

Large-Scale Sparse Regression for Multiple Responses with Applications to UK Biobank

Junyang Qian¹, Yosuke Tanigawa², Ruilin Li³,
Robert Tibshirani^{1,2}, Manuel A. Rivas^{*2} and Trevor Hastie^{†1,2}

¹Department of Statistics, Stanford University

²Department of Biomedical Data Science, Stanford University

³Institute for Computational and Mathematical Engineering, Stanford University

Abstract

In high-dimensional regression problems, often a relatively small subset of the features are relevant for predicting the outcome, and methods that impose sparsity on the solution are popular. When multiple correlated outcomes are available (multitask), reduced rank regression is an effective way to borrow strength and capture latent structures that underlie the data. Our proposal is motivated by the UK Biobank population-based cohort study, where we are faced with large-scale, ultrahigh-dimensional features, and have access to a large number of outcomes (phenotypes): lifestyle measures, biomarkers, and disease outcomes. We are hence led to fit sparse reduced-rank regression models, using computational strategies that allow us to scale to problems of this size. We use an iterative algorithm that alternates between solving the sparse regression problem and solving the reduced rank decomposition. For the sparse regression component, we propose a scalable iterative algorithm based on adaptive screening that leverages the sparsity assumption and enables us to focus on solving much smaller sub-problems. The full solution is reconstructed and tested via an optimality condition to make sure it is a valid solution for the original problem. We further extend the method to cope with practical issues such as the inclusion of confounding variables and imputation of missing values among the phenotypes. Experiments on both synthetic data and the UK Biobank data demonstrate the effectiveness of the method and the algorithm. We present `multiSnynet` package, available at <http://github.com/junyangq/multiSnynet> that works on top of PLINK2 files, which we anticipate to be a valuable tool for generating polygenic risk scores from human genetic studies.

Contents

1	Introduction	3
1.1	Reduced-Rank Regression for Multiple Responses	4
1.2	Sparse Models in High-Dimensional Problems	6

*Corresponding author: mrivas@stanford.edu

†Corresponding author: hastie@stanford.edu

33	2 Sparse Reduced-Rank Regression	6
34	3 Fast Algorithms for Large-Scale and Ultrahigh-Dimensional Problems	7
35	3.1 Alternating Minimization	7
36	3.2 Variable Screening for Ultrahigh-Dimensional Problems	8
37	3.2.1 Screening Strategies	9
38	3.2.2 Optimality Condition	10
39	3.3 Computational Considerations	11
40	3.3.1 Initialization and Warm Start	11
41	3.3.2 Early Stopping	12
42	3.4 Extensions	12
43	3.4.1 Standardization	12
44	3.4.2 Weighting	12
45	3.4.3 Adjustment Covariates	13
46	3.4.4 Missing Values	14
47	3.4.5 Lazy Reduced Rank Regression	15
48	3.5 Full Algorithm	15
49	4 Convergence Analysis	17
50	5 Simulation Studies	18
51	6 Real Data Application: UK Biobank	19
52	6.1 Asthma and 7 Blood Biomarkers	21
53	6.2 35 Biomarkers	22
54	7 Related Work	25
55	8 Summary and Discussion	26
56	References	27
57	A Additional Proofs	32
58	A.1 Proof of Lemma 1	32
59	A.2 Proof of Lemma 2	32
60	A.3 Proof of Theorem 2	32
61	B Connection with CCA	33
62	C Additional Experiments	34
63	D Additional Information on the Methods	39
64	D.1 Compliance with ethical regulations and informed consent	39
65	D.2 Population stratification in UK Biobank	40
66	D.3 Variant annotation and quality control	40

1 Introduction

The past two decades have witnessed rapid growth in the amount of data available to us. Many areas such as genomics, neuroscience, economics and Internet services have been producing increasingly larger datasets that have high dimension, large sample size, or both. A variety of statistical methods and computational tools have been developed to accommodate this change so that we are able to extract valuable information and insight from these massive datasets (Hastie et al., 2009; Efron, Hastie, 2016; Dean, Ghemawat, 2008; Zaharia et al., 2010; Abadi et al., 2016).

One major motivating application for this work is the study of data from population-scale cohorts like UK Biobank with genetic data from over one million genetic variants and phenotype data from thousands of phenotypes in over 500,000 individuals (Bycroft et al., 2018). These data present unprecedented opportunities to explore very comprehensive genetic relationships with phenotypes of interest. In particular, the subset of tasks we are interested in is the prediction of a person's phenotype value, such as disease affection status, based on his or her genetic variants.

Genome-wide association studies (GWAS) is a very powerful and widely used framework for identifying genetic variants that are associated with a given phenotype. See, for example, Visscher et al. (2017) and the references therein. It is based on the results of univariate marginal regression over all candidate variants and tries to find a subset of significant ones. While being computationally efficient and easy to interpret, GWAS has fairly limited prediction performance because at most one predictor can present in the model. If prediction performance is our main concern, it is natural to consider the class of multivariate methods, i.e. that which considers multiple variants simultaneously. In the past, *wide* data were prevalent where only a limited number, like thousands, of samples were available. In this regime, some sophisticated multivariate methods could be applicable, though they have to more or less deal with dimension reduction or variable selection. In this setting, we handle hundreds of thousands samples and even more variables. In such cases, statistical methods and computational algorithms become equally important because only efficient algorithmic design will allow for the application of sophisticated statistical modeling. Recently, we introduced some algorithms addressing these challenges. In particular, Qian et al. (2019) proposed an iterative screening framework that is able to fit the exact lasso/elastic-net solution path in large-scale and ultrahigh-dimensional settings, and demonstrate competitive computational efficiency and superior prediction performance over previous methods.

In this paper, we consider the scenarios where multivariate responses are available in addition to the multiple predictors, and propose a suite of statistical methods and efficient algorithms that allow us to further improve the statistical performance in this large n and large p regime. Some characteristics we want to leverage and challenges we want to solve include:

Statistics There are thousands of phenotypes available in the UK Biobank. Many of them are highly correlated with each other and can have a lot of overlap in their driving factors. By treating them separately, we lose this information that could have been used to stabilize our model estimation. The benefit of building a joint model can be seen from the following simplified model. Suppose all the outcomes $\mathbf{y}^k, k = 1, \dots, q$ are independent noisy observations of a shared factor $\mathbf{u} = \mathbf{X}\beta$ such that $\mathbf{y}^k = \mathbf{u} + \mathbf{e}^k$. It is easy to see that by taking an average across all the outcomes, we obtain a less noisy response $\bar{\mathbf{y}}$, and this will give us more accurate parameter estimation and better prediction than the model built on any of the single outcome. The assumption of such latent structure is an important approach to capturing the correlation structure among the outcomes and can bring in a significant reduction in variance if the data indeed behave in a similar way. We will

111 formalize this belief and build a model on top of it. In addition, in the presence of high-dimensional
112 features, we will follow the “bet on sparsity” principle (Hastie et al., 2009), and assume that only
113 a subset of the predictors are relevant to the prediction.

114 Therefore, the statistical model we will build features two major assumptions: **low-rank** in the
115 signal and **sparse effect**. Furthermore, we will introduce integrated steps to systematically deal
116 with confounders and missing values.

117 **Computation** On a large-scale dataset, building a multivariate model can pose great computa-
118 tional challenges. For example, loading the entire UK Biobank dataset into memory with double
119 precision will take more than one terabyte of space, while typically most existing statistical com-
120 puting tools assume that the data are already sitting in memory. Even if large memory is available,
121 one can always encounter data or construct features so that it becomes insufficient. Hence, instead
122 of expecting sufficient memory space, we would like to find a scalable solution that is less restricted
123 by the size of physical memory.

124 There is a dynamic data access mechanism provided by the operating system called memory
125 mapping (Bovet, Cesati, 2005) that allows for easy access to larger-than-memory data on the disk.
126 In essence, it carries a chunk of data from disk to memory when needed and swap some old chunks
127 of data out of memory when it is full. In principle, we could add a layer of memory mapping on
128 top of all the procedures and then access the data as if they were in memory. However, there is
129 one important practical component that should never be ignored: disk I/O. This is known to be
130 expensive in the operating system and can greatly delay the computation if frequent disk I/Os are
131 involved. For this reason, we do not pursue first-order gradient-based methods such as stochastic
132 gradient descent (Bottou, 2010) or dual averaging (Xiao, 2010; Duchi et al., 2011) because it can
133 take a large number of passes over the data for the objective function to converge to the optimum.

134 To address this, we design the algorithm so that it needs as few full passes over the data as
135 possible while solving the exact objective. In particular, by leveraging the sparsity assumption,
136 we propose an adaptive screening approach that allows us to strategically select a small subset of
137 variables into memory, do intensive computation on the subset, and then verify the validity of all
138 the left-out variables. The last step is important because we want to guarantee that the solution
139 obtained from the algorithm is a valid solution to the original full problem.

140 1.1 Reduced-Rank Regression for Multiple Responses

In the standard multivariate linear regression model, given a model matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$
and a multivariate response matrix $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_q) \in \mathbb{R}^{n \times q}$, we assume that

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

141 where each row of $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_q)$ is assumed to be an independent sample from some multivariate
142 Gaussian distribution $\mathbf{E}^{(i)} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_E)$. When $n \geq q$, it is easy to see that a maximum likelihood
143 estimator (MLE) can be found by solving a least squares problem with multiple outcomes, i.e.

$$\hat{\mathbf{B}} \in \underset{\mathbf{B} \in \mathbb{R}^{p \times q}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2, \quad (1)$$

144 where $\|\mathbf{A}\|_F^2 = \sum_{i=1}^n \sum_{j=1}^m \mathbf{A}_{ij}^2$ is the squared Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$. When $n \geq p$
145 and \mathbf{X} has full rank, (1) has the closed-form solution $\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$. Notice that this is

146 equivalent to solving q single-response regression problems separately.

147

However, in many scenarios, there can be some correlation structure in the signals that we can capture to improve the statistical efficiency of the estimator. One approach to modeling the correlation is to assume that there is a set of latent factors that act as the drivers for all the outcomes. When we assume that the dependencies of the latent factors on the raw features and the outcomes on the latent factors are both linear, it is equivalent to making a low-rank assumption on the coefficient matrix. Reduced-rank regression (Anderson, 1951, hereafter RRR) assumes that the coefficient matrix \mathbf{B} has a fixed rank $r \leq \min(p, q)$, or

$$\mathbf{B} = \mathbf{U}\mathbf{V}^\top,$$

148 where $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_r) \in \mathbb{R}^{p \times r}$, $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_q)^\top \in \mathbb{R}^{q \times r}$.¹ With the decomposed coefficient
 149 matrices, an alternative way to express the multivariate model is to assume that there exists a set
 150 of latent factors $\{\mathbf{z}_\ell \in \mathbb{R}^n : 1 \leq \ell \leq r\}$ such that for each ℓ ,

$$\begin{aligned} \mathbf{z}_\ell &= \mathbf{X}\mathbf{u}_\ell, \\ \mathbf{y}_k &= \mathbf{Z}\mathbf{v}_k + \mathbf{e}_k. \end{aligned}$$

151 Figure 1 gives a visualization of the dependency structure described above. It can also be seen as a a
 152 multilayer perceptron (MLP) with linear activation and one hidden layer, or multitask learning with
 153 bottleneck. We notice that under the decomposition, the parameters are not identifiable. In fact, if
 154 we apply any nonsingular linear transformation $\mathbf{M} \in \mathbb{R}^{r \times r}$ such that $\mathbf{V}' = \mathbf{V}\mathbf{M}^\top$ and $\mathbf{U}' = \mathbf{U}\mathbf{M}^{-1}$,
 155 it yields the same model but different parameters. As a result, we also have an infinite number of
 156 MLEs.

157 Under the rank constraint, an explicit global solution can be obtained. Let $\mathbf{M}\mathbf{D}\mathbf{N}^\top$ be the singular
 158 value decomposition (SVD) of $(\mathbf{X}^\top\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^\top\mathbf{Y}$, a set of solution is given by $\hat{\mathbf{U}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\mathbf{N}$,
 159 $\hat{\mathbf{V}} = \mathbf{N}$. Velu, Reinsel (2013) has a comprehensive discussion on the model under classical large n
 160 settings.

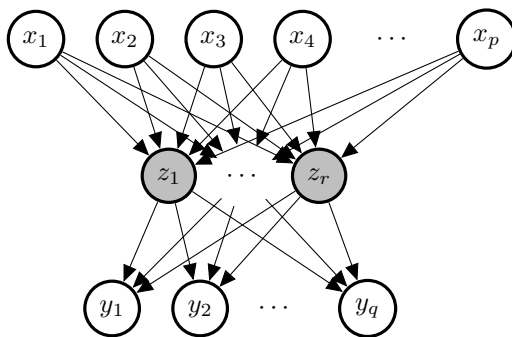


Figure 1: Diagram of the reduced rank regression. The nodes in grey are latent variables. The arrows represent the dependency structure. Known as *multitask* learning in the machine learning community.

¹We use \mathbf{v}_k^\top to represent the k th row of \mathbf{V} for convenience.

1.2 Sparse Models in High-Dimensional Problems

In the setting of high-dimensional problems where $p > n$, the original low-rank coefficient matrix \mathbf{B} can be unidentifiable. Often sparsity is assumed in the coefficients to model the belief that only a subset of the features are relevant to the outcomes. To find such a sparse estimate of the coefficients, a widely used approach is to add an appropriate non-smooth penalty to the original objective function to encourage the desired sparsity structure. Common choices include the lasso penalty (Tibshirani, 1996), the elastic-net penalty (Zou, Hastie, 2005) or the group lasso penalty (Yuan, Lin, 2006). There has been a great amount of work studying the consistency of estimation and model selection under such settings. See Greenshtein, Ritov (2004); Meinshausen, Bühlmann (2006); Zhao, Yu (2006); Bach (2008); Wainwright (2009); Bickel et al. (2009); Obozinski et al. (2011); Bühlmann, Van De Geer (2011) and references therein. In particular, the group lasso, as the name suggests, encourages group-level sparsity induced by the following penalty term:

$$P_g(\beta) = \sum_{j=1}^J \|\beta_j\|_2,$$

where $\beta_j \in \mathbb{R}^{p_j}$ is the subvector corresponding the j th group of variables and $\|\beta_j\|_2 = \sqrt{\sum_{\ell=1}^{p_j} \beta_{j,\ell}^2}$ is the vector ℓ_2 -norm. The ℓ_2 -norm enforces that if the fitted model has $\|\hat{\beta}_j\|_2 = 0$, all the elements in $\hat{\beta}_j$ will be 0, and otherwise with probability one all the elements will be nonzero. This yields a desired group-level selection in many applications. Throughout the paper, we will adopt the group lasso penalty, defining each predictor's coefficients across all outcomes as a distinct group, in order to achieve homogeneous sparsity across multiple outcomes. In addition to variable selection for better prediction and interpretation, we will also see the computational advantages we leverage to develop an efficient algorithm.

2 Sparse Reduced-Rank Regression

Given a rank r , we are going to solve the following penalized rank-constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{B}_{j\cdot}\|_2, \\ \text{s.t.} \quad & \text{rank}(\mathbf{B}) \leq r. \end{aligned} \tag{2}$$

Alternatively, we can decompose the matrix explicitly as $\mathbf{B} = \mathbf{U}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{p \times r}$, $\mathbf{V} \in \mathbb{R}^{q \times r}$. It can be shown that the problem above is equivalent to the Sparse Reduced Rank Regression (SRRR) proposed by Chen, Huang (2012):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}_{j\cdot}\|_2, \\ \text{s.t.} \quad & \mathbf{V}^\top \mathbf{V} = \mathbf{I}. \end{aligned} \tag{3}$$

Alternating minimization was proposed by Chen, Huang (2012) to solve this non-convex optimization problem, where two algorithms were considered: subgradient descent and a variational method.

177 The subgradient method was shown to be faster when $p \gg n$ and the variational method faster
 178 when $n \gg p$. However, in each iteration, the computational complexity of either method is at
 179 least quadratic in the number of variables p . It makes the problem almost intractable in ultrahigh-
 180 dimensional problems, which is common, for example, in modern genetic studies. Moreover, to
 181 obtain a model with good prediction performance, we are interested in solving the problem over
 182 multiple λ 's rather than a single one. For such purposes, we design a path algorithm with adaptive
 183 variable screening that will be both memory and computationally efficient.

184 3 Fast Algorithms for Large-Scale and Ultrahigh-Dimensional 185 Problems

186 First, we present a naive version of the path solution, which will be the basis of our subsequent
 187 development. The path is defined on a decreasing sequence of λ values $\lambda_{\max} = \lambda_1 > \lambda_2 > \dots >$
 188 $\lambda_L \geq 0$, where λ_{\max} is often defined by one that leads to the trivial (e.g. all zero) solution and the
 189 rest are often determined by an equally spaced array on the log scale. In particular, for Problem
 190 (2), we are able to figure out the exact lower bound of λ_{\max} for which the solution is trivial.

Lemma 1. *In problem (2), if $r > 0$, the maximum λ that results in a nontrivial solution $\hat{B}(\lambda)$ is*

$$\lambda_{\max} = \max_{1 \leq j \leq p} \|\mathbf{x}_j^\top \mathbf{Y}\|_2.$$

191 The proof is straightforward, which is a result of the Karush–Kuhn–Tucker (KKT) condition
 192 (See Boyd et al. (2004) for more details). We present the full argument in Appendix A.1. The
 193 naive path algorithm tries to solve the problem independently across different λ values.

194 3.1 Alternating Minimization

195 The algorithm is described in Algorithm 1. For each λ value, it applies alternating minimization
 196 to Problem (3) till convergence.

