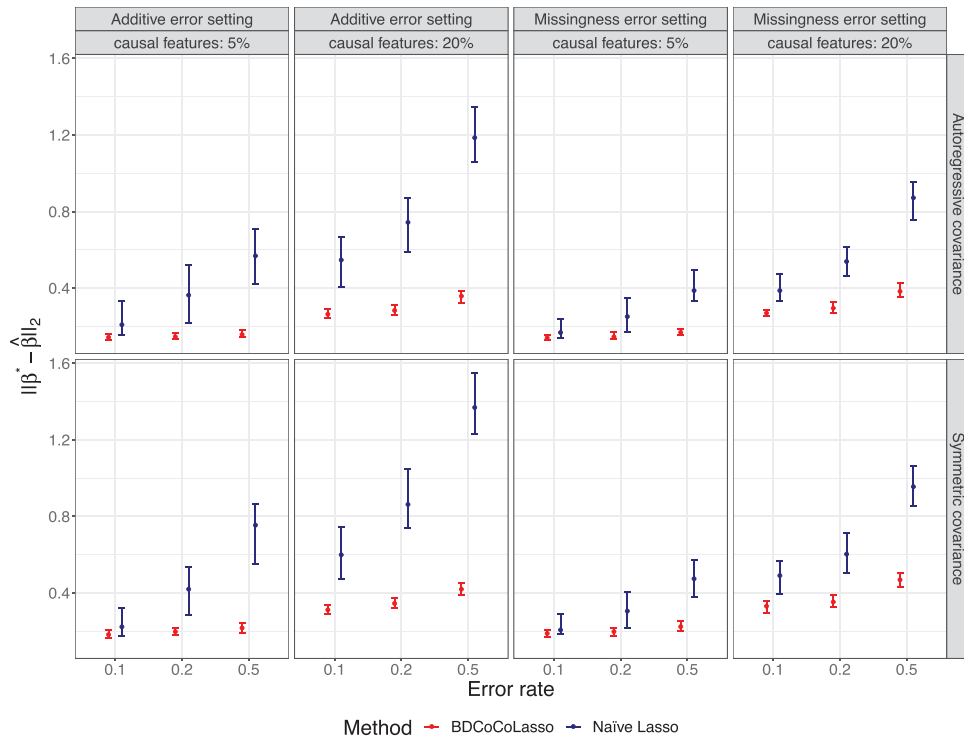
**FIGURE 2** Squared bias of *BDCoCoLasso* and *Lasso* with higher error rates and weaker signals (rows 5 and 6 in Table 1). Dots and triangles denote median total-mean-square error and error bars denote interquartile range based on 100 replicates in each simulation setting. Error rates denote the fractions of features measured with either additive error or missing data. Causal features denote the fractions of features assigned to be causal. Effect sizes of causal features were sampled from a standardized normal distribution. All simulations were based on 10,000 observations of 200 features. BDCoCoLasso, Block coordinate Descent Convex Conditioned Lasso
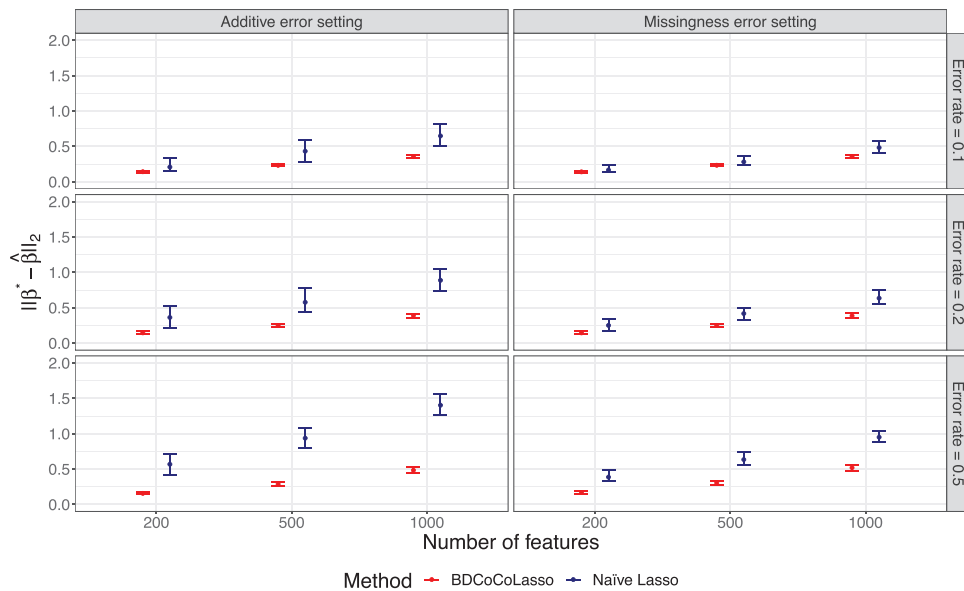


