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

Junyang Qian[1], Yosuke Tanigawa[2], Ruilin Li[3],
Robert Tibshirani[1,2], Manuel A. Rivas*[2] and Trevor Hastie†[1,2]

[1]Department of Statistics, Stanford University
[2]Department of Biomedical Data Science, Stanford University
[3]Institute 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 subproblems. 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 `multiSnpnet` package, available at `http://github.com/junyangq/multiSnpnet` 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

*Corresponding author: mrivas@stanford.edu
†Corresponding author: hastie@stanford.edu

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 $\mathbf{y}^k, k = 1, \ldots, q$ 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

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

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

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

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

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

1.1 Reduced-Rank Regression for Multiple Responses

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

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

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

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

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 $n \geq p$ 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

equivalent to solving q single-response regression problems separately.

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

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

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

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

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

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

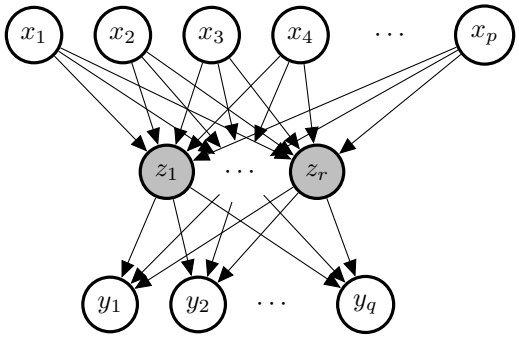

Figure 1: Diagram of the reduced rank regression. The nodes in grey are latent variables. The arrows represent the dependency structure. Known as *multitask* learning in the machine learning community.

[1] We use $\mathbf{v}_k^\top$ to represent the kth 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 jth 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:

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

Alternatively, we can decompose the matrix explicitly as $\mathbf{B} = \mathbf{UV}^\top$ 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):

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

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.

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

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

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

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

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

3.1 Alternating Minimization

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

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

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

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

Proof. See in Appendix A.2. □

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

Algorithm 1 Alternating Minimization

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

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

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

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

    $k = k + 1$
  end while
end for

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

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

3.2 Variable Screening for Ultrahigh-Dimensional Problems

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

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

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

3.2.1 Screening Strategies

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

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

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

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

- (Multi-Gaussian) Solve the full-rank convex problem at λ_k and use its active set as the candidates for the low-rank settings. The main advantage is that the screening is always stable due to the convexity. However this approach often overselects and brings extra burden to the computation. By assuming a higher rank than necessary, the effective number of responses would become more than that of a low-rank model. As a result, more variables would potentially be needed to serve for an enlarged set of responses.

- (Rank-Less) Find variables that have large $c_j = \|\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\mathbf{U}(\lambda_{k-1})\mathbf{V}(\lambda_{k-1})^\top)\|_2$. This is analogous to the strong rules applied to the vanilla multi-response lasso ignoring the rank constraint.

- (Fix-V) Find variables that have large $c'_j = \|\mathbf{X}_j^T(\mathbf{Y}\mathbf{V}(\lambda_{k-1}) - \mathbf{X}\mathbf{U}(\lambda_{k-1}))\|_2$. This is similar to the strong rules applied in the $\mathbf{U}$-step with $\mathbf{V}$ assumed fixed. To see the rationale better, 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

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

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

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

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

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

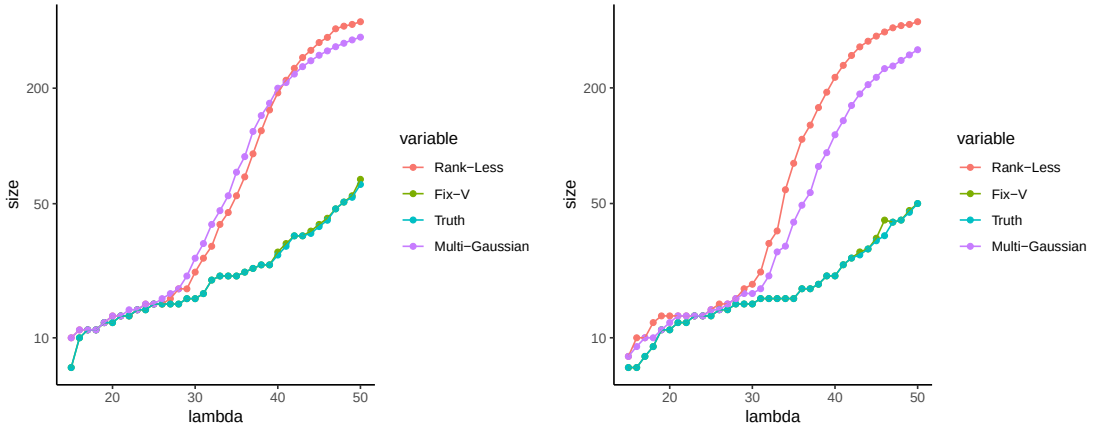

Figure 2: Size of screened set under different strategies. Left: independent design. Right: equi-correlated design with $\rho = 0.5$. Signal-to-noise ratio (SNR) = 1, and we use the true rank = 3.

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

3.2.2 Optimality Condition

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

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

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 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 problem. The argument is supported by the following lemma.

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 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* $j \in S^c$*,*

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

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 objective function in (12) and ∂f is its subdifferential. For the vector ℓ_2-norm, we know that the 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 $\mathbf{X}_S \hat{\mathbf{U}}_S = \mathbf{X}\hat{\mathbf{U}}$ by the definition of $\hat{\mathbf{U}}$. Since we have an exact solution on S, we know $\mathbf{0} \in \partial f(\hat{\mathbf{U}})_j$. for all $j \in S$. On the other hand, for $j \in S^c$, $0 \in \partial f(\hat{\mathbf{U}})$ if and only if $\mathbf{0} \in \{\mathbf{x}_j^\top (\mathbf{X}\hat{\mathbf{U}} - \mathbf{Y}\mathbf{V}) + \lambda \mathbf{s}_j : \|\mathbf{s}_j\|_2 \leq 1\}$, which is further equivalent to $\|\mathbf{x}_j^\top (\mathbf{Y}\mathbf{V} - \mathbf{X}_S \hat{\mathbf{U}}_S)\|_2 = \|\mathbf{x}_j^\top (\mathbf{Y}\mathbf{V} - \mathbf{X}\hat{\mathbf{U}})\|_2 \leq \lambda$ □

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

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

3.3 Computational Considerations

3.3.1 Initialization and Warm Start

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

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

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

3.3.2 Early Stopping

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

3.4 Extensions

3.4.1 Standardization

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

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

3.4.2 Weighting

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

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

equivalently,

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

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

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

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

3.4.3 Adjustment Covariates

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

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

The main components don't change except two adjustments. When determining the starting λ 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*

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

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

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

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 $\hat{\mathbf{W}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top(\mathbf{Y} - \mathbf{XUV}^\top)$. Plug in and we find that the problem to be solved can be written as

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

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 standard group lasso problem and can be solved by using, for example, the **glmnet** package with the `mgaussian` family.