197 In the V-step (4), we will be solving the orthogonal Procrustes problem given a fixed $\mathbf{U}^{(k)}$.
 198 An explicit solution can be constructed from the singular value decomposition, as detailed in the
 199 following Lemma.

Lemma 2. *Suppose $p \geq r$ and $\mathbf{Z} \in \mathbb{R}^{p \times r}$. Let $\mathbf{Z} = \mathbf{M}\mathbf{D}\mathbf{N}^\top$ be its (skinny) singular value decom-
 position, where $\mathbf{M} \in \mathbb{R}^{p \times r}$, $\mathbf{D} = \mathbb{R}^{r \times r}$ and $\mathbf{N} \in \mathbb{R}^{r \times r}$. An optimal solution to*

$$\underset{\mathbf{V}: \mathbf{V}^\top \mathbf{V} = \mathbf{I}}{\text{maximize}} \text{Tr}(\mathbf{Z}^\top \mathbf{V})$$

200 *is given by $\hat{\mathbf{V}} = \mathbf{M}\mathbf{N}^\top$, and the objective function has optimal value $\|\mathbf{Z}\|_*$, the nuclear norm of \mathbf{Z} .*

201 *Proof.* See in Appendix A.2. □

202 To analyze the computational complexity of the algorithm, we see a one-time computation of
 203 $\mathbf{Y}^\top \mathbf{X}$ that costs $O(npq)$. In each iteration, there is $O(pqr)$ complexity for the matrix multiplication
 204 $\mathbf{Y}^\top \mathbf{X}\mathbf{U}^{(k)}$ and $O(qr^2)$ for computing the SVD and the final solution. Therefore, the per-iteration
 205 computational complexity for the V-step is $O(pqr + qr^2)$, or $O(pqr)$ when $p \gg q$.

Algorithm 1 Alternating Minimization

- 1: Define a sequence of λ values $\lambda_1 > \dots > \lambda_L \geq 0$.
- 2: **for** $\ell = 1$ **to** L **do**
- 3: Let $k = 0$, and initialize $\mathbf{U}^{(0)}, \mathbf{V}^{(0)}$.
- 4: **while** $k = 0$ **or** $\|\mathbf{U}^{(k)}\mathbf{V}^{(k)\top} - \mathbf{U}^{(k-1)}\mathbf{V}^{(k-1)\top}\| > \epsilon$ **do**
- 5: **V-step:** Fix $\mathbf{U}^{(k)}$, solve \mathbf{V} : the orthogonal Procrustes problem

$$\underset{\mathbf{V}: \mathbf{V}^\top \mathbf{V} = \mathbf{I}}{\text{minimize}} \|\mathbf{Y} - \mathbf{X}\mathbf{U}^{(k)}\mathbf{V}^\top\|_F^2. \quad (4)$$

- Let $\mathbf{Y}^\top \mathbf{X}\mathbf{U}^{(k)} = \mathbf{M}\mathbf{D}\mathbf{N}^\top$ (skinny SVD) and solve $\mathbf{V}^{(k+1)} = \mathbf{M}\mathbf{N}^\top$.
- 6: **U-step:** Fix $\mathbf{V}^{(k+1)}$, solve \mathbf{U} : the group lasso problem

$$\underset{\mathbf{U}}{\text{minimize}} \frac{1}{2} \|\mathbf{Y}\mathbf{V}^{(k+1)} - \mathbf{X}\mathbf{U}\|_F^2 + \lambda_\ell \sum_{j=1}^p \|\mathbf{U}_j\|_2. \quad (5)$$

- 7: $k = k + 1$
 - 8: **end while**
 - 9: **end for**
-

206 In the U-step, we are solving a group lasso problem. Computing $\mathbf{Y}\mathbf{V}^{(k+1)}$ takes $O(nqr)$ time.
 207 The group-lasso problem can be solved by **glmnet** (Friedman et al., 2010) with the **mgaussian**
 208 family. With coordinate descent, its complexity is $O(\tilde{k}pqn)$, where \tilde{k} is the number of iterations
 209 until convergence and is expected to be small with a reasonable initialization, for example, provided
 210 by warm start. Thus, the per-iteration complexity for the U-step is $O(nqr + \tilde{k}npq)$, which is $O(\tilde{k}pqn)$
 211 when $p \gg r$.

212 Therefore, the overall computational complexity scales at least linearly with the number of
 213 features, and will have a large multiplier if the sample size is large as well. While subsampling
 214 can effectively reduce the computational cost, in high-dimensional settings, it is critical to have
 215 sufficient samples for the quality of estimation. Instead, we seek for computational techniques that
 216 can lower the actual number of features involved in expensive iterative computation without giving
 217 up any statistical efficiency. Thanks to the induced sparsity by the objective function, we are able
 218 to achieve it by variable screening.

219 3.2 Variable Screening for Ultrahigh-Dimensional Problems

220 In this section, we discuss strategic ways to find a good subset of variables to focus on in the
 221 computation that would allow us to reconstruct the full solution easily. In particular, we would like
 222 to iterate through the following steps for each λ :

- 223 1. **Screen** a strong set S and treat all the left-out variables S^c as null variables that potentially
 224 have zero coefficients;
- 225 2. **Solve** a significantly smaller problem on the subset of variables S ;
- 226 3. **Check** an optimality condition to guarantee the constructed full solution $\hat{\mathbf{B}} = (\hat{\mathbf{B}}_S, \hat{\mathbf{B}}_{S^c})$ with

227 $\hat{\mathbf{B}}_{S^c} = 0$ is indeed a valid solution to the original problem. If the condition is not satisfied,
 228 go back to the first step with an expanded set S .

229 3.2.1 Screening Strategies

230 We have seen Lemma 1 that determines the entry point of any nonzero coefficient on the solution
 231 path. Furthermore, there is evidence that the variables entering the model (as one decreases the λ
 232 value) tend to have large values by this criterion. Tibshirani et al. (2012) developed on this idea
 233 and proposed the strong rules as a sequential variable screening mechanism. The strong rules state
 234 that in a standard lasso problem with the model matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ and a single
 235 response $\mathbf{y} \in \mathbb{R}^n$, assume $\hat{\beta}(\lambda_{k-1})$ is the lasso solution at λ_{k-1} , then the j th predictor is discarded
 236 at λ_k if

$$|\mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_{k-1}))| < \lambda_k - (\lambda_{k-1} - \lambda_k). \quad (6)$$

237 The key idea is that the inner product above is almost “non-expansive” in terms of λ . As a result,
 238 the KKT condition suggests that the variables to be discarded by (6) would have coefficient 0 at
 239 λ_k . However it is not a guarantee. The strong rules can fail, though failures occur rarely when
 240 $p > n$. In any case, the KKT condition is checked to ensure the exact solution is found. Although
 241 Tibshirani et al. (2012) focused mostly on the lasso-type problem, they also suggested extension to
 242 general objective functions and penalties. For general objective function $f(\beta)$ with p_j -norm penalty
 243 $\|\beta_j\|_{p_j}$ for the j th group, the screening criterion will be based on the dual norm of its gradient
 244 $\|\nabla_j f(\beta)\|_{q_j}$ where $1/p_j + 1/q_j = 1$.

245 Inspired by the general strong rules, we propose three sequential screening strategies for the
 246 sparse reduced rank objective (3), named after their respective characteristics: Multi-Gaussian,
 247 Rank-Less and Fix-V. They are based either on the solution of a relaxed convex problem at the
 248 same λ_k or on the exact solution at the previous λ_{k-1} .

- 249 • (Multi-Gaussian) Solve the full-rank convex problem at λ_k and use its active set as the candi-
 250 dates for the low-rank settings. The main advantage is that the screening is always stable due
 251 to the convexity. However this approach often overselects and brings extra burden to the com-
 252 putation. By assuming a higher rank than necessary, the effective number of responses would
 253 become more than that of a low-rank model. As a result, more variables would potentially
 254 be needed to serve for an enlarged set of responses.
- 255 • (Rank-Less) Find variables that have large $c_j = \|\mathbf{X}_j^\top (\mathbf{Y} - \mathbf{X}\mathbf{U}(\lambda_{k-1})\mathbf{V}(\lambda_{k-1})^\top)\|_2$. This is
 256 analogous to the strong rules applied to the vanilla multi-response lasso ignoring the rank
 257 constraint.
- 258 • (Fix-V) Find variables that have large $c'_j = \|\mathbf{X}_j^\top (\mathbf{Y}\mathbf{V}(\lambda_{k-1}) - \mathbf{X}\mathbf{U}(\lambda_{k-1}))\|_2$. This is similar
 259 to the strong rules applied in the \mathbf{U} -step with \mathbf{V} assumed fixed. To see the rationale better,
 260 we take another perspective. The squared error in SRRR (3) can also be written as

$$\|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|_F^2 = \mathbf{Tr}(\mathbf{Y}^\top \mathbf{Y}) - 2\mathbf{Tr}(\mathbf{Y}^\top \mathbf{X}\mathbf{U}\mathbf{V}^\top) + \mathbf{Tr}(\mathbf{X}\mathbf{U}\mathbf{V}^\top \mathbf{V}\mathbf{U}^\top \mathbf{X}^\top)$$

Since $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$, the optimization problem becomes

$$\underset{\mathbf{U}, \mathbf{V}: \mathbf{V}^\top \mathbf{V} = \mathbf{I}}{\text{minimize}} \frac{1}{2} \|\mathbf{X}\mathbf{U}\|_F^2 - \mathbf{Tr}(\mathbf{Y}^\top \mathbf{X}\mathbf{U}\mathbf{V}^\top) + \lambda \sum_{j=1}^p \|\mathbf{U}_j\|_2$$

For any given \mathbf{U} , we can solve $\mathbf{V} = \mathbf{M}\mathbf{N}^\top$, where $\mathbf{Y}^\top \mathbf{X}\mathbf{B} = \mathbf{M}\mathbf{D}\mathbf{N}^\top$ is its singular value decomposition. Let $f(\mathbf{U}) = \frac{1}{2}\|\mathbf{X}\mathbf{U}\|_F^2 - \|\mathbf{Y}^\top \mathbf{X}\mathbf{U}\|_*$. The problem is reduced to

$$\underset{\mathbf{U}}{\text{minimize}} f(\mathbf{U}) + \lambda \sum_{j=1}^p \|\mathbf{U}_{j\cdot}\|_2$$

The general strong rule tells us to screen based on the gradient; that is

$$\nabla_{\mathbf{B}} f(\mathbf{B}) = \mathbf{X}^\top \mathbf{X}\mathbf{U} - \mathbf{X}^\top \mathbf{Y}\mathbf{M}\mathbf{N}^\top = \mathbf{X}^\top (\mathbf{X}\mathbf{U} - \mathbf{Y}\mathbf{V}).$$

261 Therefore, the general strong rules endorse the use of this screening rule.

262 We do some experiments to compare the effectiveness of the rules. We simulate the model matrix
 263 under an independent design and an equi-correlated design with correlation $\rho = 0.5$. The true
 264 solution path is computed using Algorithm 1 with several random initializations and the convex
 265 relaxation-based initialization (as in the Multi-Gaussian rule). Let $S(\lambda)$ be the true active set at
 266 λ . For each method ℓ above, we can find, based on either the exact solution at λ_{k-1} or the full-
 267 rank solution at λ_k , the threshold it needs so that by the screening criterion, the selected subset
 268 $\hat{S}(\lambda_k)^{(\ell)}$ contains the true subset at λ_k , i.e. $\hat{S}(\lambda_k|\lambda_{k-1})^{(\ell)} \supseteq S(\lambda_k)$. This demonstrates how deep
 269 each method has to search down the variable list to include all necessary variables, and thus how
 270 accurate the screening mechanism is — the larger the subset size, the worse the method is.

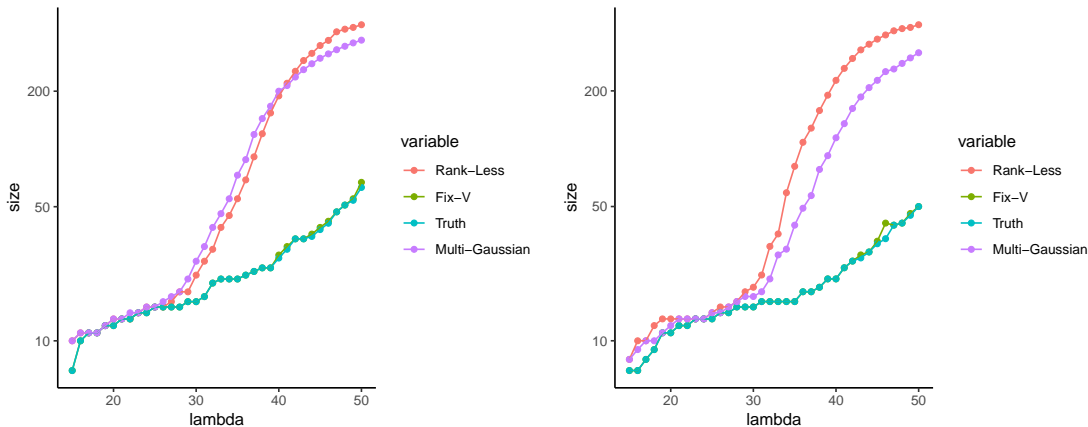


Figure 2: Size of screened set under different strategies. Left: independent design. Right: equi-correlated design with $\rho = 0.5$. Signal-to-noise ratio (SNR) = 1, and we use the true rank = 3.

271 We see from both plots that the curve of the Fix-V method is able to track that of the exact
 272 subset fairly well, while the Rank-Less and Multi-Gaussian methods both choose a much larger
 273 subset in order to cover the subset of active variables in the exact solution. In the rest of the paper,
 274 we will adopt the Fix-V method to do variable screening.

275 3.2.2 Optimality Condition

276 Although the Fix-V method turns out to be most effective in choosing the subset of variables, in
 277 practice we have no access to the true subset and have to take an estimate. Instead of trying to find

278 a sophisticated threshold, we will do batch screening at a fixed size (this size can change adaptively
 279 though). Given a size K , we will take the K variables that rank the top under this criterion. Clearly
 280 we can make mistakes by having left out some important variables in the screening stage. In order
 281 to make sure that our solution is exact rather than approximate in terms of the original problem,
 282 we need to check the optimality condition and take in more variables when necessary.

283 Suppose we find a solution $\hat{\mathbf{U}}_S, \hat{\mathbf{V}}_S$ on a subset of variables \mathbf{X}_S by alternating minimization. We
 284 will verify the assembled solution $\hat{\mathbf{U}} = (\hat{\mathbf{U}}_S, \mathbf{0}), \hat{\mathbf{V}} = \hat{\mathbf{V}}_S$ is a limit point of the original optimization
 285 problem. The argument is supported by the following lemma.

286 **Lemma 3.** *In the U-step (12), given \mathbf{V} and λ , if we have an exact solution $\hat{\mathbf{U}}_S$ for the sub-problem
 287 with \mathbf{X}_S , then $\hat{\mathbf{U}} = (\hat{\mathbf{U}}_S, \mathbf{0})$ is a solution to the full problem if and only if for all $j \in S^c$,*

$$\|\mathbf{x}_j^\top (\mathbf{YV} - \mathbf{X}_S \hat{\mathbf{U}}_S)\|_2 \leq \lambda. \quad (7)$$

288 *Proof.* Since this is a convex problem, $\hat{\mathbf{U}}$ is a solution if and only if $\mathbf{0} \in \partial f(\hat{\mathbf{U}})$ where f is the
 289 objective function in (12) and ∂f is its subdifferential. For the vector ℓ_2 -norm, we know that the
 290 subdifferential of $\|\mathbf{x}\|_2$ is $\{\mathbf{s} \in \mathbb{R}^p : \|\mathbf{s}\|_2 \leq 1\}$ if $\mathbf{x} = \mathbf{0}$ and $\{\mathbf{x}/\|\mathbf{x}\|_2\}$ if $\mathbf{x} \neq \mathbf{0}$. Notice that
 291 $\mathbf{X}_S \hat{\mathbf{U}}_S = \mathbf{X} \hat{\mathbf{U}}$ by the definition of $\hat{\mathbf{U}}$. Since we have an exact solution on S , we know $\mathbf{0} \in \partial f(\hat{\mathbf{U}})_j$
 292 for all $j \in S$. On the other hand, for $j \in S^c$, $\mathbf{0} \in \partial f(\hat{\mathbf{U}})$ if and only if $\mathbf{0} \in \{\mathbf{x}_j^\top (\mathbf{X} \hat{\mathbf{U}} - \mathbf{YV}) + \lambda \mathbf{s}_j : \|\mathbf{s}_j\|_2 \leq 1\}$, which is further equivalent to $\|\mathbf{x}_j^\top (\mathbf{YV} - \mathbf{X}_S \hat{\mathbf{U}}_S)\|_2 = \|\mathbf{x}_j^\top (\mathbf{YV} - \mathbf{X} \hat{\mathbf{U}})\|_2 \leq \lambda$ \square

294 Therefore, once we obtain a solution $\hat{\mathbf{U}}_S, \hat{\mathbf{V}}_S$ for the sub-problem and get condition (7) verified,
 295 we know in the V-step, by the lemma above, $\hat{\mathbf{U}} = (\hat{\mathbf{U}}_S, \mathbf{0})$ is the solution given $\hat{\mathbf{V}} = \hat{\mathbf{V}}_S$. In the
 296 U-step, since $\mathbf{X} \hat{\mathbf{U}} = \mathbf{X}_S \hat{\mathbf{U}}_S$, $\hat{\mathbf{U}}$ is the solution to the full problem. We see that $(\hat{\mathbf{U}}, \hat{\mathbf{V}})$ is a limiting
 297 point of the alternating minimization algorithm for the original problem. However if the condition
 298 fails, we expand the screened set or bring in the violated variables, and do the fit again. We should
 299 note that when we say an exact solution to the original problem, we do not claim it to be a local
 300 minimum or global minimum, unless under some regularity conditions as will be briefly discussed
 301 later. It is a limiting point of the vanilla alternating minimization algorithm, i.e. Algorithm 1.
 302 In other words, if we start from the constructed solution (with zero coefficients for the leftout
 303 variables), the algorithm should converge in one iteration and return the same solution.