**FIGURE 3** Squared bias of *BDCoCoLasso* and *Lasso* with high-dimensional feature sets of 500 or 1000 features (rows 7 and 8 in Table 1). Dots denote median total-mean-square error and error bars denote interquartile ranges based on 100 replicates in each simulation setting. Error rates denote the fractions of features measured with either additive error or missing data. In all simulation settings, 5% of the features were assigned to be causal with effect sizes sampled from a standardized normal distribution. Features were simulated to have an autoregressive covariance matrix. BDCoCoLasso, Block coordinate Descent Convex Conditioned Lasso
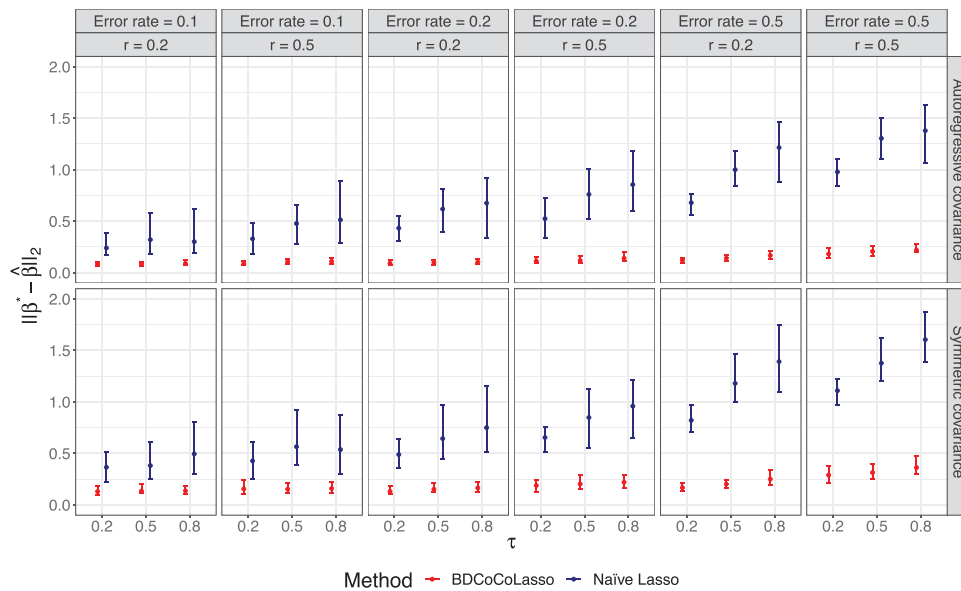
**FIGURE 4** Squared bias of *BDCoCoLasso* and *Lasso* in the mixed error setting using a three-block coordinate descent algorithm. Dots and triangles denote median total-mean-square error and error bars denote interquartile range based on 100 replicates in each simulation setting. Additive error rates and missing error rates were set to be equivalent taking values in 0.1, 0.2, or 0.5. When both the additive error rate and the missing error rate are 0.5, all features are measured with error and a two-block coordinate descent algorithm supplants the three-block coordinate descent algorithm. All simulations were based on 10,000 observations of 200 features. A 5% of the features was assigned to be causal with effect sizes sampled from a standardized normal distribution. BDCoCoLasso, Block coordinate Descent Convex Conditioned Lasso

previous figures, the *BDCoCoLasso* achieved higher estimation accuracy than *Lasso* in all combinations of $\tau$, $r$, and error rates. Its advantage became more prominent when the covariate matrix was more corrupted with a higher $\tau$, $r$, and error rate. In particular, when all features were measured with error, a two-block coordinate descent iterating between the additive-error block and the missing-error block retained its superiority over the naïve *Lasso*.

## 4 | REAL DATA APPLICATION EXAMPLES IN THE UK BIOBANK

The UK Biobank provides deep genetic and phenotypic data collected from nearly 500,000 participants between 2006 and 2010, and has enabled many important advances in human genetics and health care (Bycroft et al., 2018). One important advance is the development of genetic risk scores, which have demonstrated the potential in improving risk screening and possibly guiding prevention and intervention (Khera et al., 2018; Lu et al., 2020; Lu, Forgetta, Keller-Baruch, et al., 2021; Lu, Forgetta, Wu, et al., 2021; Lu, Zhou, et al., 2021). Notably, several genetic risk scores have been developed using *Lasso* (Lu, Forgetta, Keller-Baruch, et al., 2021; Lu, Forgetta, Wu, et al., 2021; Lu, Zhou, et al., 2021). Similar to most large-scale cohort studies, measurement errors,

especially missingness, affected a substantial proportion of clinical and lifestyle variables. We thus tested whether the *BDCoCoLasso* could help improve the predictive performance and clinical utility of covariate-adjusted genetic risk scores compared with the naïve *Lasso* or the adaptive *Lasso*.

For the purpose of testing *BDCoCoLasso* in a reasonably large high-dimensional setting, we randomly selected 4500 unrelated individuals from the UK Biobank of white British ancestry with self-reported age, sex, measured body mass index, bone mineral density, maternal and paternal living status or age of death, and 30 clinical and lifestyle variables (Figure 5a). We randomly split this data set into a training data set, including 3000 individuals possibly with missing data, and a test data set, including 1500 individuals without missing data. For all three examples, genotypes had been imputed to the Haplotype Reference Consortium panel (McCarthy et al., 2016).