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 $\Omega \subseteq \{1, \ldots, n\} \times \{1, \ldots, p\}$. Let $\mathcal{P}_\Omega : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ be such that

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

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

$$\begin{aligned} \text{minimize} \qquad & \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y}) - \mathcal{P}_\Omega(\mathbf{XUV}^\top)\|_F^2 + \lambda \sum_{j=1}^{p} \|\mathbf{U}^j\|_2, \\ \text{s.t.} \qquad & \mathbf{V}^\top\mathbf{V} = \mathbf{I}. \end{aligned} \tag{11}$$

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

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

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

3.4.5 Lazy Reduced Rank Regression

There is an alternative way to find a low-rank coefficient profile for the multivariate regression. Instead of pursing 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.,

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

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

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

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

3.5 Full Algorithm

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

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

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

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

    [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 $\mathbf{Y}^{(0)} = \mathbf{Y}$.
    while $k = 0$ **or** $\|\mathbf{U}^{(k)}{\mathbf{V}^{(k)}}^\top - \mathbf{U}^{(k-1)}{\mathbf{V}^{(k-1)}}^\top\| > \epsilon$ **do**
      V-step: Fix $\mathbf{U}^{(k)}$, solve $\mathbf{V}$: the orthogonal Procrustes problem

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

      Let $(\mathbf{Y}^{(k)} - \mathbf{ZW}^{(k)})^\top\mathbf{X}_{\mathcal{A}(\lambda_\ell)}\mathbf{U}^{(k)} = \mathbf{MDN}^\top$ (skinny SVD) and solve $\mathbf{V}^{(k+1)} = \mathbf{MN}^\top$.
      U-step: Fix $\mathbf{V}^{(k+1)}$, solve $\mathbf{U}$ and $\mathbf{W}$: the group lasso problem

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

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

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

      $k = k + 1$
    end while
    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 $\mathbf{Y} = \mathbf{Y}^{(k)}$.
    [KKT Check] Check the criterion for all $j \in \Omega \setminus \mathcal{A}(\lambda_\ell)$,

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

    $t = t + 1$
  end while
end for

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:

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

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

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

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

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.

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

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

5 Simulation Studies

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

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

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

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

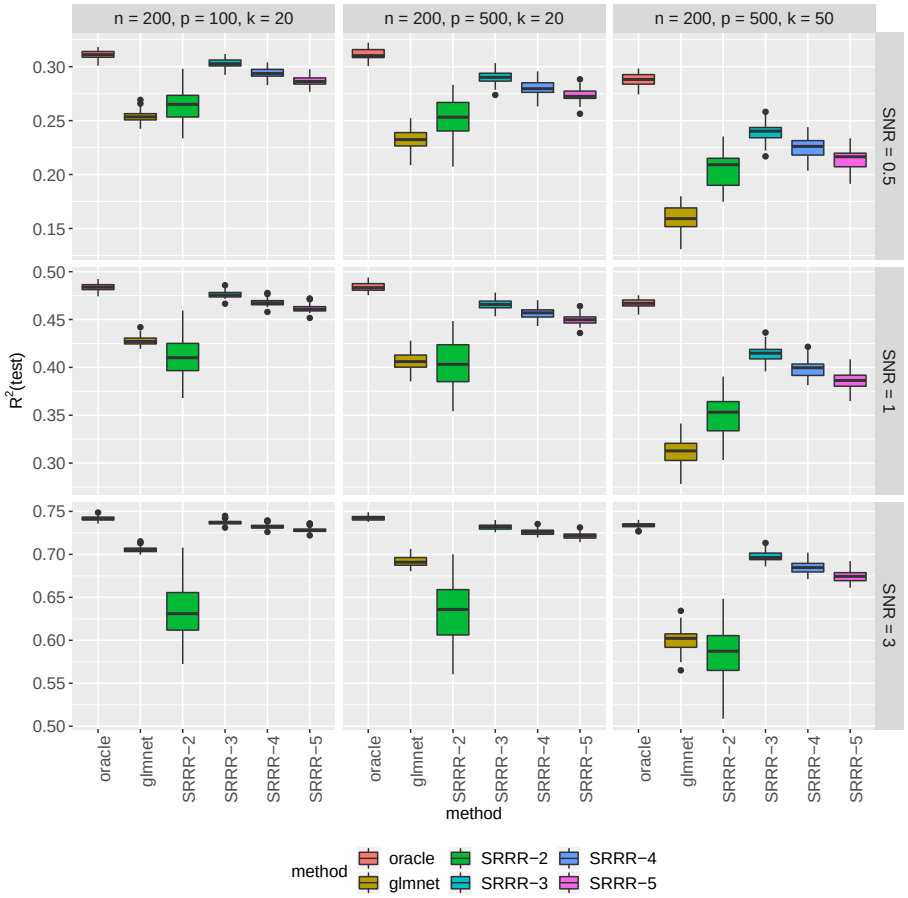

Figure 3: R^2 each run is evaluated on a test set of size 5000. "oracle" is the result where we know the true active variables and solve on this subset of variables. "glmnet" fits the responses separately. "SRRR-r" indicates the SRRR results with assumed rank r.

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

6 Real Data Application: UK Biobank

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

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

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

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

The number of samples is large in the UK Biobank dataset, so we afford to set aside an independent validation set without resorting to costly cross-validation to find an optimal regularization parameter. We also leave out a subset of observations as test set to evaluate the final model. In particular, we randomly partition the original dataset so that 70% is used for training, 10% for validation and 20% for test. The solution path is fit on the training set, whereas the desired 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 $\{(x_i, y_i), 1 \le i \le n\}$, it is defined as

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

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

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

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

6.1 Asthma and 7 Blood Biomarkers

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

We apply the SRRR to the set of phenotypes and expect some performance improvement by leveraging the correlation structure. Choice of the phenotypes: monocyte count, neutrophill count, eosinophill count, basophill count, forced vital capacity (FVC), peak expiratory flow (PEF), and forced expiratory volume in 1 second (FEV1).

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

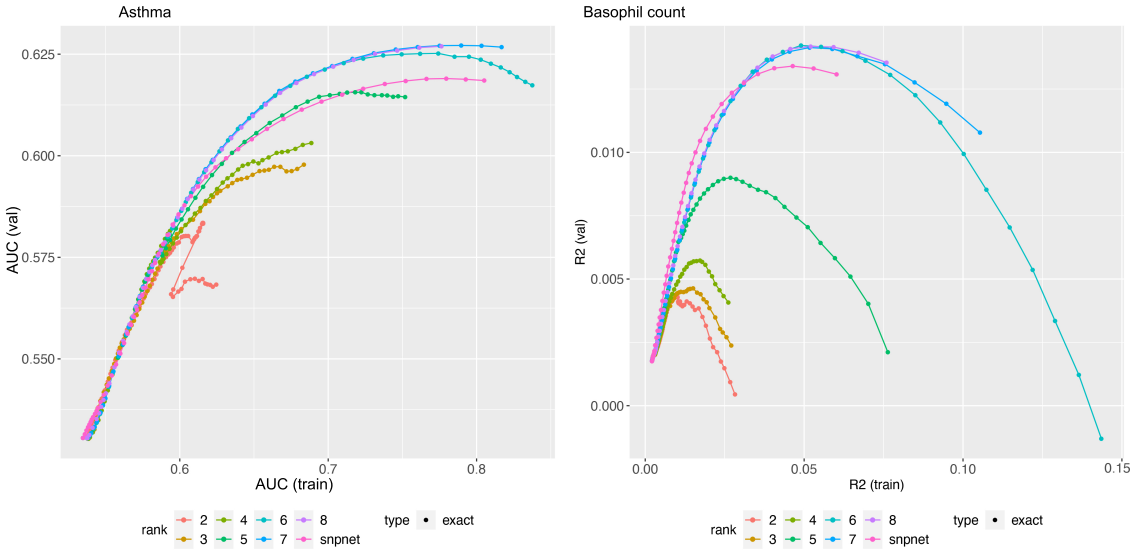

Figure 4: Asthma and Basophil count prediction performance plots. Different colors correspond to lower rank predictive performance across (x-axis) training data set and (y-axis) validation data set for (left) asthma and (right) basophil count.