304 We have seen the main ingredients of the iterative algorithm: screening, solving and checking.
 305 Next we discuss some useful practical considerations and extensions.

306 3.3 Computational Considerations

307 3.3.1 Initialization and Warm Start

308 Recall that in the training stage our goal is to fit an SRRR solution path across different λ values.
 309 It is easy to see that with a careful choice of parameterization, the path is continuous in λ . To
 310 leverage this property, we adopt a warm start strategy. Specifically, we initialize the coefficients of
 311 the existing variables at λ_{k+1} using the solution at λ_k and zero-initialize the newly added variables.
 312 With warm start, much less iterations will be needed to converge to the new minimum.

313 However, this by no means guarantees that we are all on a good path. It's likely that we
 314 are trapped into a neighborhood of local optimum and end up with much higher function value
 315 than the global minimum. One way to alleviate this, if affordable, is to solve the corresponding
 316 full-rank problem first, and initialize the coefficients with low-rank approximation of the full-rank

317 solution. We can compare the limiting function values with the warm-start initialization and see
318 which converges to a better point. Although we didn't use in the actual implementation and
319 experiments, one could also do random exploration — randomly initialize some of the coefficients,
320 run the algorithm multiple times and find one that achieves the lowest function value. That said,
321 we lose the advantage of warm start though. The good news is, in the experiments we have done,
322 we didn't observe very clear suboptimal behavior by the warm start and full-rank strategies.

323 3.3.2 Early Stopping

324 Although we pre-specify a sequence of λ values $\lambda_1 > \lambda_2 > \dots > \lambda_L$ where we want to fit the SRRR
325 models, we do not have to fit them all given our goal is to find the best predictive model. Once the
326 model starts to overfit as we move down the λ list, we can stop our process since the later models
327 will have no practical use and are expensive to train. Therefore, in the actual computation, we
328 monitor the validation error along the solution path and call it a stop if it shows a clear upward
329 trend. One other point we would like to make in this regard is that the validation metric can
330 be defined either as an average MSE over all phenotypes or a subset of phenotypes we are most
331 interested in. This is because practically the best λ value can be different for different phenotypes
332 in the joint model.

333 3.4 Extensions

334 3.4.1 Standardization

335 We often want to standardize the predictors if they are not on the same scale because the penalty
336 term is not invariant to change of units of the variables. However we emphasize that some thought
337 has to be put into this before standardizing the predictors. If the predictors are already on the
338 same scale, standardizing them could bring unintended advantages to variables with smaller variance
339 themselves. It is more reasonable not to standardize in such cases.

340 In terms of the outcomes, since they can be at different scales, it is important to standardize
341 them in the training stage so that no one dominates in the objective function. At prediction (both
342 training and test time), we scale back to the original levels using their respective variances from
343 the training set. In fact, the real impact an outcome has to the overall objective is determined by
344 the proportion of unexplained variance. It would be good to weight the responses properly based
345 on this if such information is available or can be estimated, e.g. via heritability estimation for
346 phenotypes in genetic studies.

347 3.4.2 Weighting

348 Sometimes we have strong reasons or evidence to prioritize some of the predictors than the oth-
349 ers. We can easily extend the standard objective (3) and reflect this belief in a weighted penalty
350 $\lambda \sum_{j=1}^p w_j \|\mathbf{U}_j\|_2$ where the weight w_j controls inversely the relative importance of the j th variable.
351 For example, $w_j = 0$ implies j th variable will always be included in the model, while a large w_j will
352 almost exclude the variable from the model.

353 In the response space, we can also impose a weighting mechanism to prioritize the training of
354 certain responses. For a given set of nonnegative weights $w_k, 1 \leq k \leq q$, the SRRR objective (3)
355 can be modified to $(1/2) \sum_{k=1}^q w_k \|\mathbf{Y}_{\cdot k} - \mathbf{X}\mathbf{U}\mathbf{V}_k^\top\|_2^2 + \lambda \sum_{j=1}^p \|\mathbf{U}_j\|_2$ with the same constraint, or

356 equivalently,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|\mathbf{Y}\mathbf{W}^{\frac{1}{2}} - \mathbf{X}\mathbf{U}\mathbf{V}^{\top}\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}_{j\cdot}\|_2, \\ \text{s.t.} \quad & \mathbf{V}^{\top}\mathbf{W}^{-1}\mathbf{V} = \mathbf{I}, \end{aligned} \tag{8}$$

357 where the weight matrix $\mathbf{W} = \mathbf{diag}(w_1, \dots, w_q)$. To solve the problem with our alternating min-
 358 imization scheme, we can see that in the V-step, instead of solving the standard orthogonal Pro-
 359 crustes problem with an elegant analytic solution derived from the SVD, we have to deal with a
 360 so-called weighted orthogonal Procrustes problem (WOPP). Finding the solution of the WOPP is
 361 far more complicated. See, for instance, Mooijaart, Commandeur (1990), Chu, Trendafilov (1998)
 362 and Viklands (2006). An iterative procedure is often needed to compute the solution. For better
 363 computational efficiency, we instead solve the problem with the original orthonormal constraint:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|\mathbf{Y}\mathbf{W}^{\frac{1}{2}} - \mathbf{X}\mathbf{U}\mathbf{V}^{\top}\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}_{j\cdot}\|_2, \\ \text{s.t.} \quad & \mathbf{V}^{\top}\mathbf{V} = \mathbf{I}. \end{aligned} \tag{9}$$

364 That is, we amplify the magnitude of some responses so that the objective value is more sensitive
 365 to the loss incurred on these responses. When making prediction, we will need to scale them back
 366 to the original units.

367 3.4.3 Adjustment Covariates

368 In some applications such as genome-wide association studies (GWAS), there may be confounding
 369 variables $\mathbf{Z} \in \mathbb{R}^{n \times m}$ that we want to adjust for in the model. For example, population stratification,
 370 defined as the existence of a systematic ancestry difference in the sample data, is one of the common
 371 factors in GWAS that can lead to spurious discoveries. This can be controlled for by including some
 372 leading principal components of the SNP matrix as variables in the regression (Price et al., 2006).
 373 In the presence of such variables, we solve the following problem instead. With a slight abuse of
 374 notation, in this section, we use \mathbf{W} to denote the coefficient matrix for the covariates instead of a
 375 weight matrix:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|\mathbf{Y} - \mathbf{Z}\mathbf{W} - \mathbf{X}\mathbf{U}\mathbf{V}^{\top}\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}^j\|_2, \\ \text{s.t.} \quad & \mathbf{V}^{\top}\mathbf{V} = \mathbf{I}. \end{aligned} \tag{10}$$

376 The main components don't change except two adjustments. When determining the starting λ
 377 value, we use Lemma 4.

Lemma 4. *In problem (10), if $r > 0$, the maximum λ that results in a nontrivial solution $\hat{B}(\lambda)$ is*

$$\lambda_{\max} = \max_{1 \leq j \leq p} \|\mathbf{x}_j^{\top} \hat{\mathbf{R}}\|_2,$$

378 where $\hat{\mathbf{R}} = \mathbf{Y} - \mathbf{Z}\hat{\mathbf{W}}$ and $\hat{\mathbf{W}}$ is the multiple outcome regression coefficient matrix.

The proof is almost the same as before. The other nuance we should be careful about is when fitting the model, we should leave those covariates unpenalized because they serve for the adjustment purpose and should not be experiencing the selection stage. In particular, in the U-step (group lasso) given \mathbf{V} , direct computation would reduce to solving the problem

$$\underset{\mathbf{U}, \mathbf{W}}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{YV} - \mathbf{Z}\mathbf{W}\mathbf{V} - \mathbf{XU}\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}^j\|_2,$$

which is not as convenient as standard group lasso problem. Instead, we find that \mathbf{W} can always be solved explicitly in terms of other variables. In fact, the minimizer $\hat{\mathbf{W}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top (\mathbf{Y} - \mathbf{XU}\mathbf{V}^\top)$. Plug in and we find that the problem to be solved can be written as

$$\underset{\mathbf{U}}{\text{minimize}} \quad \frac{1}{2} \|(\mathbf{I} - \mathbf{H}_Z)\mathbf{YV} - (\mathbf{I} - \mathbf{H}_Z)\mathbf{XU}\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}^j\|_2,$$

379 where $\mathbf{H}_Z = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$ is the projection matrix on the column space of \mathbf{Z} . This becomes a
 380 standard group lasso problem and can be solved by using, for example, the `glmnet` package with
 381 the `mgaussian` family.

382 3.4.4 Missing Values

In practice, there can be missing values in either the predictor matrix or the outcome matrix. If we only discard samples that have any missing value, we could lose a lot of information. For the predictor matrix, we could do imputation as simple as mean imputation or something sophisticated by leveraging the correlation structure. For missingness in the outcome, there is a natural way to integrate an imputation step seamlessly with the current procedure, analogous to the `softImpute` idea in Mazumder et al. (2010). We first define a projection operator for a subset of two dimensional indices $\Omega \subseteq \{1, \dots, n\} \times \{1, \dots, p\}$. Let $\mathcal{P}_\Omega : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{n \times p}$ be such that

$$\mathcal{P}_\Omega(\mathbf{Y})_{i,j} = \begin{cases} \mathbf{Y}_{i,j}, & (i,j) \in \Omega, \\ 0, & (i,j) \notin \Omega. \end{cases}$$

383 Let Ω be the set of indices where the response values are observed; in other words, Ω^c is the set of
 384 missing locations. Instead of (3), now we solve the following problem.

$$\begin{aligned} \underset{\mathbf{U}, \mathbf{V}, \mathbf{Y}'}{\text{minimize}} \quad & \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{Y}) - \mathcal{P}_\Omega(\mathbf{XU}\mathbf{V}^\top)\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}^j\|_2, \\ \text{s.t.} \quad & \mathbf{V}^\top \mathbf{V} = \mathbf{I}. \end{aligned} \tag{11}$$

385 We can easily see that an equivalent formulation of the problem is

$$\begin{aligned} \underset{\mathbf{U}, \mathbf{V}, \mathbf{Y}'}{\text{minimize}} \quad & \frac{1}{2} \|\mathbf{Y}' - \mathbf{XU}\mathbf{V}^\top\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}^j\|_2, \\ \text{s.t.} \quad & \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \mathcal{P}_\Omega(\mathbf{Y}') = \mathcal{P}_\Omega(\mathbf{Y}). \end{aligned}$$

386 This inspires a natural projection step to deal with the additional constraint. It can be well
387 integrated with the current alternating minimization scheme. In fact, after each alternation between
388 the U-step and the V-step, we can impute the missing values from the current predictions \mathbf{XUV}^\top ,
389 and then continue into the next U-V alternation with the completed matrix.

390 3.4.5 Lazy Reduced Rank Regression

391 There is an alternative way to find a low-rank coefficient profile for the multivariate regression.
392 Instead of pursuing to solve the non-convex problem (3) directly, we can follow a two-stage procedure:

1. Solve a full-rank multi-gaussian sparse regression, i.e.,

$$\text{minimize}_{\mathbf{B}} \quad \frac{1}{2} \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{B}_{j\cdot}\|_2.$$

- 393 2. Conduct SVD of the resulting coefficient matrix $\hat{\mathbf{B}}$ and use its rank r approximation as our
394 final estimator.

395 The advantage of this approach is that it is stable. The first stage is a convex problem and can be
396 handled efficiently by, for example, `glmnet`. A variety of adaptive screening rules are also applicable
397 in this situation to assist dimension reduction. The second stage is fairly standard and efficient
398 as long as there are not too many active variables. However, the disadvantage is clear too. The
399 low-rank approximation is conducted in an unsupervised manner, so could lead to some degrade in
400 the prediction performance.

401 That said, as before, we should still evaluate the out-of-sample performance as the penalty
402 parameter λ varies and pick the best on the solution path as our final estimated model. In many
403 cases, we compute the full-rank model under the exact mode anyways, so the set of lazy models
404 can be thought of as an efficient byproduct for our choice.

405 3.5 Full Algorithm

406 We incorporate the options above and present the full algorithm in Algorithm 2.

Algorithm 2 Large-scale and Ultrahigh-dimensional Sparse Reduced Rank Regression

- 1: Standardize or weight the responses. Define a sequence of λ values $\lambda_1 > \dots > \lambda_L$. Initialize $\mathbf{U}(\lambda_0) = \mathbf{0}$, $\mathbf{V}(\lambda_0) = \mathbf{0}$ and \mathbf{Y}_{Ω^c} .
- 2: **for** $\ell = 1$ **to** L **do**
- 3: Initialize $t = 0$, $\mathbf{U}(\lambda_\ell) = \mathbf{U}(\lambda_{\ell-1})$, $\mathbf{V}(\lambda_\ell) = \mathbf{V}(\lambda_{\ell-1})$, $\mathbf{W}(\lambda_\ell) = \mathbf{W}(\lambda_{\ell-1})$, and $\mathcal{A}(\lambda_\ell)$ be the active set at $\lambda_{\ell-1}$.
- 4: **while** $t = 0$ **or** KKT Check at $t - 1$ failed **do**
- 5: **[Variable Screening]** Find M variables $S_M \subseteq \Omega \setminus \mathcal{A}(\lambda_\ell)$ with largest values in $\|\mathbf{x}_j^\top (\mathbf{Y} - \mathbf{Z}\mathbf{W}(\lambda_\ell) - \mathbf{X}_{\mathcal{A}(\lambda_\ell)}\mathbf{U}_{\mathcal{A}(\lambda_\ell)}(\lambda_\ell)\mathbf{V}(\lambda_\ell)^\top)\|$, and let

$$\mathcal{A}(\lambda_\ell) = \mathcal{A}(\lambda_\ell) \cup S_M.$$

- 6: **[Alternating Minimization]** Let $k = 0$ and $\mathbf{U}^{(0)} = \mathbf{U}_{\mathcal{A}(\lambda_\ell)}(\lambda_\ell)$, $\mathbf{V}^{(0)} = \mathbf{V}(\lambda_\ell)$, $\mathbf{W}^{(0)} = \mathbf{W}(\lambda_\ell)$ and $\mathbf{Y}^{(0)} = \mathbf{Y}$.
- 7: **while** $k = 0$ **or** $\|\mathbf{U}^{(k)}\mathbf{V}^{(k)\top} - \mathbf{U}^{(k-1)}\mathbf{V}^{(k-1)\top}\| > \epsilon$ **do**
- 8: **V-step:** Fix $\mathbf{U}^{(k)}$, solve \mathbf{V} : the orthogonal Procrustes problem

$$\underset{\mathbf{V}: \mathbf{V}^\top \mathbf{V} = \mathbf{I}}{\text{minimize}} \|\mathbf{Y}^{(k)} - \mathbf{Z}\mathbf{W}^{(k)} - \mathbf{X}_{\mathcal{A}(\lambda_\ell)}\mathbf{U}^{(k)}\mathbf{V}^\top\|_F^2.$$

Let $(\mathbf{Y}^{(k)} - \mathbf{Z}\mathbf{W}^{(k)})^\top \mathbf{X}_{\mathcal{A}(\lambda_\ell)}\mathbf{U}^{(k)} = \mathbf{M}\mathbf{D}\mathbf{N}^\top$ (skinny SVD) and solve $\mathbf{V}^{(k+1)} = \mathbf{M}\mathbf{N}^\top$.

- 9: **U-step:** Fix $\mathbf{V}^{(k+1)}$, solve \mathbf{U} and \mathbf{W} : the group lasso problem

$$\mathbf{U}^{(k+1)} = \underset{\mathbf{U}}{\text{argmin}} \frac{1}{2} \|(\mathbf{I} - \mathbf{H}_Z)\mathbf{Y}^{(k)}\mathbf{V}^{(k+1)} - (\mathbf{I} - \mathbf{H}_Z)\mathbf{X}_{\mathcal{A}(\lambda_\ell)}\mathbf{U}\|_F^2 + \lambda_\ell \sum_{j=1}^p \|\mathbf{U}_j\|_2, \quad (12)$$

and $\mathbf{W}^{(k+1)} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top (\mathbf{Y}^{(k)} - \mathbf{X}_{\mathcal{A}(\lambda_\ell)}\mathbf{U}^{(k+1)}\mathbf{V}^{(k+1)})$.