## 4.1 | Predicting body mass index with accurate genotype variables and corrupted clinical and lifestyle measurements

Obesity is a highly polygenic trait involving multiple genes of small or moderate effects (Speliotes et al., 2010; Willer et al., 2009). Previously, the genetic basis of obesity was explored
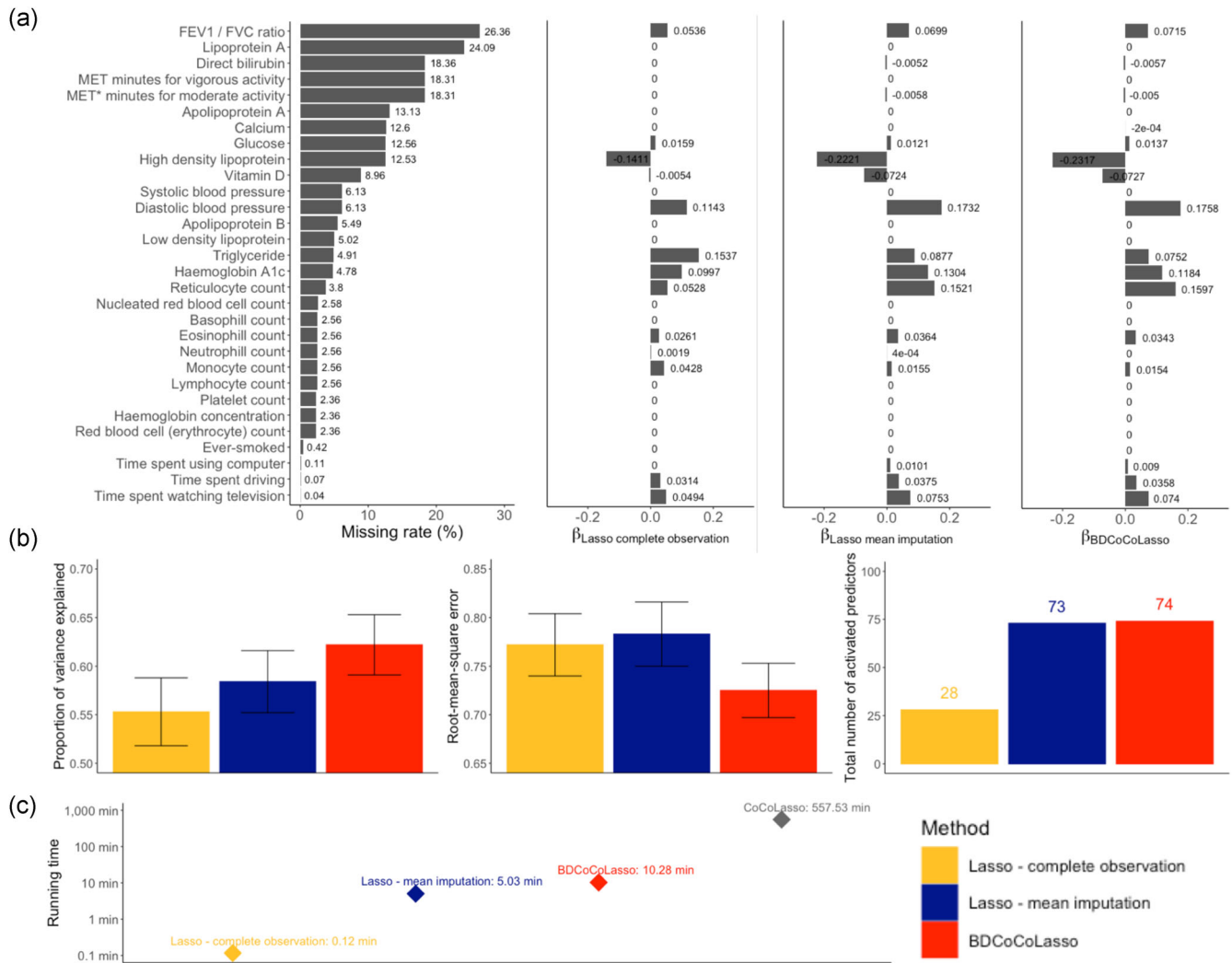
**FIGURE 5** Comparison of *Lasso* and *BDCoCoLasso* in developing a covariate-adjusted genetic risk score for the body mass index $z$ score. (a) Summary of missing rates for the covariates in the training data set (left). The test data set does not have missing data. The coefficients of these covariates estimated by *Lasso* based on complete observations (second panel; $N = 895$) or mean imputation (third panel; $N = 3000$), and by *BDCoCoLasso* (rightmost panel; $N = 3000$) on the training data set are aligned. (b) Comparison of model metrics for the *Lasso* and *BDCoCoLasso* models. Standard errors of the proportion of variance explained and root-mean-square error were generated using 100 bootstrap replicates of the test data set ($N = 1500$). The five models were evaluated on the same bootstrap replicates. (c) Comparison of running time in logarithmic scale. All methods were implemented using a 2.6-GHz quad-core processor. *BDCoCoLasso*, Block coordinate Descent Convex Conditioned Lasso; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity

by a genome-wide association study of body mass index, a widely used measure to define obesity, in 322,154 individuals of European ancestry from the Genetic Investigation of ANthropometric Traits Consortium (Locke et al., 2015). Despite tens of independent genetic risk loci identified, many clinical and lifestyle risk factors are also strongly associated with body mass index (Marti et al., 2004; Speakman, 2004), yet measurements of these risk factors may be missing in large-scale cohort studies. Such missingness limits the investigation of the joint effects of the genetic and nongenetic risk factors for obesity. We therefore applied the *BDCoCo-Lasso* to a subset of data from the UK Biobank (Bycroft et al., 2018) to examine whether incorporating variables that

previously had to be discarded due to missingness could improve the prediction of body mass index.