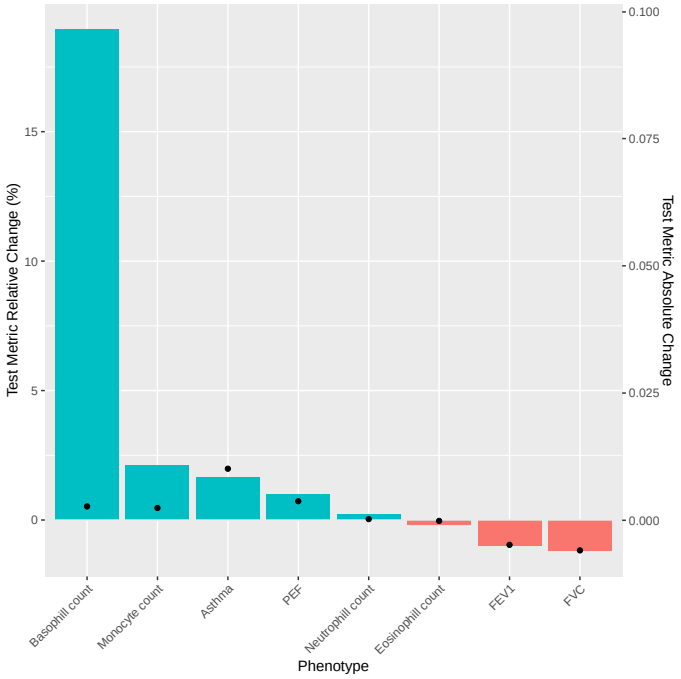

Figure 5: Change in prediction accuracy for multiresponse model compared to single response model. (top) (y-axis 1 bar) R^2 relative change (%) for each phenotype (x-axis) and R^2 absolute change (y-axis 2).

6.2 35 Biomarkers

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

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

for AST to ALT ratio (see Figure 8).

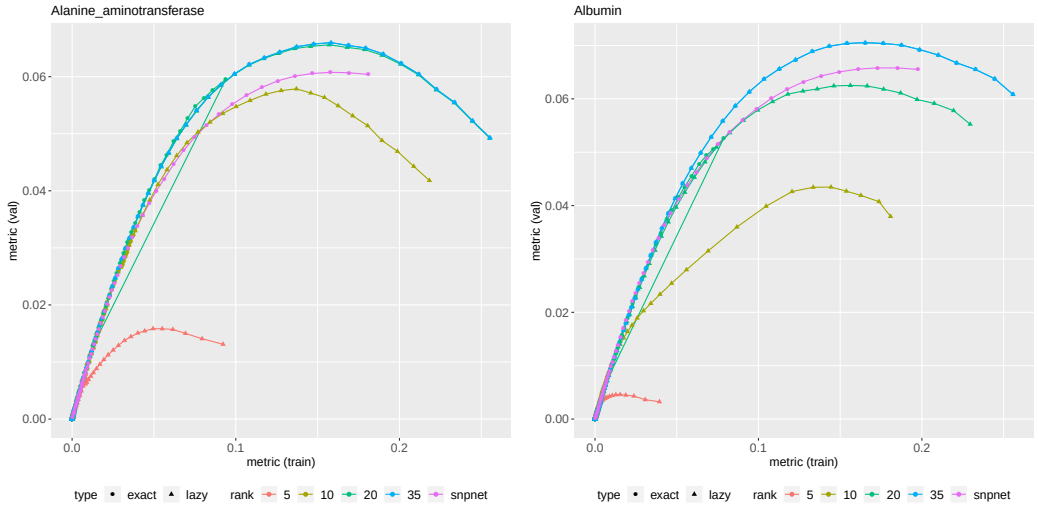

Figure 6: Alanine aminotransferase and albumin prediction performance plots. Different colors correspond to lower rank predictive performance across (x-axis) training data set and (y-axis) validation data set for (left) alanine aminotransferase and (right) albumin. For lower rank representation we applied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