- 10: **Y-step:** Impute the missing values

$$\mathbf{Y}_\Omega^{(k+1)} = \mathbf{Y}_\Omega^{(k)}, \quad \mathbf{Y}_{\Omega^c}^{(k+1)} = (\mathbf{Z}\mathbf{W}^{(k+1)} + \mathbf{X}_{\mathcal{A}(\lambda_\ell)}\mathbf{U}^{(k+1)}(\mathbf{V}^{(k+1)})^\top)_{\Omega^c}$$

- 11: $k = k + 1$
- 12: **end while**
- 13: Let $\mathbf{U}_{\mathcal{A}(\lambda_\ell)}(\lambda_\ell) = \mathbf{U}^{(k)}$, $\mathbf{U}_{\mathcal{A}(\lambda_\ell)}(\lambda_\ell) = \mathbf{0}$, $\mathbf{V}(\lambda_\ell) = \mathbf{V}^{(k)}$, $\mathbf{W}(\lambda_\ell) = \mathbf{W}^{(k)}$ and $\mathbf{Y} = \mathbf{Y}^{(k)}$.
- 14: **[KKT Check]** Check the criterion for all $j \in \Omega \setminus \mathcal{A}(\lambda_\ell)$,

$$\|\mathbf{x}_j^\top (\mathbf{Y} - \mathbf{Z}\mathbf{W}(\lambda_\ell) - \mathbf{X}_{\mathcal{A}(\lambda_\ell)}\mathbf{U}_{\mathcal{A}(\lambda_\ell)}(\lambda_\ell)\mathbf{V}(\lambda_\ell)^\top)\| \leq \lambda_\ell.$$

- 15: $t = t + 1$
 - 16: **end while**
 - 17: **end for**
-

407 4 Convergence Analysis

In this section, we present some convergence properties of the alternating minimization algorithm (Algorithm 1) on sparse reduced rank regression. Let

$$g(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}^j\|_2.$$

Theorem 1. *For any $k \geq 1$, the function values are monotonically decreasing:*

$$g(\mathbf{U}^k, \mathbf{V}^k) \geq g(\mathbf{U}^{k+1}, \mathbf{V}^k) \geq g(\mathbf{U}^{k+1}, \mathbf{V}^{k+1}).$$

Furthermore, we have the following finite convergence rate:

$$\min_{1 \leq k \leq K} g(\mathbf{U}^k, \mathbf{V}^k) - g(\mathbf{U}^{k+1}, \mathbf{V}^{k+1}) \leq \frac{1}{K} (g(\mathbf{U}^1, \mathbf{V}^1) - g^\infty),$$

408 where $g^\infty = \lim_{k \rightarrow \infty} g(\mathbf{U}^k, \mathbf{V}^k)$. It implies that the iteration will terminate in $O(1/\epsilon)$ iterations.

409 The proof is straightforward and we won't detail here. It presents the fact that alternating
410 minimization is a descent algorithm. In fact, this property holds for all alternating minimization
411 or more general blockwise coordinate descent algorithms. However it does not say how good the
412 limiting point is. In the next result, we show a local convergence result that under some regularity
413 conditions, if the initialization is closer enough to a global minimum, it will converge to a global
414 minimum at linear rate. It is based on similar results on proximal gradient descent by Dubois et al.
415 (2019). To define a local neighborhood, it would be easier if we eliminate \mathbf{V} by always setting it to
416 a minimizer given \mathbf{U} . That is, the objective function becomes $F_\lambda(\mathbf{U}) = \frac{1}{2} \|\mathbf{X}\mathbf{U}\|_2^2 - \|\mathbf{Y}^\top \mathbf{X}\mathbf{U}\|_* +$
417 $\lambda \sum_{j=1}^p \|\mathbf{U}_j\|_2$. We define a sublevel set $\mathcal{S}_c(\lambda) = \{\mathbf{U} \in \mathbb{R}^{p \times r} : F_\lambda(\mathbf{U}) \leq c\}$.

Theorem 2. *Assume $\mathbf{X}^\top \mathbf{X}$ is invertible and $\sigma_{\max}^2 \geq \sigma_{\min}^2 > 0$ be its smallest and largest eigenvalues. Let s_j be the j th singular value of $(\mathbf{X}^\top \mathbf{X})^{-\frac{1}{2}} \mathbf{X}^\top \mathbf{Y}$. There exists $\bar{\lambda} > 0$ such that for all $0 \leq \lambda < \bar{\lambda}$ and $0 \leq \mu < \sigma_{\min}^2 (1 - s_{r+1}^2 / s_r^2)$, there is a sublevel set $\mathcal{S}(\lambda, \mu)$ where the level depends on λ and μ such that if $\mathbf{U}^k \in \mathcal{S}(\lambda, \mu)$, we have*

$$\Delta(\mathbf{U}^{k+1}, \mathbf{V}^{k+1}) \leq \left(1 - \min \left(\frac{1}{2}, \frac{\mu}{\sigma_{\max}^2} \right) \right) \Delta(\mathbf{U}^k, \mathbf{V}^k),$$

418 where $\Delta(\mathbf{U}, \mathbf{V}) = g(\mathbf{U}, \mathbf{V}) - g(\mathbf{U}^*, \mathbf{V}^*)$ and $(\mathbf{U}^*, \mathbf{V}^*)$ is a global minimum.

419 From a high level, the proof is based on the fact that under the conditions, the function is strongly
420 convex near the global minima. If we starting from this region, we achieve good convergence rate
421 with alternating minimization algorithm. The full proof is given in Appendix A.3.

422 It is easy to see that the theorem above implicitly assumes the classical setting where $n \geq p$
423 since otherwise $\mathbf{X}^\top \mathbf{X}$ would not be invertible. However, it is still applicable to our algorithm. The
424 algorithm does not attempt to solve alternating minimization at the full scale, but only does it
425 after variable screening. With screening, it is very likely that we will again be working under the
426 classical setting. Moreover, with warm start, there is higher chance that the initialization lies in the
427 local region as defined above. Therefore, this theorem can provide useful guidance on the practical
428 computational performance of the algorithm.

429 5 Simulation Studies

430 We conduct some experiments to gain more insight into the method and compare with the single-
431 response lasso method. Due to space limit, we demonstrate the results in one experiment setting
432 and include results for other settings such as correlated features, deviation from the true low-rank
433 structure etc., in Appendix C. We experiment with three different sizes and three different signal-
434 to-noise ratio (SNR): $(n, p, k) = (200, 100, 20), (200, 500, 20), (200, 500, 50)$, where k is the number
435 of variables with true nonzero coefficients, and the target SNR = 0.5, 1, or 3. The number of
436 responses $q = 20$ and the true rank $r = 3$. We generate the $\mathbf{X} \in \mathbb{R}^{n \times p}$ with independent samples
437 from some multivariate Gaussian $\mathcal{N}(0, \Sigma_X)$ where $\Sigma_X = \mathbf{I}_p$ in this section. More results under
438 correlated designs are presented in the appendix. The response is generated from the true model
439 $\mathbf{Y} = \mathbf{X}\mathbf{U}\mathbf{V}^\top + \mathbf{E}$, where each entry in the support of $\mathbf{U} \in \mathbb{R}^{p \times r}$ (sparsity k) is independently drawn
440 from a standard Gaussian distribution, and $\mathbf{V} \in \mathbb{R}^{q \times r}$ takes the left singular matrix of a Gaussian
441 ensemble. Hence $\mathbf{B} = \mathbf{U}\mathbf{V}^\top$ is the true coefficient matrix. The noise matrix is generated from
442 $\mathcal{N}(0, \sigma_e^2 \mathbf{I}_q)$, where σ_e^2 is chosen such that the signal-to-noise ratio

$$\text{SNR} = \frac{\text{Tr}(\mathbf{B}^\top \Sigma_X \mathbf{B})}{\sigma_e^2 \cdot \text{Tr}(\Sigma_E)} \quad (13)$$

is set to a given level. The performance is evaluated by the test R^2 , defined as follows:

$$R^2 = 1 - \frac{\|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\|_F^2}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2}.$$

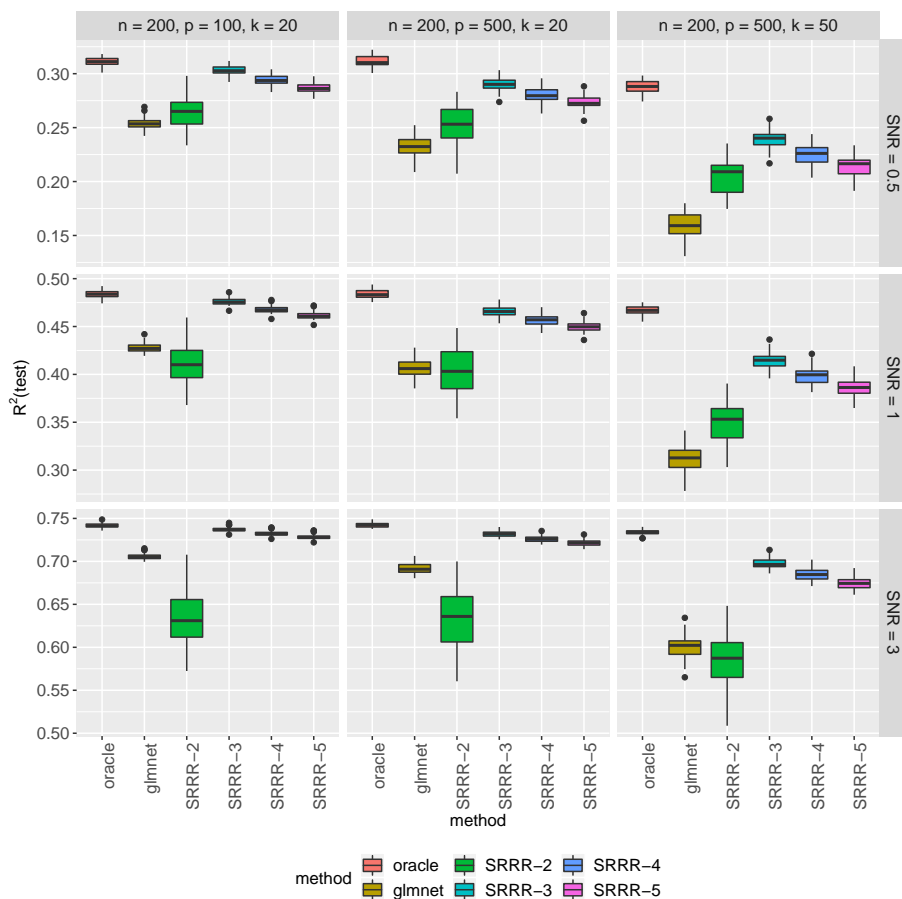


Figure 3: R^2 each run is evaluated on a test set of size 5000. “oracle” is the result where we know the true active variables and solve on this subset of variables. “glmnet” fits the responses separately. “SRRR- r ” indicates the SRRR results with assumed rank r .

443 The main insight we obtain from the experiments is that the method is more robust to over-
 444 estimating than underestimating the rank. A significant degrade in performance can be identified
 445 even if we are only off the rank by 1 from below. In contrast, the additional variance brought along
 446 by overestimating the rank doesn’t seem to be a big concern. This, in essence, can be ascribed to
 447 bias and variance decomposition. In our settings, the bias incurred in underestimating the rank
 448 and thus 1/3 loss of parameters contributes a lot more to the MSE compared with the increased
 449 variance due to 1/3 redundancy in the parameters.

450 6 Real Data Application: UK Biobank

451 The UK Biobank (Bycroft et al., 2018) is a large, prospective population-based cohort study with
 452 individuals collected from multiple sites across the United Kingdom. It contains extensive genetic
 453 and phenotypic detail such as genome-wide genotyping, questionnaires and physical measures for a

454 wide range of health-related outcomes for over 500,000 participants, who were aged 40-69 years when
455 recruited in 2006-2010. In this study, we are interested in the relationship between an individual's
456 genotype and his/her phenotypic outcomes. While genome-wide association studies (GWAS) focus
457 on identifying SNPs that may be marginally associated with the outcome using univariate tests,
458 we would like to leverage the additive effect of all SNPs to make good prediction. Recently there
459 is a line of work (Qian et al., 2019; Sinnott-Armstrong et al., 2019; Lello et al., 2018) that builds a
460 lasso solution on the large dataset and shows that the prediction is much improved over previous
461 methods. Furthermore, a number of phenotypes present nontrivial correlation structures and we
462 would like to further improve the prediction and stabilize the variable selection by building a joint
463 model for multiple outcomes.

464 We focused on 337,199 White British unrelated individuals out of the full set of over 500,000 from
465 the UK Biobank dataset (Bycroft et al., 2018) that satisfy the same set of population stratification
466 criteria as in DeBoever et al. (2018). Each individual has up to 805,426 measured variants, and
467 each variant is encoded by one of the four levels where 0 corresponds to homozygous major alleles, 1
468 to heterozygous alleles, 2 to homozygous minor alleles and NA to a missing genotype. In addition,
469 we have available covariates such as age, sex, and forty pre-computed principal components of the
470 SNP matrix. Among them, we use age, sex and the top 10 PCs for the adjustment of population
471 stratification (Price et al., 2006).

472 There are binary responses in the data such as many disease outcomes. Although in principle
473 we can solve for a mixture of Gaussian and binomial likelihood using Newton's method, for ease of
474 computation in this large-scale setting, it is a reasonable approximation to treat them as continuous
475 responses and fit the standard SRRR model. However, after the model is fit, we will refit a logistic
476 regression on the predicted score to obtain a probability estimation. Notice that the refit is still
477 trained on the training set at each λ value.

478 The number of samples is large in the UK Biobank dataset, so we afford to set aside an inde-
479 pendent validation set without resorting to costly cross-validation to find an optimal regularization
480 parameter. We also leave out a subset of observations as test set to evaluate the final model.
481 In particular, we randomly partition the original dataset so that 70% is used for training, 10%
482 for validation and 20% for test. The solution path is fit on the training set, whereas the desired
483 regularization is selected on the validation set, and the final model is evaluated on the test set.

In the experiment, we compare the performance of the multivariate-response SRRR model with
the single-response lasso model. To fit the lasso model, we rely on fast implementation of the
snpnet package (Qian et al., 2019), and we also refer to the lasso results as **snpnet** in the results
section. For continuous responses, we evaluate the prediction by R-squared (R^2). Given a linear
coefficient vector $\hat{\beta}$ (fitted on the training set) and a subset of data $\{(x_i, y_i), 1 \leq i \leq n\}$, it is defined
as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - x_i^\top \hat{\beta})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

484 We compute R^2 respectively on the training, validation and test sets. For binary response, mis-
485 classification error could be used but it would depend on the calibration. Instead the receiver
486 operating characteristic (ROC) curve provides more information and demonstrates the tradeoff be-
487 tween true positive and false positive rates under different thresholds. The area under the curve
488 (AUC) computes the area under the ROC curve — a larger value indicates a generally better classi-
489 fier. Therefore, we will evaluate AUCs on the training, validation and test sets for binary responses.
490 When comparing different methods, we evaluate both absolute change and relative change over the

491 baseline method (in particular the already competitive lasso in our case), where the relative change
492 for a given metric is defined as $(\text{metric}_{\text{new}} - \text{metric}_{\text{lasso}}) / |\text{metric}_{\text{lasso}}|$.

493 Computationally, in the UK Biobank experiments, the SNP data are stored in a compressed
494 PLINK format with two-bit encodings. PLINK 2.0 (Chang et al., 2015) provides an extensive set
495 of efficient operations including very fast, multithreaded matrix multiplication. In particular, this
496 matrix multiplication module is heavily used in the steps of screening and KKT check in this work
497 and other lasso-based results (Li et al., 2020; Qian et al., 2019) on the UK Biobank.

498 6.1 Asthma and 7 Blood Biomarkers

499 Here, we defined asthma based on a mixture of self-reported questionnaire data and hospital in-
500 patient record data described in DeBoever et al. (2018); Tanigawa et al. (2019). Furthermore, we
501 focused on 7 additional blood count measurements from Category 100081 in UK Biobank containing
502 results of haematological assays that were performed on whole blood.

503 We apply the SRRR to the set of phenotypes and expect some performance improvement by
504 leveraging the correlation structure. Choice of the phenotypes: monocyte count, neutrophil count,
505 eosinophil count, basophil count, forced vital capacity (FVC), peak expiratory flow (PEF), and
506 forced expiratory volume in 1 second (FEV1).

507 Overall, we see small rank representation can maintain predictive power for specific phenotypes
508 (see Figure 4) and that overall the multiresponse model improves the prediction over the single-
509 response lasso model (see Figure 5).

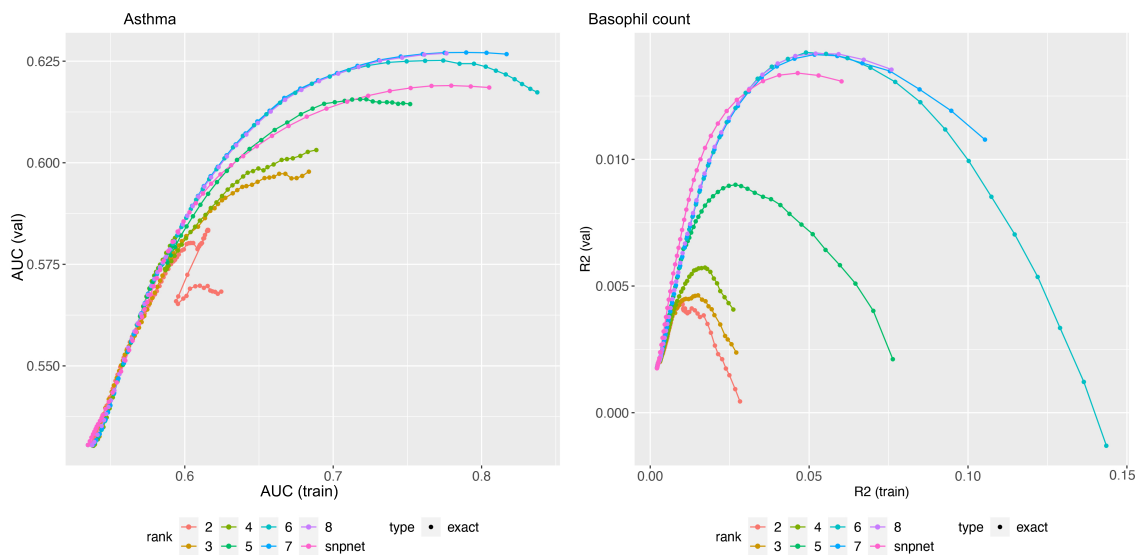


Figure 4: Asthma and Basophil count prediction performance plots. Different colors correspond to lower rank predictive performance across (x-axis) training data set and (y-axis) validation data set for (left) asthma and (right) basophil count.