An existing large meta-analysis of genome-wide association studies identified 1882 SNPs strongly associated with body mass index ($p < 5 \times 10^{-8}$) (Locke et al., 2015), hence we retrieved genotypes of these SNPs as candidate genetic predictors. Among these SNPs that were representative of causal signals, many were correlated due to linkage disequilibrium. Thus, $L1$ penalty was adopted for variable selection. Genetic variants together with age and sex were considered features measured without error or missing data, while the other clinical or lifestyle features containing missing values were further processed

by *BDCoCoLasso*. In addition, we compared the performance of *BDCoCoLasso* with two types of implementation of naïve *Lasso* and adaptive *Lasso* (with adaptive weights obtained from fivefold cross-validation Ridge regression as in the simulation study), respectively: We built *Lasso* models either on only 895 individuals with no missing data in the training data set, or on the entire mean-imputed training data set. These five models were constructed based on the same fivefold cross-validation, such that in each fold the optimization was high dimensional, and were evaluated on the independent test data set. Proportion of variance explained and prediction root-mean-square error were examined based on 100 bootstrap replicates.

As anticipated, because of a largely compromised sample size, the *Lasso* models relying on only the complete observations explained the least proportion of variance in body mass index in the test data set with the least number of predictors activated (Figure 5b). On the other hand, the naïve *Lasso* model and the adaptive *Lasso* model with mean imputation derived similar estimates for clinical and lifestyle covariates as the *BDCoCoLasso* model, that were substantially different from those estimated by the *Lasso* models on complete observations (Figure 5a). However, the *BDCoCoLasso* achieved significantly higher proportion of variance explained (0.622 vs. 0.584 of the naïve *Lasso*, paired $t$ test $p$ value of bootstrap replicates $= 5.5 \times 10^{-46}$; 0.622 vs. 0.583 of the adaptive *Lasso*, paired $t$ test $p = 1.1 \times 10^{-49}$) and significantly lower prediction root-mean-square error (0.725 vs. 0.783 of the naïve *Lasso*, paired $t$ test $p$ value of bootstrap replicates $= 9.1 \times 10^{-71}$; 0.725 vs. 0.784 of the adaptive *Lasso*, paired $t$ test $p = 1.6 \times 10^{-73}$) than this *Lasso* model on the test data set. Notably, the *BDCoCoLasso* only required twice as much the running time as the mean-imputed naïve *Lasso* model, whereas the *CoCoLasso* without the block coordinate descent procedures had more than 100 times higher time cost to yield the same parameter estimates (Figure 5c).

## 4.2 | Predicting bone mineral density and fracture risk

Osteoporotic fractures affect up to 1 in 3 women and 1 in 5 men aged above 50 years, and incur a heavy socioeconomic burden among elderly populations (Kanis et al., 2000). Therefore, good predictions of the risk of osteoporotic fracture are essential to public health management. Bone mineral density is a key indicator of bone mass and bone quality, and has been included in successful risk factor-based fracture risk prediction tools, such as FRAX (Kanis, 2002; Kanis et al., 2008). Recently, it has been shown that, when combined with clinical risk factors, genetically predicted bone mineral density could significantly improve the

predictive performance in identifying individuals at an elevated risk of fracture (Lu, Forgetta, Keller-Baruch, et al., 2021). Therefore, we attempted to leverage *BDCoCoLasso* to further improve the prediction of bone mineral density and fracture risk.

We retrieved 7307 SNPs strongly associated with bone mineral density (estimated by quantitative ultrasound speed of sound and broadband ultrasound attenuation); SNPs had demonstrated $p < 5 \times 10^{-8}$ in a previous genome-wide association study (Morris et al., 2019). We implemented *BDCoCoLasso*, the naïve *Lasso* (with complete data or mean-imputed data) and the adaptive *Lasso* (with complete data or mean-imputed data) as in Section 4.1 with the same training data set and the same clinical and lifestyle features. We found that the covariate-adjusted genetic risk score constructed using the *BDCoCoLasso* again had the highest proportion of variance explained and the lowest prediction root-mean-square error for bone mineral density on the independent test data set, and its computational cost was tremendously reduced compared with the *CoCoLasso* (Figure S6).

Moreover, among the 1500 individuals in the test data set, 170 self-reported or had a medical record of major osteoporotic fractures affecting hip, radius/ulna, humerus, or vertebrae upon recruitment. The score constructed by *BDCoCoLasso* also exhibited the strongest discriminative power in identifying individuals who experienced fractures, with an area under the receiver operating characteristic curve (AUROC) of 0.571 and an area under the precision-recall curve (AUPRC) of 0.132 (Figure 6). In contrast, the naïve *Lasso* with mean imputation achieved the best performance among the four naïve *Lasso* or adaptive *Lasso* implementations, but only obtained an AUROC of 0.554 and an AUPRC of 0.123 (Figure 6), respectively.