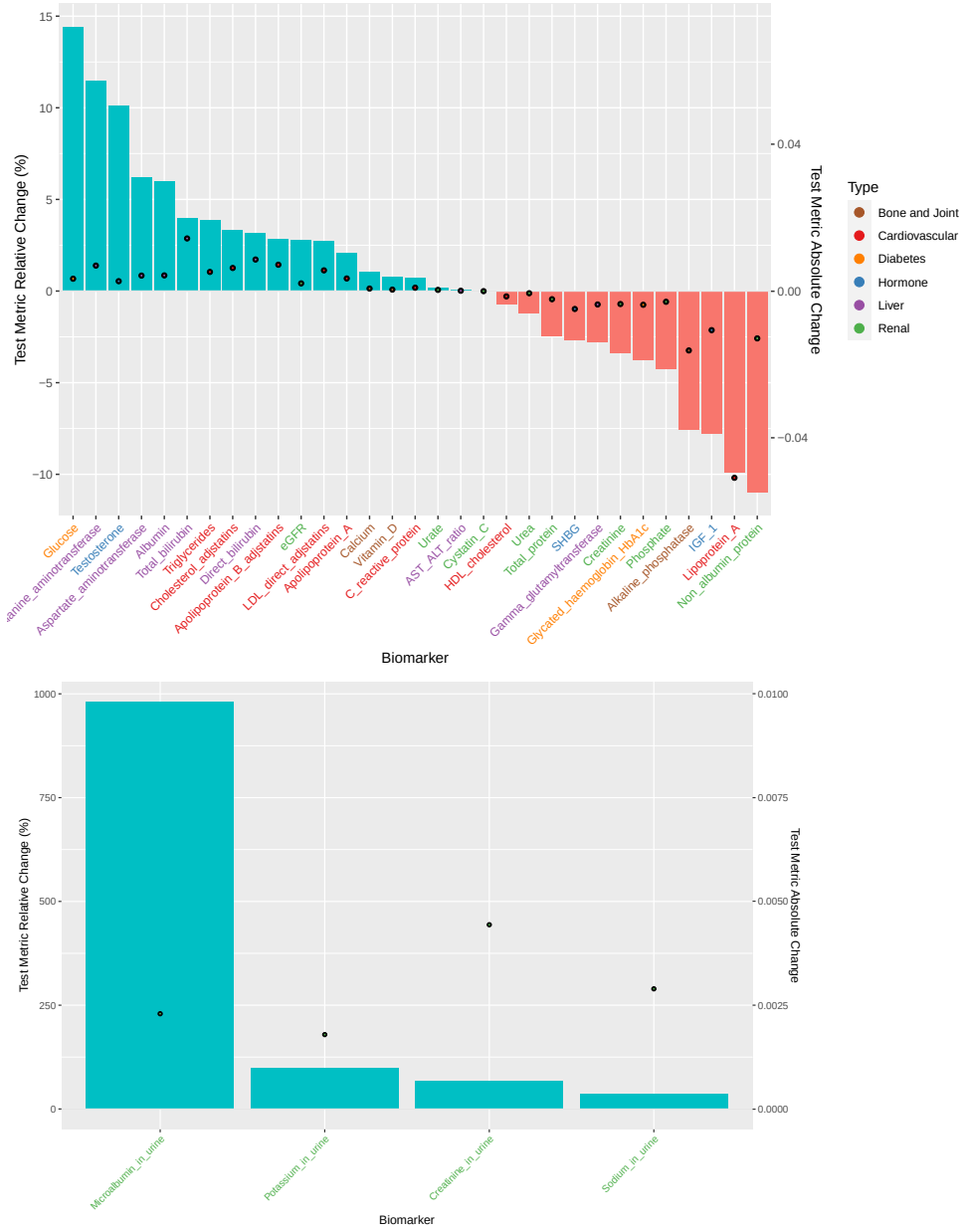

Figure 7: Change in prediction accuracy for multiresponse model compared to single response model. (top) (y-axis 1 bar) R^2 relative change (%) for each biomarker (x-axis) across different biomarker category (color) and R^2 absolute change (y-axis 2). (bottom) Change in predictive accuracy for multiresponse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

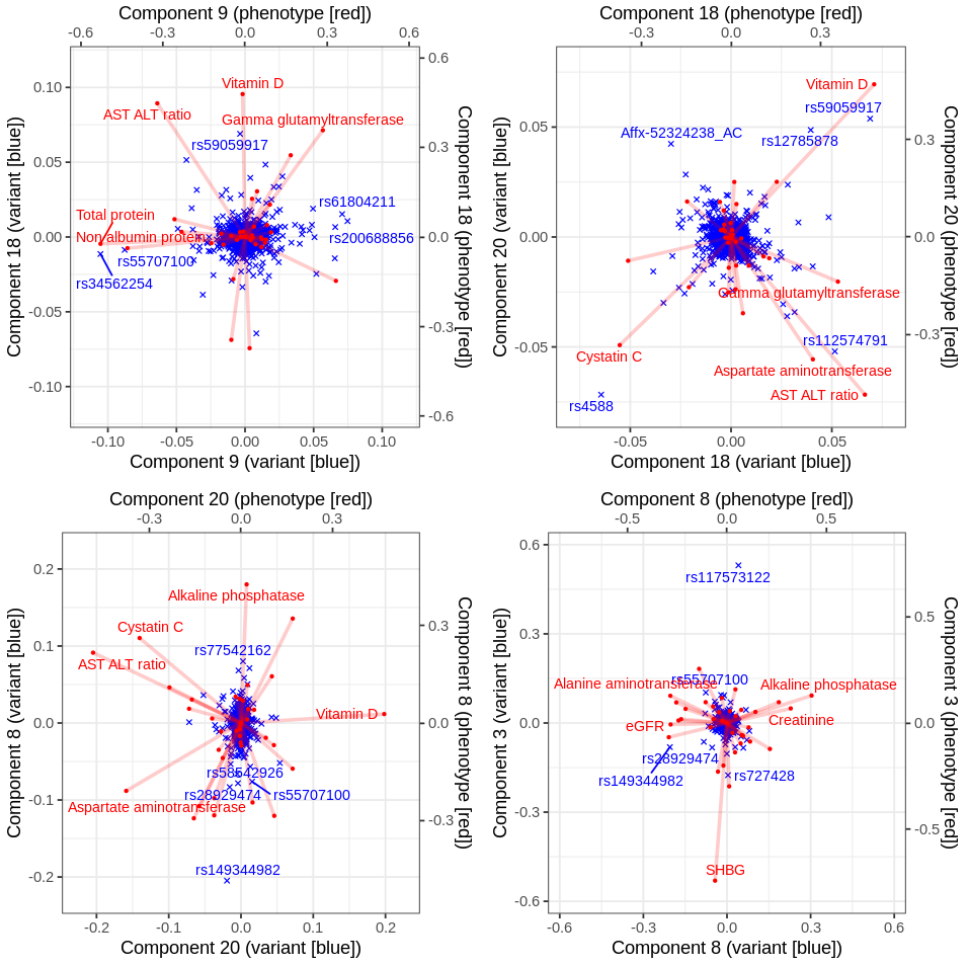

Figure 8: The latent structures of the the top five key SRRR components for AST to ALT ratio. Using trait squared cosine score described in Tanigawa et al. (2019), the top five key SRRR components for AST to ALT ratio (components 9, 18, 20, 8, and 3) are identified from a full-rank SVD of coefficient 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.

7 Related Work

There are many other methods that were proposed for multivariate regression in high-dimensional settings. Chen, Huang (2012) compares the SRRR with rank-free methods including L_2SVS Similä, Tikka (2007), L_∞SVS (Turlach et al., 2005) that replaces the ℓ_2-norm with ℓ_∞-norm of each row, and RemMap (Peng et al., 2010) that imposes an additional elementwise sparsity of the coefficient

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

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

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

8 Summary and Discussion

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

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

Acknowledgement

This research has been conducted using the UK Biobank Resource under Application Number 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 on the information provided in Protocol 44532 the Stanford IRB has determined that the research does not involve human subjects as defined in 45 CFR 46.102(f) or 21 CFR 50.3(g). All participants of UK Biobank provided written informed consent (more information is available at `https://www.ukbiobank.ac.uk/2018/02/gdpr/`). We thank all the participants in the UK Biobank. M.A.R. is supported by Stanford University and a National Institutes of Health (NIH) Center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080). Y.T. is supported by a Funai Overseas Scholarship from the Funai Foundation for Information Technology and the Stanford University School of Medicine. Research reported in this publication was supported by the National Human Genome Research Institute of the NIH under Award Number R01HG010140 (M.A.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. R.T. was partially supported by NIH grant 5R01 EB001988-16 and NSF grant 19 DMS1208164. T.H. was partially supported by grant DMS-1407548 from the National Science Foundation, and grant 5R01 EB 001988-21 from the National Institutes of Health.

References

Abadi Martín, Barham Paul, Chen Jianmin, Chen Zhifeng, Davis Andy, Dean Jeffrey, Devin Matthieu, Ghemawat Sanjay, Irving Geoffrey, Isard Michael, Kudlur Manjunath, Levenberg Josh, Monga Rajat, Moore Sherry, Murray Derek G., Steiner Benoit, Tucker Paul, Vasudevan Vijay, Warden Pete, Wicke Martin, Yu Yuan, Zheng Xiaoqiang. **TensorFlow**: A System for Large-scale Machine Learning // Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. Berkeley, CA, USA: USENIX Association, 2016. 265–283. (OSDI'16).