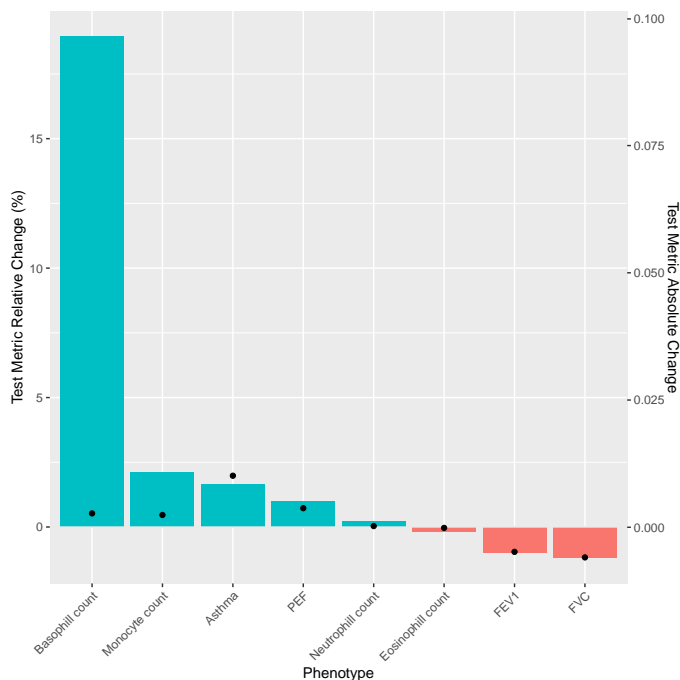


Figure 5: Change in prediction accuracy for multiresponse model compared to single response model. (top) (y-axis 1 bar) R^2 relative change (%) for each phenotype (x-axis) and R^2 absolute change (y-axis 2).

510 6.2 35 Biomarkers

511 In addition, we used 35 biomarkers from the UK Biobank biomarker panel in Sinnott-Armstrong
512 et al. (2019), and apply SRRR to the dataset. Noticeably, for the liver biomarkers including alanine
513 aminotransferase and albumin, and the urinary biomarkers including Microalbumin in urine and
514 Sodium in urine, we see an improvement in prediction performance for the SRRR application beyond
515 the single-response snpnet models (see Figures 6 and 7).

516 We can represent the lower rank representation as a biplot of the singular value decomposition
517 of the coefficient matrix (Gower et al., 2011; Gabriel, 1971; Tanigawa et al., 2019). Specifically, we
518 display phenotypes projected on phenotype principal components as a scatter plot. We also show
519 variants projected on variant principal components as a separate scatter plot and added phenotype
520 singular vectors as arrows on the plot using sub-axes. In scatter plot with biplot annotation, the
521 inner product of a genetic variant and a phenotype represents the direction and the strength of the
522 projection of the genetic association of the variant-phenotype pair on the displayed latent compo-
523 nents. For example, when a variant and a phenotype share the same direction on the annotated
524 scatter plot, that means the projection of the genetic associations of the variant-phenotype pair on
525 the displayed latent components is positive. When a variant-phenotype pair is projected on the
526 same line, but on the opposite direction, the projection of the genetic associations on the shown
527 latent components is negative. When the variant and phenotype vectors are orthogonal or one of
528 the vectors are of zero length, the projection of the genetic associations of the variant-phenotype
529 pair on the displayed latent components is zero. We focused on the top five key SRRR components

530 for AST to ALT ratio (see Figure 8).

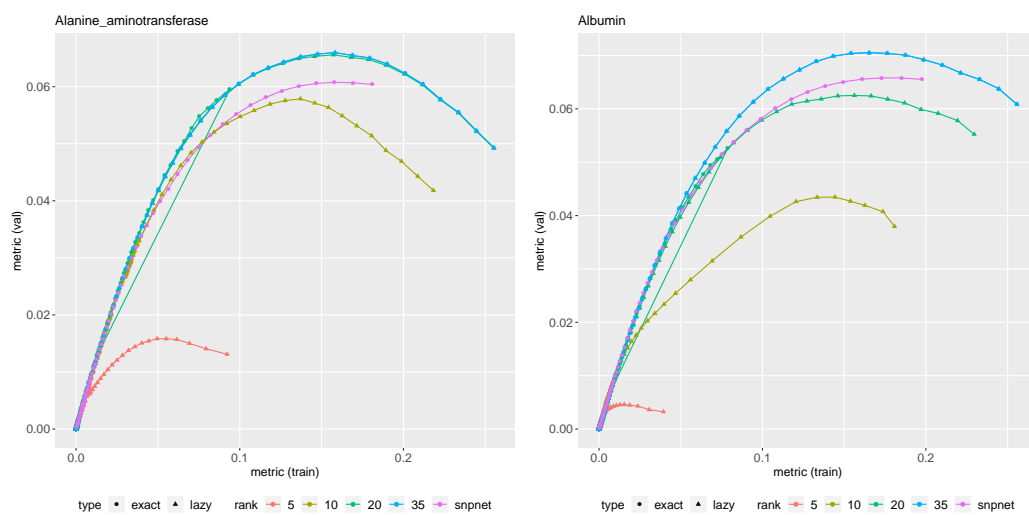


Figure 6: Alanine aminotransferase and albumin prediction performance plots. Different colors correspond to lower rank predictive performance across (x-axis) training data set and (y-axis) validation data set for (left) alanine aminotransferase and (right) albumin. For lower rank representation we applied lazy rank evaluation.

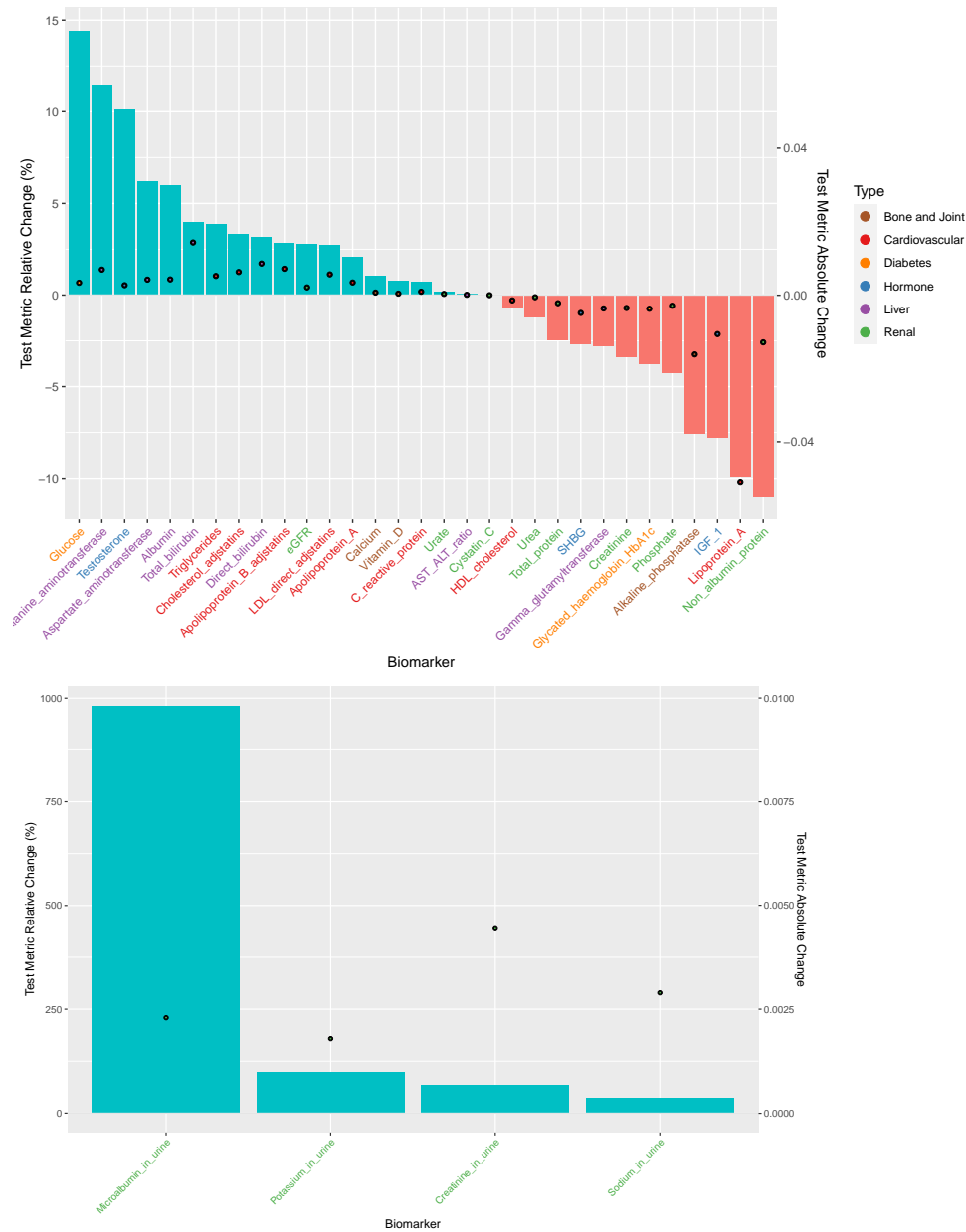


Figure 7: Change in prediction accuracy for multiresponse model compared to single response model. (top) (y-axis 1 bar) R^2 relative change (%) for each biomarker (x-axis) across different biomarker category (color) and R^2 absolute change (y-axis 2). (bottom) Change in predictive accuracy for multiresponse model compared to single response model for urinary biomarkers.

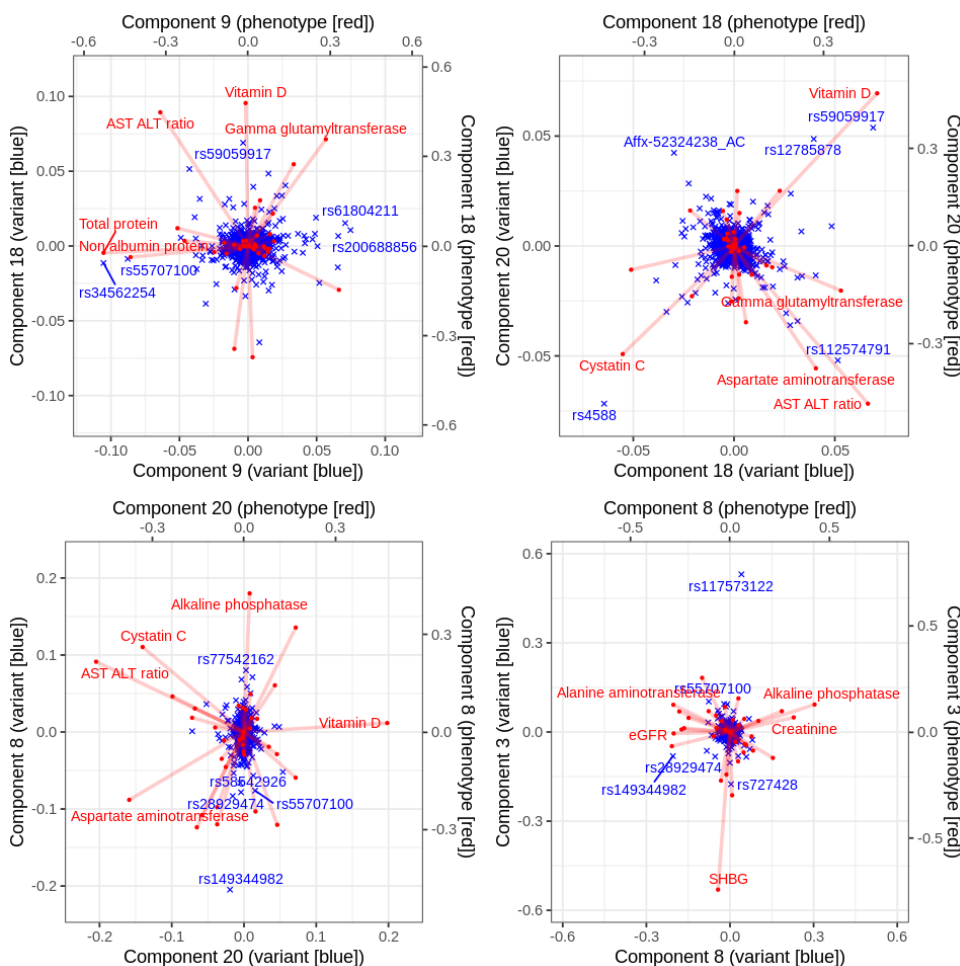


Figure 8: The latent structures of the the top five key SRRR components for AST to ALT ratio. Using trait squared cosine score described in Tanigawa et al. (2019), the top five key SRRR components for AST to ALT ratio (components 9, 18, 20, 8, and 3) are identified from a full-rank SVD of coefficient matrix C from SRRR ($C = UDV^T$) and shown as a series of biplots. In each panel, principal components of genetic variants (rows of UD) are shown in blue as scatter plot using the main axis and singular vectors of traits (rows of V) are shown in red dots with lines using the secondary axis, for the identified key components. The five traits and variants with the largest distance from the center of origin are annotated with their name.

531 7 Related Work

532 There are many other methods that were proposed for multivariate regression in high-dimensional
 533 settings. Chen, Huang (2012) compares the SRRR with rank-free methods including L_2 SVS Similä,
 534 Tikka (2007), L_∞ SVS (Turlach et al., 2005) that replaces the l_2 -norm with l_∞ -norm of each row,
 535 and RemMap (Peng et al., 2010) that imposes an additional elementwise sparsity of the coefficient

536 matrix. It also compares with the SPLS Chun, Keleş (2010) and points out that the latter does not
537 target directly on prediction of the responses so the performance turns out not as good. Another
538 important category of methods Canonical Correlation Analysis (CCA) (Hotelling, 1936) that tries to
539 constructed uncorrelated components in both the feature space and the response space to maximize
540 their correlation coefficients also falls short in the aspect, even though some connection can be
541 established with the reduced rank regression as seen in Appendix B.

542 More recently, there is a line of new advances in sparse and low-rank regression problems. For
543 example, Ma, Sun (2014) proposed a subspace assisted regression with row sparsity and studied
544 its near-optimal estimation properties. Ma et al. (2020) furthered this work to a two-way sparsity
545 setting, where nonzero entries are present only on a few rows and columns. Li et al. (2019) proposed
546 an integrative multi-view reduced-rank regression that encourages group-wise low-rank coefficient
547 matrices with a composite nuclear norm. Dubois et al. (2019) developed a fast first-order proximal
548 gradient algorithm on the SRRR objective reparameterized by a single matrix and proves linear
549 local convergence. Luo et al. (2018) proposed a mixed-outcome reduced-rank regression method
550 that deals with different types of responses and also missing data, though it does not aim for
551 high-dimensional settings with variable selection.

552 In genetics, some approaches proposed to decompose genetic associations from summary level
553 data using LD-pruning along with p-value thresholding for variable selection in an approach referred
554 to as DeGAs (Tanigawa et al., 2019) and MetaPhat (Lin et al., 2019). DeGAs was extended for
555 genetic risk prediction and to "paint" an individual's risk to a disease based on genetic component
556 loadings in an approach referred to as DeGAs-risk (Aguirre et al., 2019).

557 8 Summary and Discussion

558 In this paper, we propose a method that takes into account both sparsity in high-dimensional regres-
559 sion problems and low-rank structure when multiple correlated outcomes are present. A screening-
560 based alternating minimization algorithm is designed to deal with large-scale and ultrahigh-dimensional
561 applications, such as the UK Biobank population cohort. We demonstrate the effectiveness of the
562 method on both synthetic and real datasets focusing on asthma and 7 related blood count biomark-
563 ers, in addition to the 35 biomarker panel made available by UK Biobank (Sinnott-Armstrong et al.,
564 2019). We anticipate that the approach presented here will generalize to thousands of phenotypes
565 that are currently being measure in UK Biobank, e.g. metabolomics and imaging data that are
566 currently being generated in over 100,000 individuals.

567 Methodologically, in the UK Biobank experiments, we use continuous approximation to binary
568 outcomes. This is a reasonable assumption but ideally one would like to solve the exact problem
569 based on their respective likelihood. In principle, there is no theoretical challenge in the algorithmic
570 design. We can use Newton's method and enclose the procedure with an outer loop that conducts
571 quadratic approximation of the objective function. However, the quadratic problem involving both
572 penalty and low-rank constraint can be very messy. We might need some heuristics to find a more
573 convenient approximation. We see this as future work along with extending the SRRR algorithm
574 to other families including time-to-event multiple responses that can be used for survival analysis.
575 Furthermore, for an individual we can project a variant and phenotype loading across the reduced
576 rank to their risk to arrive at a similar analysis of outlier individuals with unusual painting of
577 genetic risk and to quantify the overall contribution of a component which may aid in disease risk
578 interpretation. Overall, we see the method and algorithms presented here as an important toolkit
579 to the prediction problem in human genetics.