## 4.3 | Predicting human lifespan

Longevity is a highly complex trait in which genetics plays a debatable role (van den Berg et al., 2017). It was only recently that genes and genetic variants influencing extreme longevity (Deelen et al., 2019) or human lifespan (Timmers et al., 2019) have been systematically identified in large-scale genome-wide association studies. We tested whether a covariate-adjusted genetic risk score could predict lifespan and inform lifetime risk of death.

We retrieved 462 SNPs strongly associated with human lifespan ($p < 5 \times 10^{-8}$) identified in a recent genome-wide association study (Timmers et al., 2019) as candidate genetic predictors. Because the majority of the UK Biobank participants were alive at the time of the latest
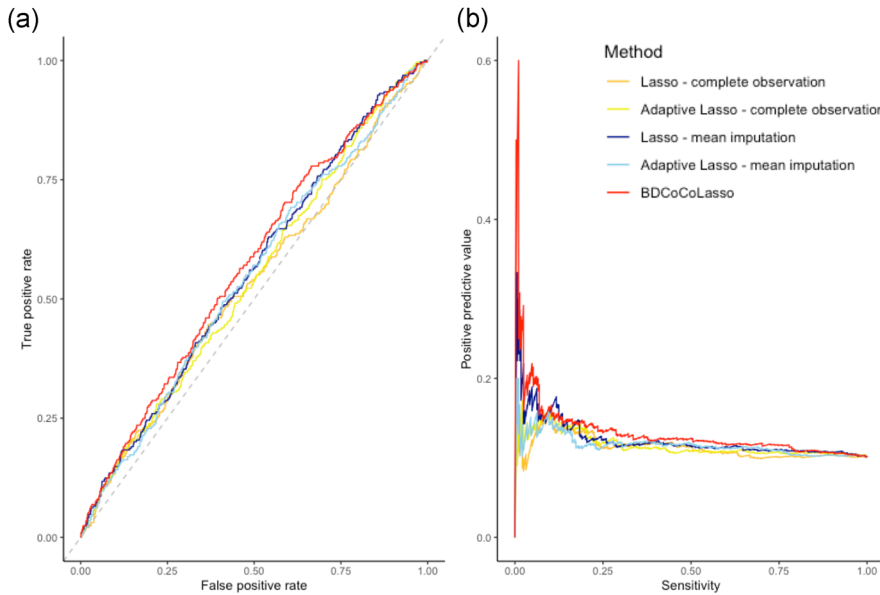
(a)

(b)



**FIGURE 6** Comparison of predictive performance of covariate-adjusted genetic risk scores for bone mineral density in identifying individuals who had fractures. (a) Receiver operating characteristic curves and (b) precision-recall curves. Scores were evaluated based on the test data set ($N = 1500$). Other model metrics are provided in Figure S6. BDCoCoLasso, Block coordinate Descent Convex Conditioned Lasso

follow-up, we sought to predict parental lifespan instead. Since *BDCoCoLasso* has not been adapted to time-to-event outcomes, we created two subtraining data sets containing 1814 individuals whose mother had died and 2254 individuals whose father had died, from the original training data set of 3000 individuals. We trained models to predict maternal or paternal age of death separately, to account for potential sex-specific effects (Timmers et al., 2019). We again implemented *BDCoCoLasso*, the naïve *Lasso* (with complete data or mean-imputed data), and the adaptive *Lasso* (with complete data or mean-imputed data) as above with the same clinical and lifestyle features. Notably, the naïve *Lasso* models with complete data did not select any of the predictors, probably due to the reduced sample size (Figure S7).

Next, we tested the predictive performance of these covariate-adjusted genetic risk scores based on the test data set using Cox regression models. Of the 1500 individuals in the test data set, 902 mothers, and 1109 fathers had died upon recruitment. Although a genetic risk score based on offspring genotypes is not an ideal way to estimate parental genetic predispositions, our *BDCoCo-Lasso*-based scores achieved modest discriminative power in identifying individuals whose parents lived longer in the test data set, and outperformed the other naïve *Lasso* or adaptive *Lasso* models (Figure 7). Specifically, a one standard deviation decrease in the maternal score (corresponding to a shorter predicted lifespan) was associated with a lifetime hazard ratio for time to death of 1.104 (95% CI, 1.032–1.181) while a one standard deviation decrease in the paternal score was associated with a lifetime hazard ratio of 1.071 (95% CI, 1.009–1.137). In contrast, the runner-up score for maternal lifespan using an adaptive *Lasso* with mean imputation had a hazard

ratio of 1.084 (95% CI, 1.013–1.161) per standard deviation increase and the runner-up score for paternal lifespan using naïve *Lasso* with mean imputation had a hazard ratio of 1.068 (95% CI, 1.006–1.134) per standard deviation increase.