Aguirre Matthew, Tanigawa Yosuke, Venkataraman Guhan, Tibshirani Rob, Hastie Trevor, Rivas Manuel A. Polygenic risk modeling with latent trait-related genetic components // BioRxiv. 2019. 808675.

Anderson T. W. Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions // Ann. Math. Statist. 09 1951. 22, 3. 327–351.

Bach Francis R. Consistency of the group lasso and multiple kernel learning // Journal of Machine Learning Research. 2008. 9, Jun. 1179–1225.

Bickel Peter J., Ritov Ya'acov, Tsybakov Alexandre B. Simultaneous analysis of Lasso and Dantzig selector // Ann. Statist. 08 2009. 37, 4. 1705–1732.

Bottou Léon. Large-scale machine learning with stochastic gradient descent // Proceedings of COMPSTAT'2010. 2010. 177–186.

Bovet Daniel P, Cesati Marco. Understanding the Linux Kernel: from I/O ports to process management. 2005.

Boyd Stephen, Boyd Stephen P, Vandenberghe Lieven. Convex optimization. 2004.

Bühlmann Peter, Van De Geer Sara. Statistics for high-dimensional data: methods, theory and applications. 2011.

Bycroft Clare, Freeman Colin, Petkova Desislava, Band Gavin, Elliott Lloyd T., Sharp Kevin, Motyer Allan, Vukcevic Damjan, Delaneau Olivier, O?Connell Jared, Cortes Adrian, Welsh Samantha, Young Alan, Effingham Mark, McVean Gil, Leslie Stephen, Allen Naomi, Donnelly Peter, Marchini Jonathan. The UK Biobank Resource with Deep Phenotyping and Genomic Data // Nature. 2018. 562, 7726. 203–209.

Chang Christopher C, Chow Carson C, Tellier Laurent CAM, Vattikuti Shashaank, Purcell Shaun M, Lee James J. Second-generation PLINK: rising to the challenge of larger and richer datasets // GigaScience. 02 2015. 4, 1.

Chen Lisha, Huang Jianhua Z. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection // Journal of the American Statistical Association. 2012. 107, 500. 1533–1545.

Chu Moody T, Trendafilov Nickolay T. On a differential equation approach to the weighted orthogonal Procrustes problem // Statistics and Computing. 1998. 8, 2. 125–133.

Chun Hyonho, Keleş Sündüz. Sparse partial least squares regression for simultaneous dimension reduction and variable selection // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010. 72, 1. 3–25.

DeBoever Christopher, Tanigawa Yosuke, Lindholm Malene E., McInnes Greg, Lavertu Adam, Ingelsson Erik, Chang Chris, Ashley Euan A., Bustamante Carlos D., Daly Mark J., Rivas Manuel A. Medical Relevance of Protein-Truncating Variants across 337,205 Individuals in the UK Biobank Study // Nature Communications. 2018. 9, 1. 1612.

Dean Jeffrey, Ghemawat Sanjay. MapReduce: Simplified Data Processing on Large Clusters // Commun. ACM. I 2008. 51, 1. 107–113.

Dubois Benjamin, Delmas Jean-François, Obozinski Guillaume. Fast Algorithms for Sparse Reduced-Rank Regression // Proceedings of Machine Learning Research. 89. 16–18 Apr 2019. 2415–2424. (Proceedings of Machine Learning Research).

Duchi John C, Agarwal Alekh, Wainwright Martin J. Dual averaging for distributed optimization: Convergence analysis and network scaling // IEEE Transactions on Automatic control. 2011. 57, 3. 592–606.

Efron Bradley, Hastie Trevor. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. 5. 2016.

Friedman Jerome, Hastie Trevor, Tibshirani Rob. Regularization Paths for Generalized Linear Models via Coordinate Descent. 2010. 1–22.

Gabriel Karl Ruben. The biplot graphic display of matrices with application to principal component analysis // Biometrika. 1971. 58, 3. 453–467.

Gower John C, Lubbe Sugnet Gardner, Le Roux Niel J. Understanding biplots. 2011.

Greenshtein Eitan, Ritov Ya'Acov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization // Bernoulli. 12 2004. 10, 6. 971–988.

Hastie Trevor, Tibshirani Robert, Friedman Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. 2009. (Springer series in statistics).

Hotelling Harold. Relations Between Two Sets of Variates // Biometrika. 1936. 28, 3/4. 321–377.

Lello Louis, Avery Steven G, Tellier Laurent, Vazquez Ana I, Campos Gustavo de los, Hsu Stephen DH. Accurate genomic prediction of human height // Genetics. 2018. 210, 2. 477–497.

Li Gen, Liu Xiaokang, Chen Kun. Integrative multi-view regression: Bridging group-sparse and low-rank models // Biometrics. 2019. 75, 2. 593–602.

Li Ruilin, Chang Christopher, Justesen Johanne Marie, Tanigawa Yosuke, Qian Junyang, Hastie Trevor, Rivas Manuel A, Tibshirani Robert j. Fast Lasso method for Large-scale and Ultrahigh-dimensional Cox Model with applications to UK Biobank // BioRxiv. 2020.

Lin Jake, Tabassum Rubina, Ripatti Samuli, Pirinen Matti. MetaPhat: Detecting and decomposing multivariate associations from univariate genome-wide association statistics // bioRxiv. 2019. 661421.

Luo Chongliang, Liang Jian, Li Gen, Wang Fei, Zhang Changshui, Dey Dipak K, Chen Kun. Leveraging mixed and incomplete outcomes via reduced-rank modeling // Journal of Multivariate Analysis. 2018. 167. 378–394.

Ma Zhuang, Ma Zongming, Sun Tingni. Adaptive Estimation in Two-way Sparse Reduced-rank Regression // Statistica Sinica. 01 2020.

Ma Zongming, Sun Tingni. Adaptive sparse reduced-rank regression // arXiv preprint arXiv:1403.1922. 2014.

Mazumder Rahul, Hastie Trevor, Tibshirani Robert. Spectral regularization algorithms for learning large incomplete matrices // Journal of Machine Learning Research. 2010. 11, Aug. 2287–2322.

Meinshausen Nicolai, Bühlmann Peter. High-dimensional graphs and variable selection with the Lasso // Ann. Statist. 06 2006. 34, 3. 1436–1462.

Mooijaart Ab, Commandeur Jacques JF. A general solution of the weighted orthonormal Procrustes problem // Psychometrika. 1990. 55, 4. 657–663.

Obozinski Guillaume, Wainwright Martin J, Jordan Michael I, others . Support union recovery in high-dimensional multivariate regression // The Annals of Statistics. 2011. 39, 1. 1–47.

Peng Jie, Zhu Ji, Bergamaschi Anna, Han Wonshik, Noh Dong-Young, Pollack Jonathan R, Wang Pei. Regularized multivariate regression for identifying master predictors with application to 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.

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 in Cloud Computing. Berkeley, CA, USA: USENIX Association, 2010. 10–10. (HotCloud'10).