580 Acknowledgement

581 This research has been conducted using the UK Biobank Resource under Application Number
582 24983, “Generating effective therapeutic hypotheses from genomic and hospital linkage data” (<http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf>). Based
583 on the information provided in Protocol 44532 the Stanford IRB has determined that the re-
584 search does not involve human subjects as defined in 45 CFR 46.102(f) or 21 CFR 50.3(g). All
585 participants of UK Biobank provided written informed consent (more information is available
586 at <https://www.ukbiobank.ac.uk/2018/02/gdpr/>). We thank all the participants in the UK
587 Biobank. M.A.R. is supported by Stanford University and a National Institutes of Health (NIH)
588 Center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01
589 HG009080). Y.T. is supported by a Funai Overseas Scholarship from the Funai Foundation for
590 Information Technology and the Stanford University School of Medicine. Research reported in this
591 publication was supported by the National Human Genome Research Institute of the NIH under
592 Award Number R01HG010140 (M.A.R.). The content is solely the responsibility of the authors
593 and does not necessarily represent the official views of the NIH. R.T. was partially supported by
594 NIH grant 5R01 EB001988-16 and NSF grant 19 DMS1208164. T.H. was partially supported by
595 grant DMS-1407548 from the National Science Foundation, and grant 5R01 EB 001988-21 from the
596 National Institutes of Health.
597

598 References

- 599 *Abadi Martín, Barham Paul, Chen Jianmin, Chen Zhifeng, Davis Andy, Dean Jeffrey, Devin*
600 *Mathieu, Ghemawat Sanjay, Irving Geoffrey, Isard Michael, Kudlur Manjunath, Levenberg Josh,*
601 *Monga Rajat, Moore Sherry, Murray Derek G., Steiner Benoit, Tucker Paul, Vasudevan Vijay,*
602 *Warden Pete, Wicke Martin, Yu Yuan, Zheng Xiaoqiang.* **TensorFlow: A System for Large-scale**
603 **Machine Learning** // Proceedings of the 12th USENIX Conference on Operating Systems Design
604 and Implementation. Berkeley, CA, USA: USENIX Association, 2016. 265–283. (OSDI’16).
- 605 *Aguirre Matthew, Tanigawa Yosuke, Venkataraman Guhan, Tibshirani Rob, Hastie Trevor, Rivas*
606 *Manuel A.* Polygenic risk modeling with latent trait-related genetic components // BioRxiv.
607 2019. 808675.
- 608 *Anderson T. W.* Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal
609 Distributions // *Ann. Math. Statist.* 09 1951. 22, 3. 327–351.
- 610 *Bach Francis R.* Consistency of the group lasso and multiple kernel learning // *Journal of Machine*
611 *Learning Research.* 2008. 9, Jun. 1179–1225.
- 612 *Bickel Peter J., Ritov Ya’acov, Tsybakov Alexandre B.* Simultaneous analysis of Lasso and Dantzig
613 selector // *Ann. Statist.* 08 2009. 37, 4. 1705–1732.
- 614 *Bottou Léon.* Large-scale machine learning with stochastic gradient descent // Proceedings of
615 COMPSTAT’2010. 2010. 177–186.
- 616 *Bovet Daniel P, Cesati Marco.* Understanding the Linux Kernel: from I/O ports to process man-
617 agement. 2005.

- 618 *Boyd Stephen, Boyd Stephen P, Vandenberghe Lieven.* Convex optimization. 2004.
- 619 *Bühlmann Peter, Van De Geer Sara.* Statistics for high-dimensional data: methods, theory and
620 applications. 2011.
- 621 *Bycroft Clare, Freeman Colin, Petkova Desislava, Band Gavin, Elliott Lloyd T., Sharp Kevin, Mo-*
622 *tyer Allan, Vukcevic Damjan, Delaneau Olivier, O'Connell Jared, Cortes Adrian, Welsh Saman-*
623 *tha, Young Alan, Effingham Mark, McVean Gil, Leslie Stephen, Allen Naomi, Donnelly Peter,*
624 *Marchini Jonathan.* The UK Biobank Resource with Deep Phenotyping and Genomic Data //
625 Nature. 2018. 562, 7726. 203–209.
- 626 *Chang Christopher C, Chow Carson C, Tellier Laurent CAM, Vattikuti Shashaank, Purcell*
627 *Shaun M, Lee James J.* Second-generation PLINK: rising to the challenge of larger and richer
628 datasets // GigaScience. 02 2015. 4, 1.
- 629 *Chen Lisha, Huang Jianhua Z.* Sparse reduced-rank regression for simultaneous dimension reduction
630 and variable selection // Journal of the American Statistical Association. 2012. 107, 500. 1533–
631 1545.
- 632 *Chu Moody T, Trendafilov Nickolay T.* On a differential equation approach to the weighted orthog-
633 onal Procrustes problem // Statistics and Computing. 1998. 8, 2. 125–133.
- 634 *Chun Hyonho, Keleş Sündüz.* Sparse partial least squares regression for simultaneous dimension
635 reduction and variable selection // Journal of the Royal Statistical Society: Series B (Statistical
636 Methodology). 2010. 72, 1. 3–25.
- 637 *DeBoever Christopher, Tanigawa Yosuke, Lindholm Malene E., McInnes Greg, Lavertu Adam,*
638 *Ingelsson Erik, Chang Chris, Ashley Euan A., Bustamante Carlos D., Daly Mark J., Rivas*
639 *Manuel A.* Medical Relevance of Protein-Truncating Variants across 337,205 Individuals in the
640 UK Biobank Study // Nature Communications. 2018. 9, 1. 1612.
- 641 *Dean Jeffrey, Ghemawat Sanjay.* MapReduce: Simplified Data Processing on Large Clusters //
642 Commun. ACM. I 2008. 51, 1. 107–113.
- 643 *Dubois Benjamin, Delmas Jean-François, Obozinski Guillaume.* Fast Algorithms for Sparse
644 Reduced-Rank Regression // Proceedings of Machine Learning Research. 89. 16–18 Apr 2019.
645 2415–2424. (Proceedings of Machine Learning Research).
- 646 *Duchi John C, Agarwal Alekh, Wainwright Martin J.* Dual averaging for distributed optimization:
647 Convergence analysis and network scaling // IEEE Transactions on Automatic control. 2011. 57,
648 3. 592–606.
- 649 *Efron Bradley, Hastie Trevor.* Computer Age Statistical Inference: Algorithms, Evidence, and Data
650 Science. 5. 2016.
- 651 *Friedman Jerome, Hastie Trevor, Tibshirani Rob.* Regularization Paths for Generalized Linear
652 Models via Coordinate Descent. 2010. 1–22.
- 653 *Gabriel Karl Ruben.* The biplot graphic display of matrices with application to principal component
654 analysis // Biometrika. 1971. 58, 3. 453–467.

- 655 *Gower John C, Lubbe Sugnet Gardner, Le Roux Niel J.* Understanding biplots. 2011.
- 656 *Greenshtein Eitan, Ritov Ya'Acov.* Persistence in high-dimensional linear predictor selection and
657 the virtue of overparametrization // *Bernoulli*. 12 2004. 10, 6. 971–988.
- 658 *Hastie Trevor, Tibshirani Robert, Friedman Jerome.* The Elements of Statistical Learning: Data
659 Mining, Inference, and Prediction, 2nd Edition. 2009. (Springer series in statistics).
- 660 *Hotelling Harold.* Relations Between Two Sets of Variates // *Biometrika*. 1936. 28, 3/4. 321–377.
- 661 *Lello Louis, Avery Steven G, Tellier Laurent, Vazquez Ana I, Campos Gustavo de los, Hsu
662 Stephen DH.* Accurate genomic prediction of human height // *Genetics*. 2018. 210, 2. 477–
663 497.
- 664 *Li Gen, Liu Xiaokang, Chen Kun.* Integrative multi-view regression: Bridging group-sparse and
665 low-rank models // *Biometrics*. 2019. 75, 2. 593–602.
- 666 *Li Ruilin, Chang Christopher, Justesen Johanne Marie, Tanigawa Yosuke, Qian Junyang, Hastie
667 Trevor, Rivas Manuel A, Tibshirani Robert j.* Fast Lasso method for Large-scale and Ultrahigh-
668 dimensional Cox Model with applications to UK Biobank // *BioRxiv*. 2020.
- 669 *Lin Jake, Tabassum Rubina, Ripatti Samuli, Pirinen Matti.* MetaPhat: Detecting and decomposing
670 multivariate associations from univariate genome-wide association statistics // *bioRxiv*. 2019.
671 661421.
- 672 *Luo Chongliang, Liang Jian, Li Gen, Wang Fei, Zhang Changshui, Dey Dipak K, Chen Kun.*
673 Leveraging mixed and incomplete outcomes via reduced-rank modeling // *Journal of Multivariate
674 Analysis*. 2018. 167. 378–394.
- 675 *Ma Zhuang, Ma Zongming, Sun Tingni.* Adaptive Estimation in Two-way Sparse Reduced-rank
676 Regression // *Statistica Sinica*. 01 2020.
- 677 *Ma Zongming, Sun Tingni.* Adaptive sparse reduced-rank regression // *arXiv preprint
678 arXiv:1403.1922*. 2014.
- 679 *Mazumder Rahul, Hastie Trevor, Tibshirani Robert.* Spectral regularization algorithms for learning
680 large incomplete matrices // *Journal of Machine Learning Research*. 2010. 11, Aug. 2287–2322.
- 681 *Meinshausen Nicolai, Bühlmann Peter.* High-dimensional graphs and variable selection with the
682 Lasso // *Ann. Statist.* 06 2006. 34, 3. 1436–1462.
- 683 *Mooijaart Ab, Commandeur Jacques JF.* A general solution of the weighted orthonormal Procrustes
684 problem // *Psychometrika*. 1990. 55, 4. 657–663.
- 685 *Obozinski Guillaume, Wainwright Martin J, Jordan Michael I, others .* Support union recovery in
686 high-dimensional multivariate regression // *The Annals of Statistics*. 2011. 39, 1. 1–47.
- 687 *Peng Jie, Zhu Ji, Bergamaschi Anna, Han Wonshik, Noh Dong-Young, Pollack Jonathan R, Wang
688 Pei.* Regularized multivariate regression for identifying master predictors with application to
689 integrative genomics study of breast cancer // *The Annals of Applied Statistics*. 2010. 4, 1. 53.

- 690 *Price Alkes L., Patterson Nick J., Plenge Robert M., Weinblatt Michael E., Shadick Nancy A., Reich*
691 *David.* Principal Components Analysis Corrects for Stratification in Genome-Wide Association
692 *Studies // Nature Genetics.* 2006. 38. 904.
- 693 *Qian Junyang, Du Wenfei, Tanigawa Yosuke, Aguirre Matthew, Tibshirani Robert, Rivas Manuel A,*
694 *Hastie Trevor.* A Fast and Flexible Algorithm for Solving the Lasso in Large-scale and Ultrahigh-
695 *dimensional Problems // BioRxiv.* 2019. 630079.
- 696 *Similä Timo, Tikka Jarkko.* Input selection and shrinkage in multiresponse linear regression //
697 *Computational Statistics & Data Analysis.* 2007. 52, 1. 406–422.
- 698 *Sinnott-Armstrong Nasa, Tanigawa Yosuke, Amar David, Mars Nina J, Aguirre Matthew,*
699 *Venkataraman Guhan Ram, Wainberg Michael, Ollila Hanna M, Pirruccello James P, Qian*
700 *Junyang, others .* Genetics of 38 blood and urine biomarkers in the UK Biobank // BioRxiv.
701 2019. 660506.
- 702 *Tanigawa Yosuke, Li Jiehan, Justesen Johanne M, Horn Heiko, Aguirre Matthew, DeBoever*
703 *Christopher, Chang Chris, Narasimhan Balasubramanian, Lage Kasper, Hastie Trevor, others*
704 *.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight
705 *adipocyte biology // Nature communications.* 2019. 10, 1. 1–14.
- 706 *Tibshirani Robert.* Regression Shrinkage and Selection via the Lasso // Journal of the Royal
707 *Statistical Society. Series B (Methodological).* 1996. 58, 1. 267–288.
- 708 *Tibshirani Robert, Bien Jacob, Friedman Jerome, Hastie Trevor, Simon Noah, Taylor Jonathan,*
709 *Tibshirani Ryan J.* Strong Rules for Discarding Predictors in Lasso-Type Problems // Journal
710 *of the Royal Statistical Society. Series B (Statistical Methodology).* 2012. 74, 2. 245–266.
- 711 *Turlach Berwin A, Venables William N, Wright Stephen J.* Simultaneous variable selection //
712 *Technometrics.* 2005. 47, 3. 349–363.
- 713 *Velu Raja, Reinsel Gregory C.* Multivariate reduced-rank regression: theory and applications. 136.
714 2013.
- 715 Algorithms for the Weighted Orthogonal Procrustes Problem and other Least Squares Problems.
716 // . 2006.
- 717 *Visscher Peter M., Wray Naomi R., Zhang Qian, Sklar Pamela, McCarthy Mark I., Brown*
718 *Matthew A., Yang Jian.* 10 Years of GWAS Discovery: Biology, Function, and Translation
719 // The American Journal of Human Genetics. 2017. 101, 1. 5–22.
- 720 *Wainwright Martin J.* Sharp thresholds for High-Dimensional and noisy sparsity recovery using
721 ℓ_1 -Constrained Quadratic Programming (Lasso) // IEEE transactions on information theory.
722 2009. 55, 5. 2183–2202.
- 723 *Xiao Lin.* Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization
724 // Journal of Machine Learning Research. 2010. 11, Oct. 2543–2596.
- 725 *Yuan Ming, Lin Yi.* Model selection and estimation in regression with grouped variables // Journal
726 *of the Royal Statistical Society: Series B (Statistical Methodology).* 2006. 68, 1. 49–67.

727 *Zaharia Matei, Chowdhury Mosharaf, Franklin Michael J., Shenker Scott, Stoica Ion. Spark: Cluster Computing with Working Sets // Proceedings of the 2Nd USENIX Conference on Hot Topics*
728 *in Cloud Computing. Berkeley, CA, USA: USENIX Association, 2010. 10–10. (HotCloud’10).*
729

730 *Zhao Peng, Yu Bin. On Model Selection Consistency of Lasso // J. Mach. Learn. Res. XII 2006.*
731 *7. 2541?2563.*

732 *Zou Hui, Hastie Trevor. Regularization and variable selection via the elastic net // Journal of the*
733 *Royal Statistical Society: Series B (Statistical Methodology). 2005. 67, 2. 301–320.*

734 A Additional Proofs

735 A.1 Proof of Lemma 1

736 This is intuitively the same as one without the rank constraint because when the coefficients just
 737 start to become nonzero, the coefficient matrix is low-rank in its nature. Therefore, for the purpose
 738 of finding the maximum meaningful λ , we can ignore the rank constraint unless $r = 0$. Without the
 739 constraint, it follows from the KKT condition that having all coefficients to be zero is equivalent
 740 to setting

$$\lambda \geq \lambda_{\max} = \max_{1 \leq j \leq p} \|\mathbf{x}_j^\top \mathbf{Y}\|_2. \quad (14)$$

741 Therefore, the maximum λ that accommodates a nontrivial solution is $\lambda_{\max} = \max_{1 \leq j \leq p} \|\mathbf{x}_j^\top \mathbf{Y}\|_2$.

742 A.2 Proof of Lemma 2

743 We plug in the SVD of \mathbf{Z} and have $\text{Tr}(\mathbf{Z}^\top \mathbf{V}) = \text{Tr}(\mathbf{N} \mathbf{D} \mathbf{M}^\top \mathbf{V}) = \text{Tr}(\mathbf{D} \mathbf{M}^\top \mathbf{V} \mathbf{N}) = \sum_{k=1}^r \mathbf{D}_{kk} \mathbf{S}_{kk}$,
 744 where $\mathbf{S} = \mathbf{M}^\top \mathbf{V} \mathbf{N}$ and the last equality is due to the fact that \mathbf{D} is a diagonal matrix. Notice that
 745 by the skinny SVD, $\mathbf{S} \mathbf{S}^\top = \mathbf{M}^\top \mathbf{V}^\top \mathbf{N} \mathbf{N}^\top \mathbf{V} \mathbf{M} = \mathbf{I}$. We thus know \mathbf{S} is an orthogonal matrix and the
 746 magnitude of its diagonal elements cannot exceed 1. Since \mathbf{D}_{kk} are all non-negative. To maximize
 747 $\sum_{k=1}^r \mathbf{D}_{kk} \mathbf{S}_{kk}$, we let $\mathbf{S}_{kk} = 1$ for all $1 \leq k \leq r$. This is equivalent to setting $\mathbf{S} = \mathbf{M}^\top \mathbf{V} \mathbf{N} = \mathbf{I}$.
 748 Therefore, one solution is given by $\mathbf{V} = \mathbf{M} \mathbf{N}^\top$. The maximum value of the objective is thus
 749 $\sum_{k=1}^r \mathbf{D}_{kk} = \|\mathbf{Z}\|_*$, the nuclear norm of \mathbf{Z} .