## 5 | DISCUSSION

With the increasing availability of large population-based cohorts, developing rigorous methods for model estimation and variable selection is a pressing need in contemporary medical research. The *CoCoLasso* algorithm proposed by Datta and Zou (2017) utilizes a reformulated form of the *Lasso* objective function with a modified covariance estimator to allow for high-dimensional error-in-variables regression. More recent studies have combined the principles of the *CoCoLasso* with other techniques that render more complicated scenarios tractable. For example, Brown et al. (2019) developed a Measurement Error Boosting algorithm with a measurement error-corrected score function to enable Poisson, Gamma, and Wald. However, no algorithm to our knowledge specifically targets data that are only partially corrupted by measurement or have mixed error types, but such characteristics are common in most large-scale genomics and medical studies. In this study, we developed a block coordinate descent algorithm as an extension to the *CoCo-Lasso* algorithm to improve both computational efficiency and estimation accuracy. We also implemented an optional SCAD penalty for further improved model estimation and variable selection when the signals are strong. These adaptations make it possible to use error-in-variables penalized models for data sets with large feature
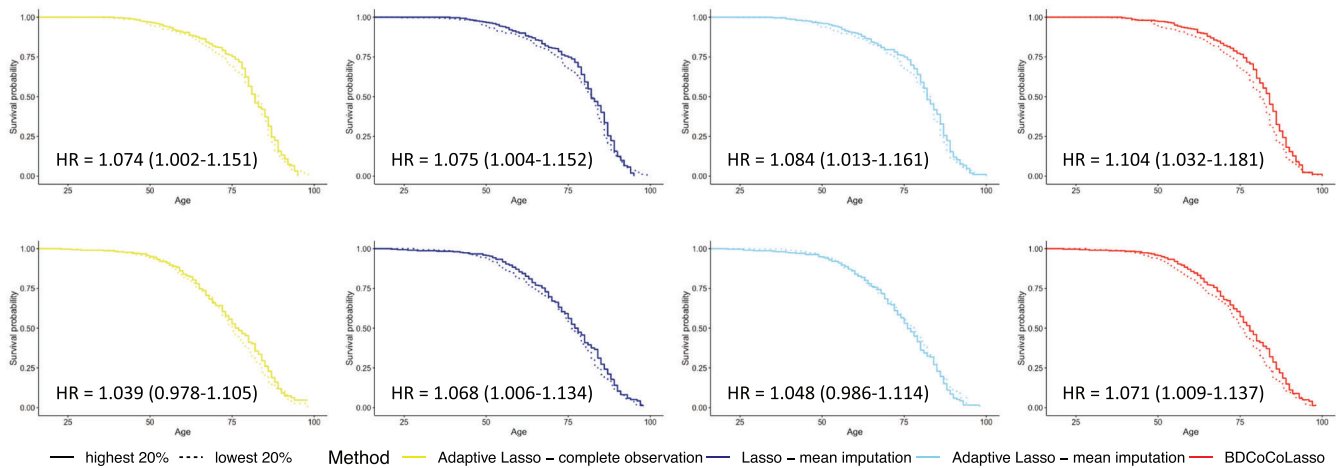
**FIGURE 7** Comparison of predictive performance of covariate-adjusted genetic risk scores for lifespan. (a) Kaplan–Meier curves for time to maternal death and (b) Kaplan–Meier curves for time to paternal death. Parents of individuals with the top 20% highest scores (predicted to be the most likely to live longer) and the top 20% lowest (predicted to be the least likely to live longer) were compared. Hazard ratios (HRs) were estimated based on standardized covariate-adjusted genetic risk score using Cox regression models. Scores were evaluated based on the test data set ($N = 1500$). Other model metrics are provided in Figure S7

dimension, as long as the number of corrupted features remains modest. Computational time depends linearly on sample size, but is cubic as a function of the number of corrupted features. Therefore, although these developments achieve an important step towards being able to analyze large-scale data, to work with data of the size of the UK Biobank, while allowing for corrupted data, additional developments would be required. Perhaps by combining these approaches with new methods for working with biobank data at scale (Bi et al., 2020; Jiang et al., 2019; Qian et al., 2020), it may be possible to achieve the orders-of-magnitude expansions required.

In multifaceted simulations, the *BDCoCoLasso* algorithm substantially outperformed the naïve *Lasso* (as expected), achieving smaller total-mean-square error, lower FPR, and higher sparsity. The *BDCoCoLasso* was also less sensitive to increases in the intensity of additive error and/or missing rate, fraction of features measured with error, dimensionality as well as reduction in the magnitude of signals. We further derived covariate-adjusted genetic risk scores for body mass index, bone mineral density, and parental lifespan in the UK Biobank and showed that the *BDCoCoLasso* leveraged more information than the naïve *Lasso* without the need to discard missing data or perform imputation, and achieved better prediction accuracy. It should be noted that, while we worked on well-genotyped and well-imputed genotypes (INFO > 0.3), poorly imputed SNPs that were filtered out before our analysis could potentially be considered as measured with error, and hence used more effectively by our algorithm. We do not pursue this here since most genetic studies analyze only well-imputed genotypes.

Considering that genomics-facilitated personalized medicine is booming, and large data sets are being rapidly released containing both accurate genotyping information and other partially corrupted features, we posit the *BDCoCoLasso* algorithm has the potential to be applied in various medical research settings and we have provided a freely available R package for public use.