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

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

A Additional Proofs

A.1 Proof of Lemma 1

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

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

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

A.2 Proof of Lemma 2

We plug in the SVD of $\mathbf{Z}$ and have $\mathbf{Tr}(\mathbf{Z}^\top \mathbf{V}) = \mathbf{Tr}(\mathbf{N}\mathbf{D}\mathbf{M}^\top \mathbf{V}) = \mathbf{Tr}(\mathbf{D}\mathbf{M}^\top \mathbf{V}\mathbf{N}) = \sum_{k=1}^r \mathbf{D}_{kk}\mathbf{S}_{kk}$, 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 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 magnitude of its diagonal elements cannot exceed 1. Since $\mathbf{D}_{kk}$ are all non-negative. To maximize $\sum_{k=1}^r \mathbf{D}_{kk}\mathbf{S}_{kk}$, we let $\mathbf{S}_{kk} = 1$ for all $1 \leq k \leq r$. This is equivalent to setting $\mathbf{S} = \mathbf{M}^\top \mathbf{V}\mathbf{N} = \mathbf{I}$. Therefore, one solution is given by $\mathbf{V} = \mathbf{M}\mathbf{N}^\top$. The maximum value of the objective is thus $\sum_{k=1}^r \mathbf{D}_{kk} = \|\mathbf{Z}\|_*$, the nuclear norm of $\mathbf{Z}$.

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 $(1/2)\|\mathbf{Y}\|_F^2$):

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

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 $F_\lambda(\mathbf{U}) = f_\lambda(\mathbf{U}) + \lambda \sum_{j=1}^p \|\mathbf{U}_{j\cdot}\|_2$. Define a local smooth approximation of F_λ as

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

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

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

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

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

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 that $\operatorname{\mathbf{Tr}}((\mathbf{U} - \mathbf{U}^k)^\top\mathbf{X}^\top\mathbf{X}(\mathbf{U} - \mathbf{U}^k)) \leq \sigma_{\max}^2\|\mathbf{U} - \mathbf{U}^k\|_F^2$, and the last equality is by the definition of $\tilde{F}_\lambda^t$ function. Therefore, if we let $\mathbf{U}^{k,+} = \operatorname{argmin}_{\mathbf{U}}[\tilde{F}_\lambda^{1/\sigma_{\max}^2}(\mathbf{U}; \mathbf{U}^k) - F_\lambda(\mathbf{U}^k)]$, we have

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

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, 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

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

and the convergence is linear.

B Connection with CCA

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

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

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

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

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

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 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 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 solves

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

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 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 due to the constraint.

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

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

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

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

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 $\Gamma = \mathbf{S}_{yy}^{-1}$ is closely related to the relaxed CCA solution. $\mathbf{U}$ is the same while $\mathbf{V}$ is the so-called reflexive inverse of $\mathbf{V}$ there.

C Additional Experiments

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

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

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

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

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

We consider several sets of experiments.

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

2. **Scenario 10-18** Same as Scenario 1-9. The true coefficient matrix is no longer exact low 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,

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

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

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

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

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

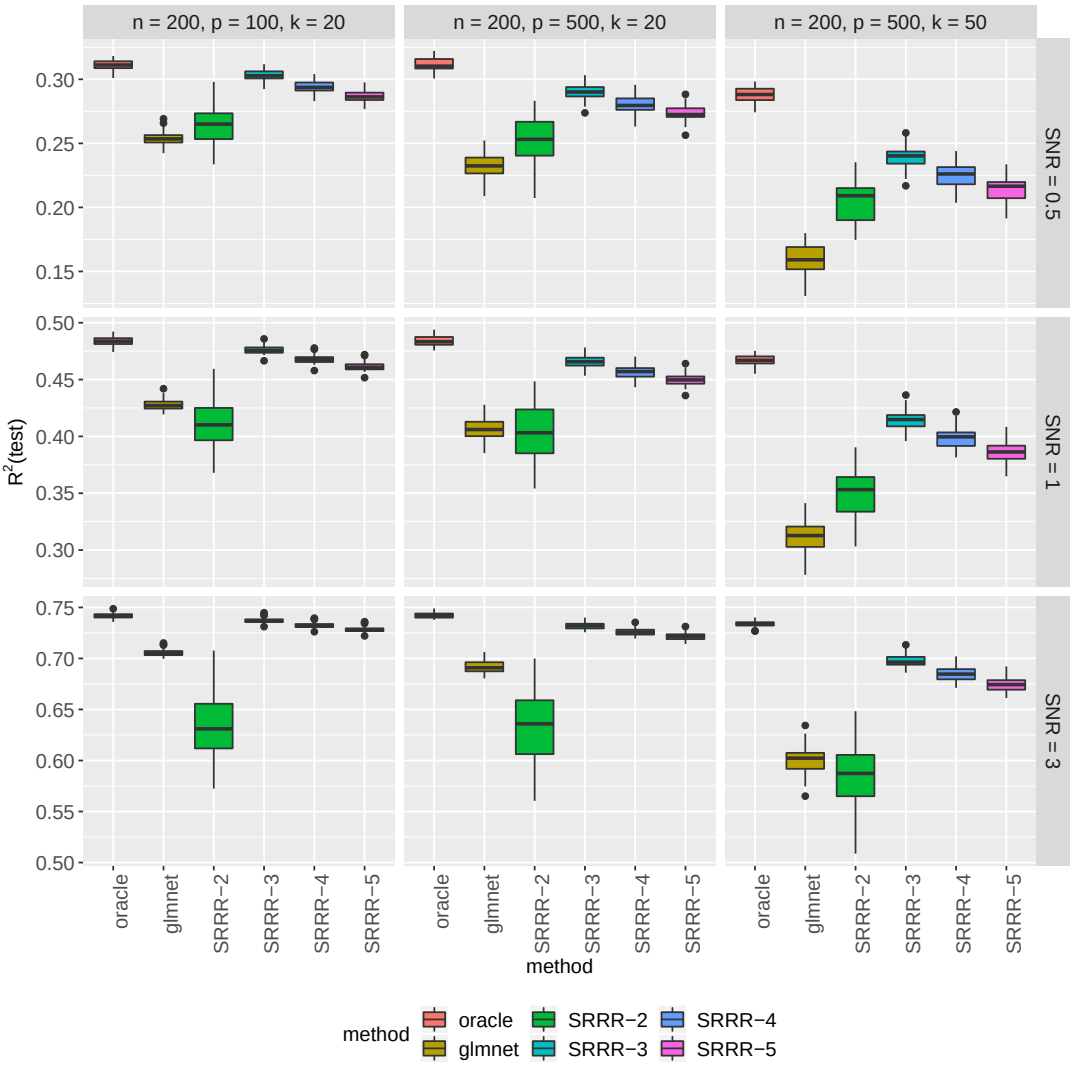

Figure C.1: Scenario 1-9. R^2 each run is evaluated on a test set of size 5000.