750 A.3 Proof of Theorem 2

We notice that in Problem (3) we can solve explicitly for \mathbf{V} and plug back into the objective function. It yields the objective function (after dropping the constant term $(1/2)\|\mathbf{Y}\|_F^2$):

$$F_\lambda(\mathbf{U}) = \frac{1}{2} \|\mathbf{X}\mathbf{U}\|_2^2 - \|\mathbf{Y}^\top \mathbf{X}\mathbf{U}\|_* + \lambda \sum_{j=1}^p \|\mathbf{U}_j\|_2,$$

We let $f_\lambda(\mathbf{U}) = (1/2)\|\mathbf{X}\mathbf{U}\|_2^2 - \|\mathbf{Y}^\top \mathbf{X}\mathbf{U}\|_*$ without the penalty term so that $F_\lambda(\mathbf{U}) = f_\lambda(\mathbf{U}) + \lambda \sum_{j=1}^p \|\mathbf{U}_j\|_2$. Define a local smooth approximation of F_λ as

$$\tilde{F}_\lambda^t(\mathbf{U}'; \mathbf{U}) = f_\lambda(\mathbf{U}) + \langle \nabla f_\lambda(\mathbf{U}), \mathbf{U}' - \mathbf{U} \rangle + (1/2t) \|\mathbf{U}' - \mathbf{U}\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{U}_j\|_2,$$

751 and $\mathbf{U}^+ = \text{argmin}_{\mathbf{U}'} [\tilde{F}_\lambda^t(\mathbf{U}'; \mathbf{U}) - F_\lambda(\mathbf{U})]$. Dubois et al. (2019) showed that if t is small enough
 752 such that $\tilde{F}_\lambda^t(\mathbf{U}^+; \mathbf{U}) \geq F_\lambda(\mathbf{U}^+)$, we have

$$F_\lambda(\mathbf{U}^+) - F_\lambda^* \leq \left(1 - \min\left(\frac{1}{2}, \mu t\right)\right) (F_\lambda(\mathbf{U}) - F_\lambda^*). \quad (15)$$

753 Consider the iterates $(\mathbf{U}^k, \mathbf{V}^k)_{k \geq 1}$ in the alternating minimization algorithm. Notice that $\nabla f_\lambda(\mathbf{U}^k) =$
 754 $\mathbf{X}^\top \mathbf{X} \mathbf{U}^k - \mathbf{X}^\top \mathbf{Y} \mathbf{V}^k$. We have

$$\begin{aligned}
 F_\lambda(\mathbf{U}^{k+1}) &= g(\mathbf{U}^{k+1}, \mathbf{V}^{k+1}) - \frac{1}{2} \|\mathbf{Y}\|_F^2 \quad (g \text{ is the SRRR objective function}) \\
 &\leq g(\mathbf{U}^{k+1}, \mathbf{V}^k) - \frac{1}{2} \|\mathbf{Y}\|_F^2 \\
 &= \min_{\mathbf{U}} g(\mathbf{U}, \mathbf{V}^k) - \frac{1}{2} \|\mathbf{Y}\|_F^2 \\
 &= \min_{\mathbf{U}} \left(\frac{1}{2} \|\mathbf{Y} - \mathbf{X} \mathbf{U}^k (\mathbf{V}^k)^\top\|_F^2 + \langle \mathbf{X}^\top (\mathbf{X} \mathbf{U}^k - \mathbf{Y} \mathbf{V}^k), \mathbf{U} - \mathbf{U}^k \rangle + \right. \\
 &\quad \left. \frac{1}{2} \text{Tr}((\mathbf{U} - \mathbf{U}^k)^\top \mathbf{X}^\top \mathbf{X} (\mathbf{U} - \mathbf{U}^k)) \right) + \lambda \sum_{j=1}^p \|\mathbf{U}_{j \cdot}\|_2 - \frac{1}{2} \|\mathbf{Y}\|_F^2 \\
 &\leq \min_{\mathbf{U}} \left(f_\lambda(\mathbf{U}^k) + \langle \nabla f_\lambda(\mathbf{U}), \mathbf{U}' - \mathbf{U} \rangle + \frac{1}{2} \sigma_{\max}^2 \|\mathbf{U} - \mathbf{U}^k\|_F^2 \right) + \lambda \sum_{j=1}^p \|\mathbf{U}_{j \cdot}\|_2 \\
 &= \min_{\mathbf{U}} \tilde{F}_\lambda^{1/\sigma_{\max}^2}(\mathbf{U}; \mathbf{U}^k),
 \end{aligned}$$

755 where the fourth line is the quadratic expansion of $g(\mathbf{U}, \mathbf{V}^k)$ at \mathbf{U}^k , the second to last is by the fact
 756 that $\text{Tr}((\mathbf{U} - \mathbf{U}^k)^\top \mathbf{X}^\top \mathbf{X} (\mathbf{U} - \mathbf{U}^k)) \leq \sigma_{\max}^2 \|\mathbf{U} - \mathbf{U}^k\|_F^2$, and the last equality is by the definition
 757 of \tilde{F}_λ^t function. Therefore, if we let $\mathbf{U}^{k,+} = \text{argmin}_{\mathbf{U}} [\tilde{F}_\lambda^{1/\sigma_{\max}^2}(\mathbf{U}; \mathbf{U}^k) - F_\lambda(\mathbf{U}^k)]$, we have

$$F_\lambda(\mathbf{U}^{k+1}) - F_\lambda^* \leq F_\lambda(\mathbf{U}^{k,+}) - F_\lambda^*. \quad (16)$$

758 We need to show that $\mathbf{U}^{k,+}$ satisfies the condition $\tilde{F}_\lambda^{1/\sigma_{\max}^2}(\mathbf{U}^{k,+}; \mathbf{U}^k) \geq F_\lambda(\mathbf{U}^{k,+})$. To see this,
 759 notice that in fact for any \mathbf{U} ,

$$\begin{aligned}
 \frac{1}{2} \|\mathbf{X} \mathbf{U}\|_2^2 &= \frac{1}{2} \|\mathbf{X} \mathbf{U}^k\|_F^2 + \langle \mathbf{X}^\top \mathbf{X} \mathbf{U}^k, \mathbf{U} - \mathbf{U}^k \rangle + \frac{1}{2} \|\mathbf{X} (\mathbf{U} - \mathbf{U}^k)\|_F^2 \\
 &\leq \frac{1}{2} \|\mathbf{X} \mathbf{U}^k\|_F^2 + \langle \mathbf{X}^\top \mathbf{X} \mathbf{U}^k, \mathbf{U} - \mathbf{U}^k \rangle + \frac{1}{2} \sigma_{\max}^2 \|\mathbf{U} - \mathbf{U}^k\|_F^2.
 \end{aligned}$$

Since $\mathbf{X}^\top \mathbf{Y} \mathbf{V}^k$ is a subgradient of $\|\mathbf{Y}^\top \mathbf{X} \mathbf{U}\|_*$ at \mathbf{U}^k , we have

$$-\|\mathbf{Y}^\top \mathbf{X} \mathbf{U}\|_* \leq -\|\mathbf{Y}^\top \mathbf{X} \mathbf{U}^k\|_* - \langle \mathbf{X}^\top \mathbf{Y} \mathbf{V}^k, \mathbf{U} - \mathbf{U}^k \rangle.$$

Adding the two inequalities up, and we have $F_\lambda(\mathbf{U}) \leq \tilde{F}_\lambda^{1/\sigma_{\max}^2}(\mathbf{U}; \mathbf{U}^k)$ for all \mathbf{U} . In particular, it
 holds for $\mathbf{U}^{k,+}$. Therefore, by (15) and (16), we have

$$F_\lambda(\mathbf{U}^{k+1}) - F_\lambda^* \leq F_\lambda(\mathbf{U}^{k,+}) - F_\lambda^* \leq \left(1 - \min \left(\frac{1}{2}, \frac{\mu}{\sigma_{\max}^2} \right) \right) (F_\lambda(\mathbf{U}^k) - F_\lambda^*),$$

760 and the convergence is linear.

761 B Connection with CCA

762 Canonical Correlation Analysis (CCA) has an internal connection with Reduced-Rank Regression
 763 (RRR). In particular, it can be shown that the low-rank components constructed on the \mathbf{X} space

764 turn out to be the same by a relaxed CCA and a generalized RRR. CCA finds linear combinations
 765 $\mathbf{XU} \in \mathbb{R}^{n \times r}$ of variables in $\mathbf{X} \in \mathbb{R}^{n \times p}$ and linear combinations $\mathbf{YV} \in \mathbb{R}^{n \times r}$ of variables in $\mathbf{Y} \in \mathbb{R}^{n \times q}$
 766 that attain the maximum correlation. We assume both \mathbf{X} and \mathbf{Y} have been centered. CCA solves
 767 the following optimization problem:

$$\begin{aligned} & \underset{\mathbf{U}, \mathbf{V}}{\text{maximize}} && \text{Tr}(\mathbf{U}^\top \mathbf{X}^\top \mathbf{Y} \mathbf{V}), \\ & \text{s.t.} && \mathbf{U}^\top \mathbf{X}^\top \mathbf{X} \mathbf{U} = \mathbf{V}^\top \mathbf{Y}^\top \mathbf{Y} \mathbf{V} = \mathbf{I}_r. \end{aligned} \quad (17)$$

768 In particular, in the one dimensional case, this reduces to the problem of maximizing our familiar
 769 correlation coefficient. An equivalent representation to (17) can be written as

$$\begin{aligned} & \underset{\mathbf{U}, \mathbf{V}}{\text{minimize}} && \|\mathbf{YV} - \mathbf{XU}\|_F^2, \\ & \text{s.t.} && \mathbf{U}^\top \mathbf{X}^\top \mathbf{X} \mathbf{U} = \mathbf{V}^\top \mathbf{Y}^\top \mathbf{Y} \mathbf{V} = \mathbf{I}_r. \end{aligned} \quad (18)$$

770 The solution to the problem is $\hat{\mathbf{U}} = \mathbf{S}_{xx}^{-1/2} \mathbf{Q}^{(r)}$, $\hat{\mathbf{V}} = \mathbf{S}_{yy}^{-1/2} \mathbf{P}^{(r)}$ where $\mathbf{P}^{(r)}$ and $\mathbf{Q}^{(r)}$ are the r
 771 leading left and right singular vectors of matrix $\mathbf{R} = \mathbf{S}_{yy}^{-1/2} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1/2}$. $\mathbf{P}^{(r)}$ is also the r leading
 772 eigenvectors of $\mathbf{S}_{yy}^{-1/2} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1/2}$. A relaxed form of CCA problem ignoring the \mathbf{U} -constraint
 773 solves

$$\begin{aligned} & \underset{\mathbf{U}, \mathbf{V}}{\text{minimize}} && \|\mathbf{YV} - \mathbf{XU}\|_F^2, \\ & \text{s.t.} && \mathbf{V}^\top \mathbf{Y}^\top \mathbf{Y} \mathbf{V} = \mathbf{I}_r. \end{aligned} \quad (19)$$

774 The solution is $\hat{\mathbf{U}} = \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1/2} \mathbf{P}^{(r)}$, $\hat{\mathbf{V}} = \mathbf{S}_{yy}^{-1/2} \mathbf{P}^{(r)}$, where $\mathbf{P}^{(r)}$ is the r leading eigenvectors
 775 of $\mathbf{S}_{yy}^{-1/2} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1/2}$. Therefore, the solution for \mathbf{V} remains unchanged, though \mathbf{U} is different
 776 due to the constraint.

777 On the other hand, in the (generalized) reduced rank regression, given a given positive-definite
 778 matrix Γ , the problem becomes

$$\underset{\mathbf{U}, \mathbf{V}}{\text{minimize}} \quad \text{Tr}(\Gamma^{1/2} (\mathbf{Y} - \mathbf{XUV}^\top)^\top (\mathbf{Y} - \mathbf{XUV}^\top) \Gamma^{1/2}). \quad (20)$$

779 This can be derived, for example, as an maximum likelihood estimator under the Gaussian assump-
 780 tion with known covariance Γ^{-1} . One solution (Velu, Reinsel, 2013) is given by

$$\begin{aligned} \hat{\mathbf{U}} &= \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \Gamma^{1/2} \mathbf{P}^{(r)}, \\ \hat{\mathbf{V}} &= \Gamma^{-1/2} \mathbf{P}^{(r)}, \end{aligned}$$

781 where $\mathbf{P}^{(r)}$ is the leading eigenvectors of $\mathbf{R} = \Gamma^{1/2} \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \Gamma^{1/2}$. We see that the solution when
 782 $\Gamma = \mathbf{S}_{yy}^{-1}$ is closely related to the relaxed CCA solution. \mathbf{U} is the same while \mathbf{V} is the so-called
 783 reflexive inverse of \mathbf{V} there.

784 C Additional Experiments

785 We conduct some experiments to gain more insight into the method and compare with other meth-
 786 ods. We generate the $\mathbf{X} \in \mathbb{R}^{n \times p}$ with independent samples from some multivariate Gaussian

787 $\mathcal{N}(0, \Sigma_X)$. For the first several cases, we generate the response from the true, most favorable model
 788 $\mathbf{Y} = \mathbf{X}\mathbf{U}\mathbf{V}^\top + \mathbf{E}$, where each entry in the support of $\mathbf{U} \in \mathbb{R}^{p \times r}$ (sparsity k) is independently drawn
 789 from a standard Gaussian distribution, and $\mathbf{V} \in \mathbb{R}^{q \times r}$ takes the left singular matrix of a Gaussian
 790 ensemble. Hence $\mathbf{B} = \mathbf{U}\mathbf{V}^\top$ is the true coefficient matrix. The noise matrix is generated from
 791 $\mathcal{N}(0, \sigma_e^2 \Sigma_E)$, where σ_e^2 is chosen such that the signal-to-noise ratio

$$\text{SNR} = \frac{\text{Tr}(\mathbf{B}^\top \Sigma_X \mathbf{B})}{\sigma_e^2 \cdot \text{Tr}(\Sigma_E)} \quad (21)$$

is set to a given level. The performance is evaluated by the test R^2 , defined as follows:

$$R^2 = 1 - \frac{\|\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\|_F^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|_F^2}.$$

792 We consider several sets of experiments.

- 793 1. **Scenario 1-9** Small experiments: $(n, p, k) = (200, 100, 20), (200, 500, 20), (200, 500, 50)$, $q =$
 794 $20, r = 3$. The X has independent design, and the noise across different responses are all
 795 independent, i.e. $\Sigma_X = \mathbf{I}_p, \Sigma_E = \mathbf{I}_q$. Target SNR = 0.5, 1, 3. The results are evaluated on
 796 test sets of size 5000.
- 797 2. **Scenario 10-18** Same as Scenario 1-9. The true coefficient matrix is no longer exact low
 798 rank. It is perturbed by Gaussian noise with mean 0 and standard deviation 0.5.
3. **Scenario 19-27** Same as Scenario 1-9, except that the predictors are correlated. In particular,

$$\text{Cov}(\mathbf{x}_j, \mathbf{x}_{j'}) = \begin{cases} 1, & j = j', \\ \rho, & j \neq j'. \end{cases}$$

799 We let $\rho = 0.5$ in this set of simulation.

- 800 4. **Scenario 28-36** Same as Scenario 10-18, except that the predictors are correlated as in
 801 Scenario 19-27.

802 From the simulations, we find that underestimating the rank can degrade the performance instantly.
 803 Overestimating the rank will give one a variance penalty, but it seems to be rather robust compared
 804 with the other direction.

805 **Scenario 1-9** Small experiments: $(n, p, k) = (200, 100, 20), (200, 500, 20), (200, 500, 50), q = 20, r =$
 806 **3.** The \mathbf{X} has independent design, and the noise across different responses are all independent, i.e.
 807 $\Sigma_X = \mathbf{I}_p, \Sigma_E = \mathbf{I}_q$. Target SNR = 0.5, 1, 3. The results are evaluated on test sets of size 5000.

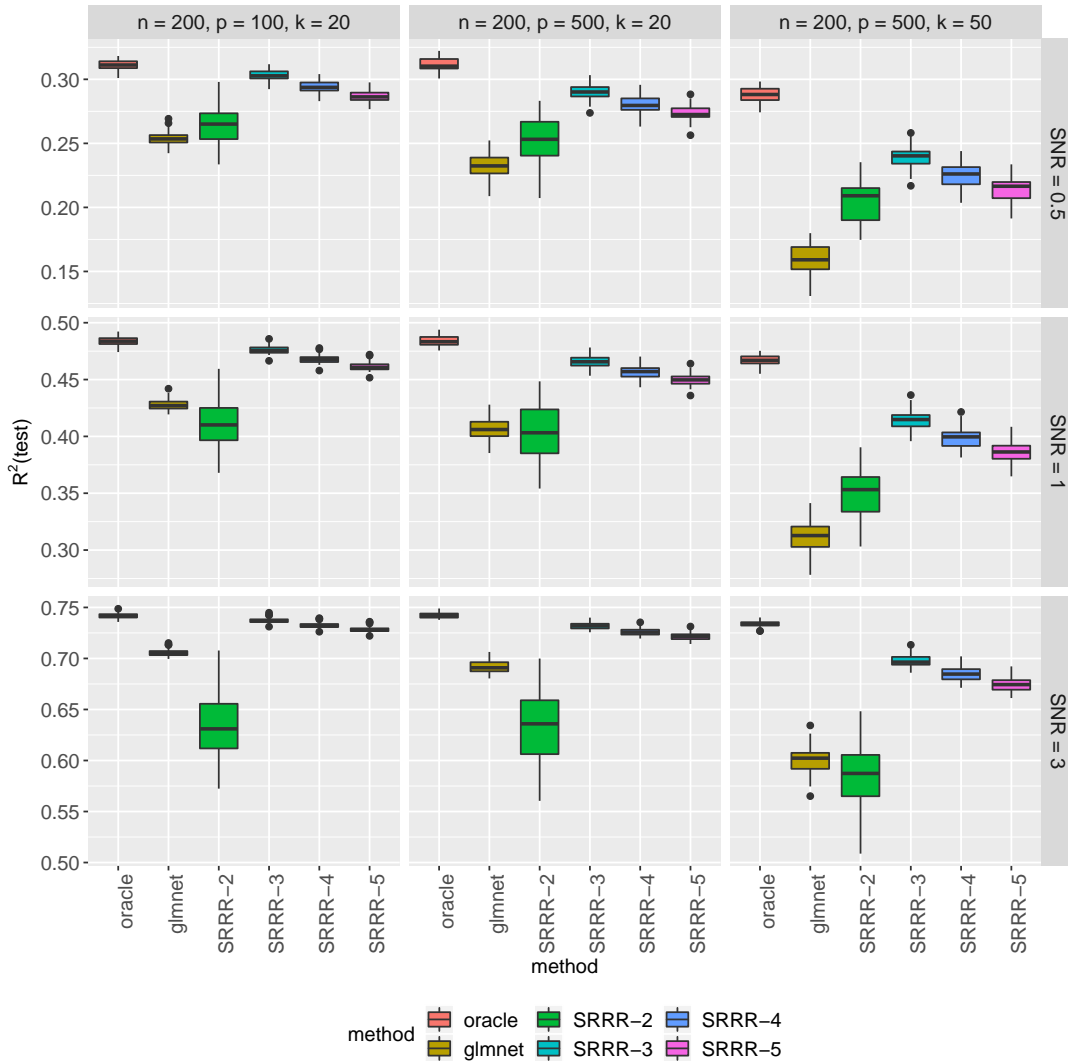


Figure C.1: Scenario 1-9. R^2 each run is evaluated on a test set of size 5000.