Since our algorithm utilizes corrupted covariates, *BDCoCoLasso* on an extremely small sample size may have less stable performance than the naïve *Lasso*. Particularly with small $n$-large $p$ situations, results should be carefully examined and data perturbed to assess stability. If cross-validation were to be employed, the number of folds should be chosen such that each fold contains sufficient observations. Our simulations with $n = 1000$ (fivefold cross-validation) experienced no trouble, but with $n = 100$ or 200 (and $p$ double these values, using fivefold cross-validation), convergence was not always achieved. Leave-one-out cross-validation may be an appropriate alternative under such circumstances. Extra caution should also be taken when implementing the SCAD penalty in a high-dimensional setting if the features are correlated, as it may introduce instability in parameter estimation or prediction.

Given that our algorithm exhibited better model sparsity in multiple simulation settings, it may be combined with various approaches for post-selection inference, including but limited to those proposed by Lee et al. (2016, with closed-form $p$ values and confidence intervals), Taylor et al. (2016, forward stepwise regression and least angle regression), and possibly in the future, Taylor and Tibshirani (2018, generalized regression models). The improved control of false discovery rate

may benefit various fields, including genetic epidemiology studies.

Our algorithm has some important limitations. First, it assumes that each feature can harbor at most one type of error (either additive or missing error) and does not cope with coexistence of both types of error in one feature. Therefore, *BDCoCoLasso* could be combined with a complete case analysis removing features with both types of error but a low missing rate, or an imputation of only the features with a low missing rate to control potential bias. Second, a useful extension of our algorithm could be to allow for varying penalty factors for different coefficient blocks, for example, $\lambda_1$ for $\beta_1$ and $\lambda_2$ for $\beta_2$ in Equation (5). However, without strong prior knowledge of the features, selecting optimal penalty factors with cross-validation becomes non-trivial and requires future investigations. Third, the ADMM algorithm becomes unstable when the missing rate is high. Replacing the max norm by a Frobenius norm when defining the nearest positive semidefinite matrix, or down-weighting features with a high missing rate in the ADMM algorithm may boost its stability; in fact, the recently developed high missing Lasso (*HMLasso*) algorithm has successfully adopted similar concepts to handle scenarios where features are subject to very high missing rates (Takada et al., 2019). Our package includes an option with *HMLasso* features, although we did not observe a clear benefit to this adaptation in our simulations. Furthermore, in the additive error setting, similar to the *CoCoLasso* (Datta et al., 2017), our algorithm requires knowledge about the variance of the error, and therefore it is essential to be able to find relevant literature, such as measures of precision of an instrument used for measurement. Lastly, we noted that with a very large feature dimension and strong correlations between features (e.g., a symmetric covariance matrix for $X$), the algorithm became time intensive. Enhanced memory handling and parallelization may assist in enabling and accelerating computation in higher-dimensional data sets with more complex correlation structures. Nevertheless, the algorithm copes extremely efficiently with large sample sizes—our UK Biobank example analyzed over thousands of samples and could easily have analyzed more.

## CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

## DATA AVAILABILITY STATEMENT

We have developed an open-sourced R package *BDCoCoLasso* available at https://mcgill.ca/statisticalgenetics/soft ware. This package includes an implementation of the *Co-CoLasso* algorithm as well as our improved algorithm with the block coordinate descent implementation, an option for the SCAD penalty, and an option for the *HMLasso* adaptation. Scripts for conducting the simulation study are available upon reasonable request to the correspondence author. Data from the UK Biobank (Bycroft et al., 2018) are available upon successful project application.

## ORCID

*Tianyuan Lu* [ID] http://orcid.org/0000-0002-5664-5698
*Karim Oualkacha* [ID] http://orcid.org/0000-0002-9911-079X
*Celia M. T. Greenwood* [ID] https://orcid.org/0000-0002-2427-5696

## REFERENCES

Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S., & Lee, S. (2020). A fast and accurate method for genome-wide time-to-event data analysis and its application to UK Biobank. *The American Journal of Human Genetics*, *107*(2), 222–233.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Brown, B., Weaver, T., & Wolfson, J. (2019). MEBoost: Variable selection in the presence of measurement error. *Statistics in Medicine*, *38*(15), 2705–2718.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(1), 1–67.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203.

Chesher, A. (1991). The effect of measurement error. *Biometrika*, *78*(3), 451–462.

Datta, A., & Zou, H. (2017). CoCoLasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, *45*(6), 2400–2426.

Deelen, J., Evans, D. S., Arking, D. E., Tesi, N., Nygaard, M., Liu, X., Wojczynski, M. K., Biggs, M. L., van Der Spek, A., Atzmon, G., Ware, E. B., Sarnowski, C., Smith, A. V.,

Seppälä, I., Cordell, H. J., Dose, J., Amin, N., Arnold, A. M., Ayers, K. L., Murabito, J. M., ... Murabito, J. M. (2019). A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nature Communications*, *10*(1), 1–14.

Dempster, A. P. (1977). Maximum likelihood estimation from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *39*, 1–38.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*(4), 352.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Jain, P., & Kar, P. (2017). Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, *10*(3–4), 142–336.

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). *A resource-efficient tool for mixed model association analysis of large-scale data* (Technical Report). Nature Publishing Group.

Kanis, J., Johnell, O., Odén, A., Johansson, H., & McCloskey, E. (2008). Frax$^{TM}$ and the assessment of fracture probability in men and women from the UK. *Osteoporosis International*, *19*(4), 385–397.