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

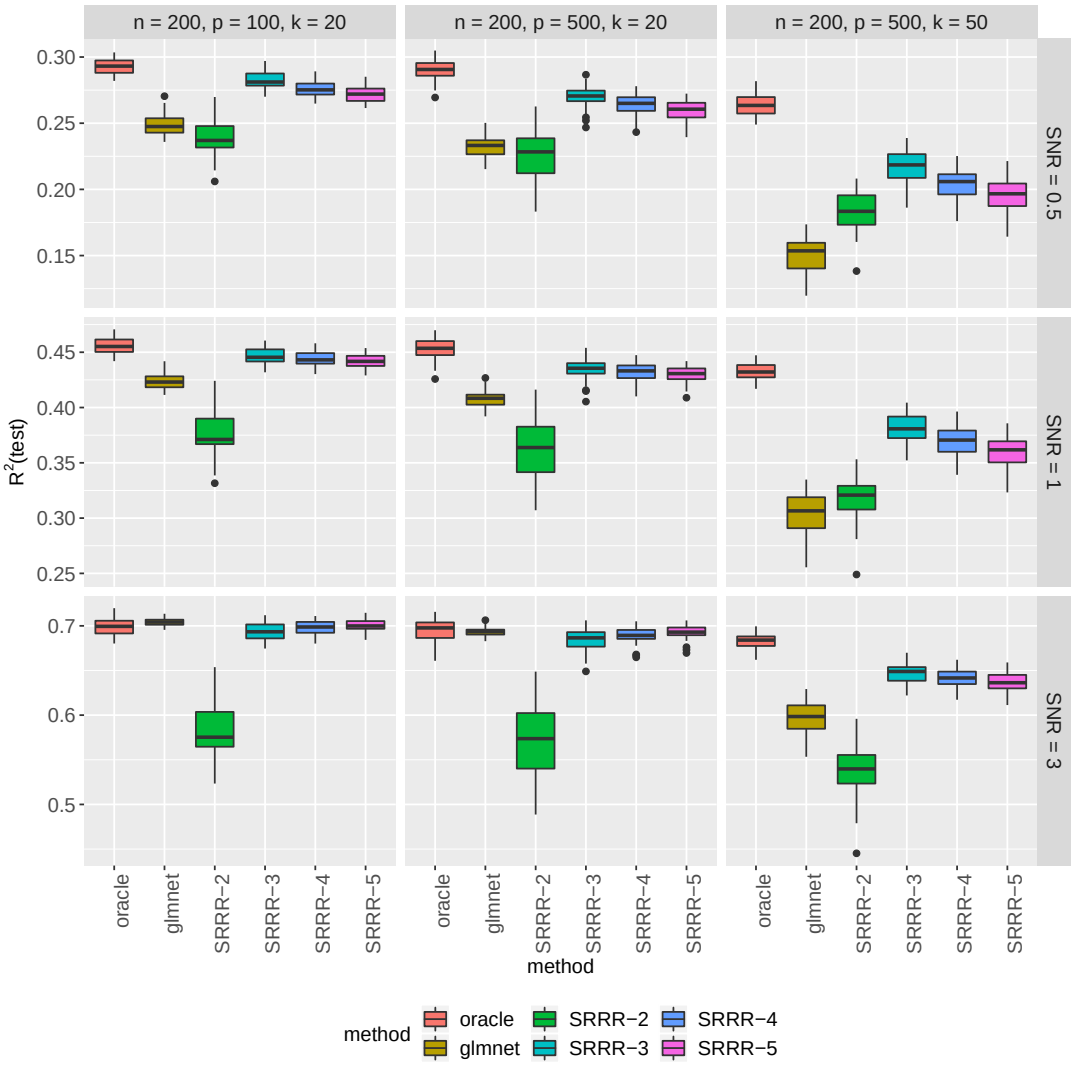

Figure C.2: Scenario 10-18. R^2 each run is evaluated on a test set of size 5000. The oracle here does not take into account the noise in true coefficient matrix, and do reduced rank regression on the true support and the true rank.

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

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

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

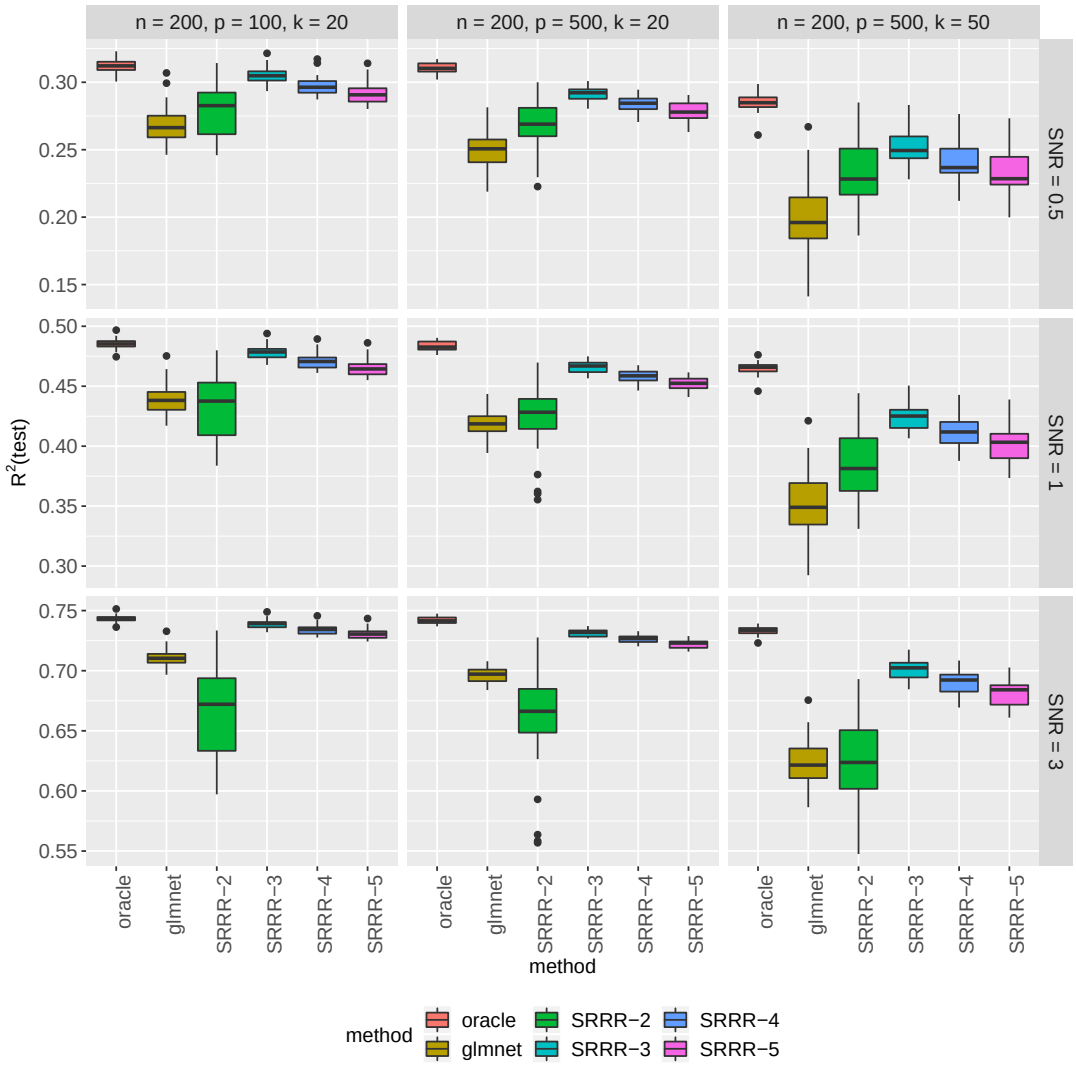

Figure C.3: Scenario 19-27. R^2 each run is evaluated on a test set of size 5000.