808 **Scenario 10-18** Same as Scenario 1-9. The true coefficient matrix is no longer exact low rank.
 809 It is perturbed by Gaussian noise with mean 0 and standard deviation 0.5.

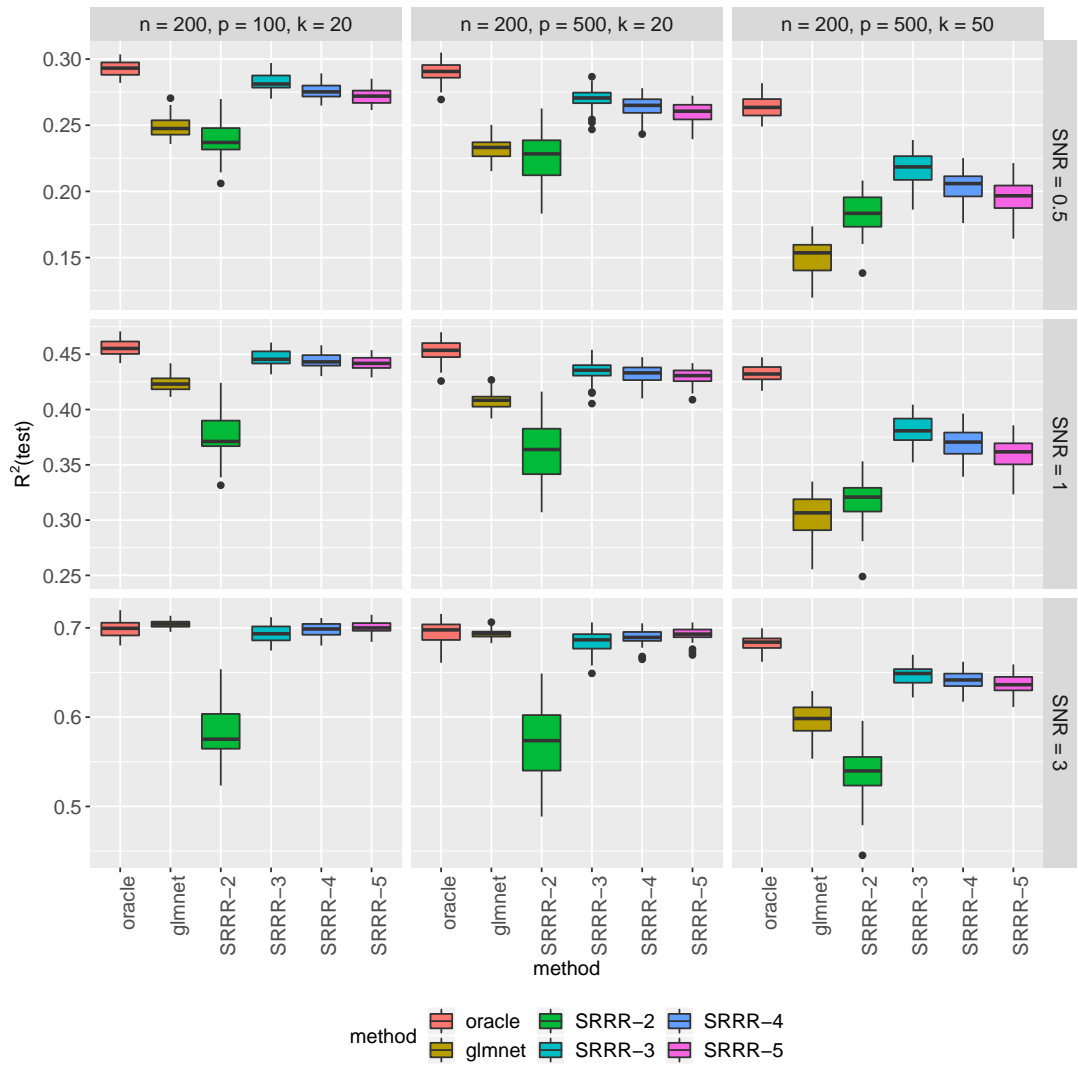


Figure C.2: Scenario 10-18. R^2 each run is evaluated on a test set of size 5000. The oracle here does not take into account the noise in true coefficient matrix, and do reduced rank regression on the true support and the true rank.

Scenario 19-27 Same as Scenario 1-9, except that the predictors are correlated. In particular,

$$\text{Cov}(\mathbf{x}_j, \mathbf{x}_{j'}) = \begin{cases} 1, & j = j', \\ \rho, & j \neq j'. \end{cases}$$

810 We let $\rho = 0.5$ in this set of simulation.

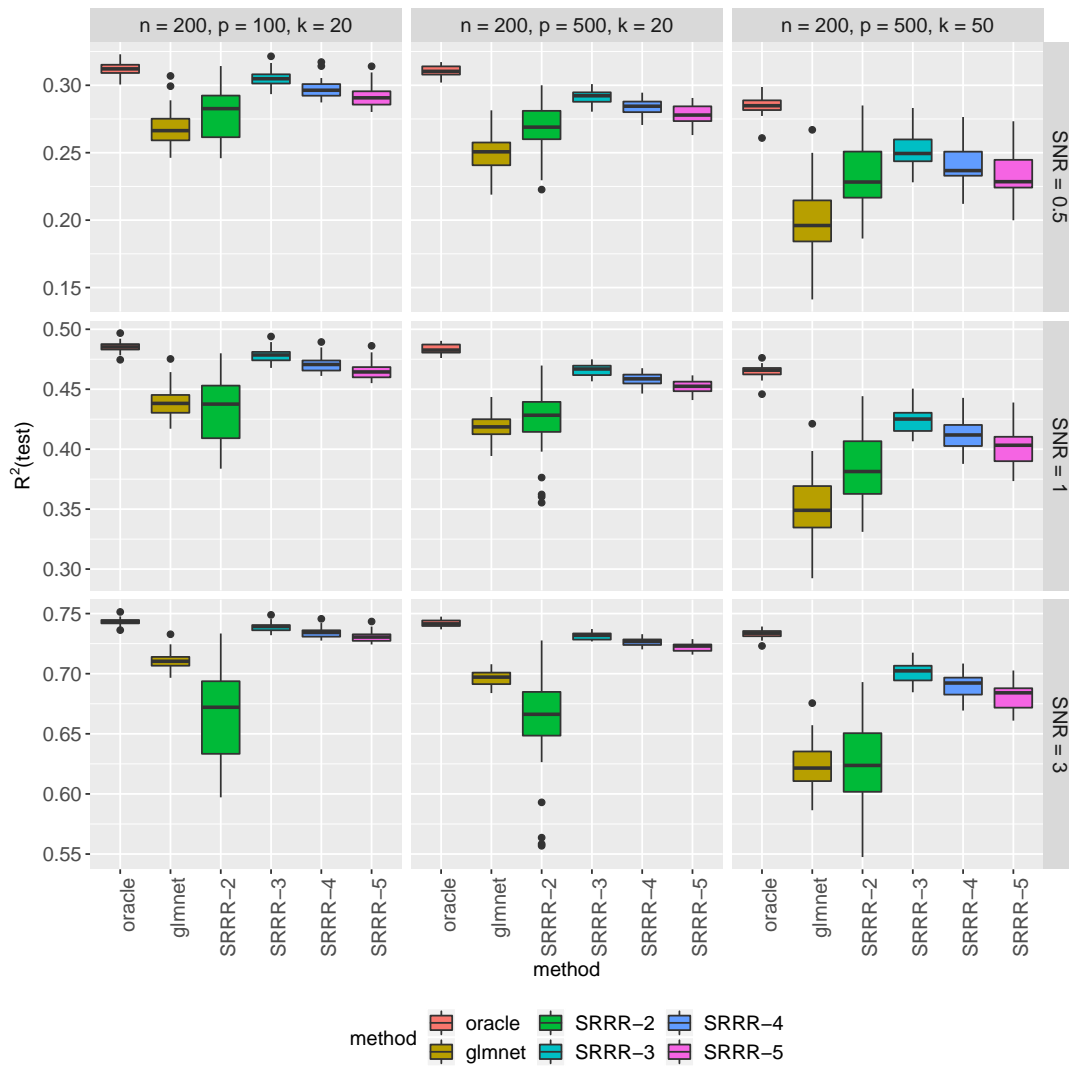


Figure C.3: Scenario 19-27. R^2 each run is evaluated on a test set of size 5000.

811 **Scenario 28-36** Same as Scenario 10-18, except that the predictors are correlated as in Scenario
 812 19-27.

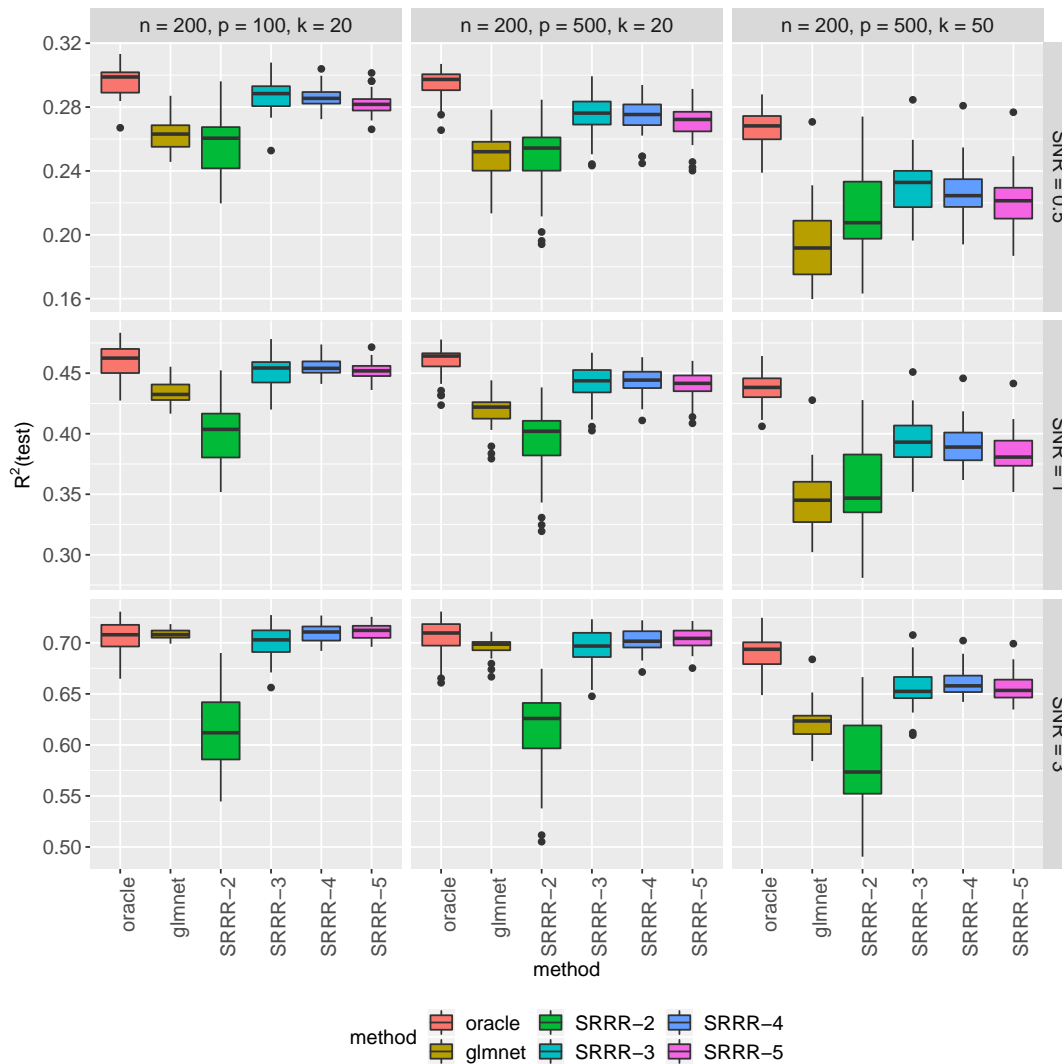


Figure C.4: Scenario 28-36. R^2 each run is evaluated on a test set of size 5000.

813 D Additional Information on the Methods

814 D.1 Compliance with ethical regulations and informed consent

815 This research has been conducted using the UK Biobank Resource under Application Number
 816 24983, “Generating effective therapeutic hypotheses from genomic and hospital linkage data” (<http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf>). Based
 817

818 on the information provided in Protocol 44532 the Stanford IRB has determined that the research
819 does not involve human subjects as defined in 45 CFR 46.102(f) or 21 CFR 50.3(g). All participants
820 of UK Biobank provided written informed consent (more information is available at [https://www.
821 ukbiobank.ac.uk/2018/02/gdpr/](https://www.ukbiobank.ac.uk/2018/02/gdpr/)).

822 D.2 Population stratification in UK Biobank

823 We used genotype data from the UK Biobank dataset release version 2 and the hg19 human genome
824 reference for all analyses in the study. To minimize the variabilities due to population structure in
825 our dataset, we restricted our analyses to include 337,151 White British individuals (Figure D.1)
826 based on the following five criteria (DeBoever et al., 2018; Tanigawa et al., 2019) reported by the
827 UK Biobank in the file “ukb_sqc.v2.txt”:

- 828 1. self- reported white British ancestry (“in_white_British_ancestry_subset” column)
- 829 2. used to compute principal components (“used_in_pca_calculation” column)
- 830 3. not marked as outliers for heterozygosity and missing rates (“het_missing_outliers” column)
- 831 4. do not show putative sex chromosome aneuploidy (“putative_sex_chromosome_aneuploidy”
832 column)
- 833 5. have at most 10 putative third-degree relatives (“excess_relatives” column).

834 D.3 Variant annotation and quality control

835 We prepared a genotype dataset by combining the directly-genotype variants, copy number variants
836 (CNVs) and HLA allelotype datasets.

837 We annotated the directly-genotyped variants using the VEP LOFTEE plugin ([https://github.
838 com/konradjk/loftee](https://github.com/konradjk/loftee)) and variant quality control by comparing allele frequencies in the UK
839 Biobank and gnomAD (gnomad.exomes.r2.0.1.sites.vcf.gz) as previously described²⁸. We focused
840 on variants outside of the major histocompatibility complex (MHC) region (chr6:25477797-36448354)
841 as previously described. We focused on the variants according to the following criteria:

- 842 • Missigness of the variant is less than 1%, considering that two genotyping arrays (the UK
843 BiLEVE array and the UK Biobank array) which covers a slightly different set of variants.
- 844 • Minor-allele frequency is greater than 0.01%, given the recent reports casting questions on
845 the reliability of ultra low-frequency variants.
- 846 • The variant is in the LD-pruned set
- 847 • Hardy-Weinberg disequilibrium test p-value is less than 1.0×10^{-7}
- 848 • Manual cluster plot inspection. We investigated the cluster plots for subset of variants and
849 removed 11 variants that have unreliable genotype calls.
- 850 • Passed the comparison of minor allele frequency with gnomAD dataset as described before

851 CNVs were called by applying PennCNV v1.0.4 on raw signal intensity data from each array
852 within each genotyping batch as previously described. We applied a filter on minor-allele frequency
853 (MAF > 0.01%), which resulted in 8,274 non-rare (MAF > 0.01%) CNVs.

854 The HLA data from the UK Biobank contains all HLA loci (one line per person) in a specific
855 order (A, B, C, DRB5, DRB4, DRB3, DRB1, DQB1, DQA1, DPB1, DPA1). We downloaded these
856 values, which were imputed via the HLA:IMP*2 program (Resource 182); the UK Biobank reports
857 one value per imputed allele, and only the best-guess alleles are reported. Out of the 362 alleles
858 reported in UKB, we used 175 alleles that were present in >0.1% of the population surveyed.

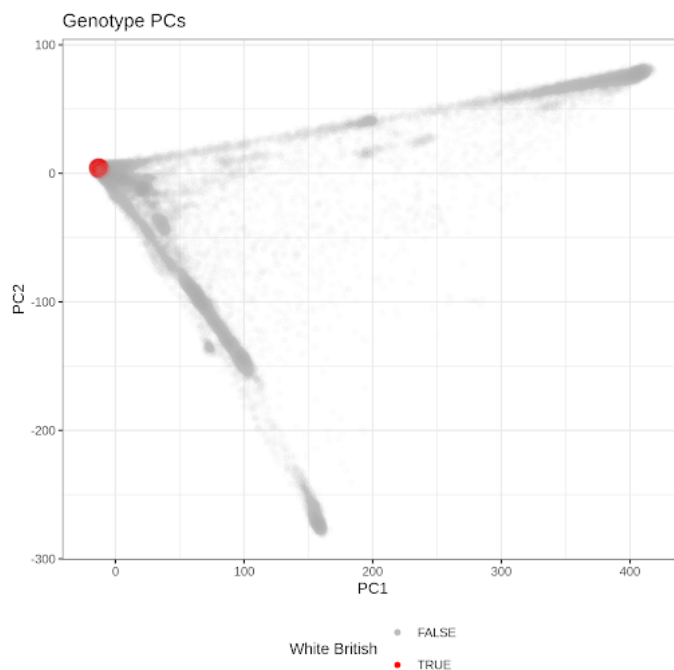


Figure D.1: The identification of unrelated White British individuals in UK Biobank. The first two genotype principal components (PCs) are shown on the x- and y-axis and the identified unrelated White British individuals (Methods) are shown in red.