Kanis, J., Johnell, O., Oden, A., Sernbo, I., Redlund-Johnell, I., Dawson, A., De Laet, C., & Jonsson, B. (2000). Long-term risk of osteoporotic fracture in malmö. *Osteoporosis International*, *11*(8), 669–674.

Kanis, J. A. (2002). Diagnosis of osteoporosis and assessment of fracture risk. *The Lancet*, *359*(9321), 1929–1936.

Kelley, C. T. (1999). *Iterative methods for optimization*. SIAM.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907–927.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J. C., ... Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206.

Loh, P. L., & Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in Neural Information Processing Systems*, *40*(3), 1637–1664.

Lu, T., Forgetta, V., Keller-Baruch, J., Nethander, M., Bennett, D., Forest, M., Bhatnagar, S., Walters, R. G., Lin, K., Chen, Z., Li, L., Karlsson, M., Mellström, D., Orwoll, E., McCloskey, E. V., Kanis, J. A., Leslie, W. D., Clarke, R. J., Ohlsson, C., ... Brent Richards, J. (2021). Improved prediction of fracture risk leveraging a genome-wide polygenic risk score. *Genome Medicine*, *13*(1), 1–15.

Lu, T., Forgetta, V., Oriana, H., Mokry, L., Gregory, M., Thanassoulis, G., Greenwood, C. M., & Richards, J. B. (2020). Polygenic risk for coronary heart disease acts through atherosclerosis in type 2 diabetes. *Cardiovascular Diabetology*, *19*(1), 1–10.

Lu, T., Forgetta, V., Wu, H., Perry, J. R., Ong, K. K., Greenwood, C. M., Timpson, N. J., Manousaki, D., & Richards, J. B. (2021). A polygenic risk score to predict future adult short stature among children. *The Journal of Clinical Endocrinology & Metabolism*, *106*(7), 1918–1928.

Lu, T., Zhou, S., Wu, H., Forgetta, V., Greenwood, C. M., & Richards, J. B. (2021). Individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening. *Genetics in Medicine*, *23*(3), 508–515.

Marti, A., Moreno-Aliaga, M., Hebebrand, J., & Martinez, J. (2004). Genes, lifestyles and obesity. *International Journal of Obesity*, *28*(3), S29–S36.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., ... Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279.

Morris, J. A., Kemp, J. P., Youlten, S. E., Laurent, L., Logan, J. G., Chai, R. C., Vulpescu, N. A., Forgetta, V., Kleinman, A., Mohanty, S. T., Marcelo Sergio, C., Quinn, J., Nguyen-Yamamoto, L., Luco, A.-L., Vijay, J., Simon, M.-M., Pramatarova, A., Medina-Gomez, C., Trajanoska, K., ... Brent Richards, J. (2019). An atlas of genetic influences on osteoporosis in humans and mice. *Nature Genetics*, *51*(2), 258–266.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, *375*(13), 1216.

Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., Rivas, M. A., & Hastie, T. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genetics*, *16*(10), e1009141.

Rosenbaum, M., & Tsybakov, A. B. (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics*, *38*(5), 2620–2651.

Rosenbaum, M., & Tsybakov, A. B. (2013). Improved matrix uncertainty selector. In M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii & M. H. Maathuis (eds.), *From probability to statistics and back: High-dimensional models and processes—A Festschrift in honor of Jon A. Wellner* (pp. 276–290). Institute of Mathematical Statistics.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Speakman, J. R. (2004). Obesity: The integrated roles of environment and genetics. *The Journal of Nutrition*, *134*(8), 2090S–2105S.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Mägi, R., Randall, J. C., Vedantam, S., Winkler, T. W., Qi, L., Workalemahu, T., Heid, I. M., Steinthorsdottir, V., Stringham, H. M., Weedon, M. N., Wheeler, E., ... Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new

loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948.

Takada, M., Fujisawa, H., & Nishikawa, T. (2019). HMLasso: Lasso with high missing rate. *Machine Learning*, *1050*, 3541–3547. https://doi.org/10.24963/ijcai.2019/491

Taylor, J., & Tibshirani, R. (2018). Post-selection inference for penalized likelihood models. *Canadian Journal of Statistics*, *46*(1), 41–61.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R. J., & Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, *39*(3), 1335–1371.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600–620.

Timmers, P. R., Mounier, N., Lall, K., Fischer, K., Ning, Z., Feng, X., Bretherick, A. D., Clark, D. W., Shen, X., Esko, T., Kutalik, Z., Wilson, J. F., & Joshi, P. K., & eQTLGen Consortium. (2019). Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife*, *8*, e39856.

van den Berg, N., Beekman, M., Smith, K. R., Janssens, A., & Slagboom, P. E. (2017). Historical demography and longevity genetics: Back to the future. *Ageing Research Reviews*, *38*, 28–39.

Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., Lettre, G., Lim, N., Lyon, H. N., McCarroll, S. A., Papadakis, K., Qi, L., Randall, J. C., Roccasecca, R. M., & Sanna, S., ... Genetic Investigation of ANthropometric Traits Consortium. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics*, *41*(1), 25.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, *36*(4), 1509.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.