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

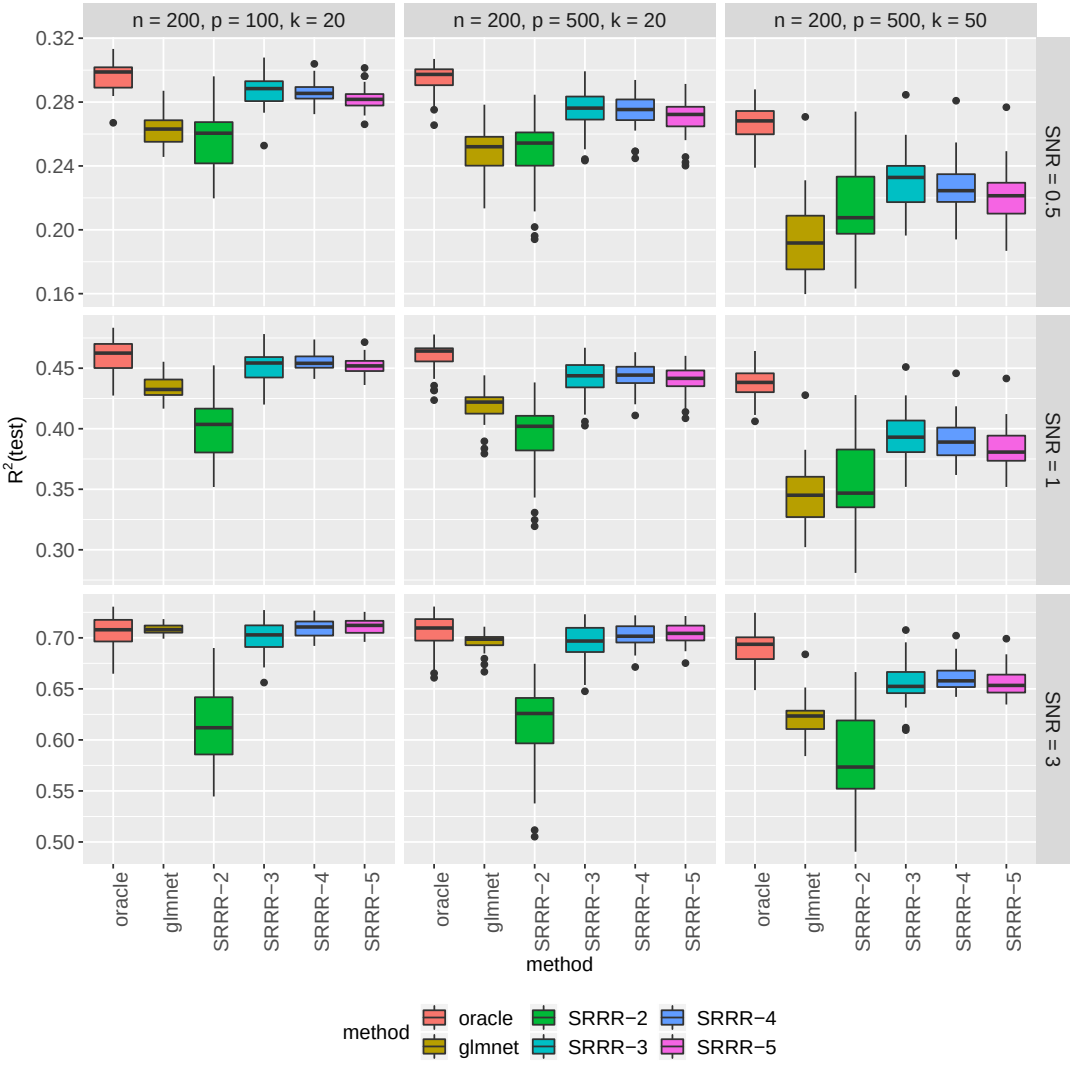

Figure C.4: Scenario 28-36. R^2 each run is evaluated on a test set of size 5000.

D Additional Information on the Methods

D.1 Compliance with ethical regulations and informed consent

This research has been conducted using the UK Biobank Resource under Application Number 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

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

D.2 Population stratification in UK Biobank

We used genotype data from the UK Biobank dataset release version 2 and the hg19 human genome reference for all analyses in the study. To minimize the variabilities due to population structure in our dataset, we restricted our analyses to include 337,151 White British individuals (Figure D.1) based on the following five criteria (DeBoever et al., 2018; Tanigawa et al., 2019) reported by the UK Biobank in the file "ukb_sqc_v2.txt":

1. self- reported white British ancestry ("in_white_British_ancestry_subset" column)
2. used to compute principal components ("used_in_pca_calculation" column)
3. not marked as outliers for heterozygosity and missing rates ("het_missing_outliers" column)
4. do not show putative sex chromosome aneuploidy ("putative_sex_chromo- some_aneuploidy" column)
5. have at most 10 putative third-degree relatives ("excess_relatives" column).

D.3 Variant annotation and quality control

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

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

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

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

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

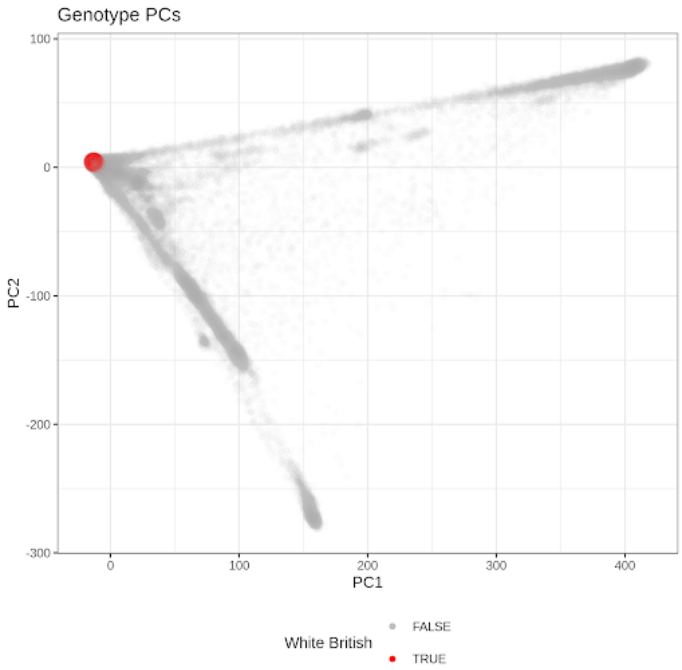

Figure D.1: The identification of unrelated White British individuals in UK Biobank. The first two genotype principal components (PCs) are shown on the x- and y-axis and the identified unrelated White British individuals (Methods) are shown in red.