Asymptotic Inference for Constrained Regression

By MADHAV SANKARANARAYANAN

Department of Biostatistics, Harvard T.H. Chan School of Public Health,
Boston, MA, USA

madhav_sankaranarayanan@g.harvard.edu

YANA HRYTSENKO

Cardiovascular Institute, Beth Israel Deaconess Medical Center,
Boston, MA, USA

yhrytsenko@gmail.com

JEROME I. ROTTER

The Institute for Translational Genomics and Population Sciences, Department of Pediatrics,
The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center
Torrance, CA, USA

jrotter@lundquist.org

TAMAR SOFER

Cardiovascular Institute, Beth Israel Deaconess Medical Center,
Boston, MA, USA

tsofer@bidmc.harvard.edu

RAJARSHI MUKHERJEE

Department of Biostatistics, Harvard T.H. Chan School of Public Health,
Boston, MA, USA

rmukherj@hsph.harvard.edu

SUMMARY

We consider statistical inference in high-dimensional regression problems under affine constraints on the parameter space. The theoretical study of this is motivated by the study of genetic determinants of diseases, such as diabetes, using external information from mediating protein expression levels. Specifically, we develop rigorous methods for estimating genetic effects on diabetes-related continuous outcomes when these associations are constrained based on external information about genetic determinants of proteins, and genetic relationships between proteins and the outcome of interest. In this regard, we discuss multiple candidate estimators and study their theoretical properties, sharp large sample optimality, and numerical qualities under a high-dimensional proportional asymptotic framework. Finally, we apply the developed methods to study the genetic determinants of BMI, fasting insulin and HbA1c, leveraging their genetic correlation with protein expression obtained from an external study.

Some key words: Asymptotic inference; Constrained linear regression; Genetic correlation; Genome wide association study; High dimensional inference; Metabolic trait.

1. Introduction

1.1. *Motivation*

In this paper, we consider improved statistical inference in high-dimensional linear regression problems by leveraging known affine constraints on the underlying unknown parameter vector drawn from external sources. Specifically, we operate under a proportional asymptotic framework that allows us to derive precise efficiency gains implied by the number of independent constraints. Our main goal is to calibrate the precise statistical gains implied by the assumed constraints, to develop estimation methods to efficiently extract additional information, and provide asymptotically valid uncertainty quantification. In practical applications, these constraints are reliant on domain knowledge or external data. A specific instance of this arises quite naturally in the context of genetic disease modeling, and we will be using these methodologies to improve inference on the genetic pathways for diabetes. Diabetes is a chronic disease that affects an individual's ability to effectively utilize insulin and regulate blood sugar levels. While the onset of diabetes, in particular, type 2 diabetes or adult onset diabetes, is influenced by external factors, there is a marked genetic component of risk (Cole & Florez, 2020). As a result, there is a pressing need for tools and techniques that leverage genetics for the early detection and prevention of diabetes. As diabetes diagnostics are based on glycemic traits, investigating genetic factors that impact glycemic and related metabolic traits can shed crucial light on the relationship between one's genetic profile and risk of diabetes. Traditional analyses rely on genome-wide association studies (GWAS) to provide evidence of associations between genetic loci and phenotypes of interest. However, proteins mediate the effect of genetic variation on disease, and thus following trends in recent research, we leverage the advent and popularization of protein quantification technologies (He et al., 2020), to utilize intermediary data and improve the effect estimation of genetic mutation on disease traits.

We operate under this very framework and incorporate pleiotropic information, i.e., use effect sizes from a genetic variant's association with a non-target trait, to enhance the efficiency of genetic effect estimation on metabolic traits of interest. Specifically, our methodological paradigm aims to leverage pleiotropic information through *genetic correlations*, the measure of correlation between traits that are dependent on genetic contribution alone. In this regard, genetic correlations between protein expression levels and metabolic traits obtained from a reference population allow us to posit a *constraint* on the effect of involved genetic variants of interest, and thereby increase statistical accuracy of effect estimation in a target population. The main results of this paper can be summarized as follows: (i) we develop theoretically rigorous procedures for incorporating protein expression data into the pipeline of genetic effect estimation; (ii) we demonstrate statistical efficiency of our methods under moderately high dimensional settings; (iii) in higher dimensional settings, we provide novel methods that provide valid inference under some additional regularity conditions; (iv) we apply these methods to real genetic datasets for causal variant discovery.

1.2. *Literature*

The work in this paper follows from the classical literature of constrained linear regressions (Aitchison & Silvey, 1958; Amemiya, 1985), more recent literature on regression with random designs (Hsu et al., 2014; Mourtada, 2022), and the application of large random matrix theory to statistical problems in proportional asymptotic regimes (Vaart, 1998; Dobriban & Sheng, 2022; Dobriban & Wager, 2018). The problem of estimating parameters in constrained spaces arises in various settings, thus there are many avenues that have been explored in prior works. The affine constraints setup is explored in Yu (2020) under a fixed design model and presents an algorithmic framework for constructing estimators under sparsity in high dimensions. Shi et al.

(2016) allows for simplicial constraints in high dimensions with sparsity, a natural consideration in microbiome data. Aside from linear constraints, elliptical constraints have been studied in fixed design contexts (Donoho, 1994) and more recently, in general random designs (Pathak et al., 2023). Also, conical constraints, arising from higher-order constraints are investigated in Yu et al. (2019). The techniques used for these constraints are fundamentally different from the techniques required for affine constraints, but provide insights on the difficulty of estimation in constrained parameter spaces. More generally, Han (2022) provides exact risk asymptotics for recovery under convex constraints with Gaussian design. We leverage the additional structure of affine constraints and general universality results from random matrix theory to obtain more precise and interpretable error bounds in our work. There is also a large body of work on more general regression setups with constraints (Klaassen & Susyanto, 2016), estimation in general random design regression (Tsybakov, 2009; Hsu et al., 2014; Györfi, 2002), estimation under inequality constraints (Shapiro, 1989), and even estimation beyond proportional asymptotic regimes (Cheng & Montanari, 2024).

Our methods provide techniques to conduct statistical inference in high-dimensional settings, providing strict minimax optimality conditions akin to Mourtada (2022), in general design setups under proportional asymptotics. The formulation is akin to flavours of transfer learning in regression setups (Li et al., 2022a,b), but with problem-specific definitions of similarity across sites.

These methods arise directly from the field of genetic disease etiology. Large-scale genetic data analysis to study disease biology is an ever-growing field of research that addresses prediction and elucidation of causal relationships between risk factors and health and outcomes (Khoury et al., 1985; Holtzman & Marteau, 2000; Cedric Gondro, 2013; Clerget-Darpoux & Elston, 2013). Particularly, polygenic risk scores have been used for disease prognostics, particularly in the case of diabetes (Hahn et al., 2022; Pemmasani et al., 2023). Moreover, ongoing research has also contributed to the literature on genetic studies of estimation and inference of quantities such as genetic correlations (Elgart et al., 2021a) and heritability (Sofer, 2017). This literature provides context and impetus for our research questions and statistical methods.

1.3. *Notations*

We use the following notations in the remainder of the paper. First, we use bold symbols (e.g. $\boldsymbol{u}$) to denote vectors, non-bold capital letters (e.g. W) to denote matrices and non-bold lowercase letters (e.g. e) to denote scalars. For a square matrix $W \in \mathbb{R}^{d \times d}$, we denote $\lambda_{\max}(W) = \lambda_1(W) \geq \lambda_2(W) \geq \ldots \geq \lambda_d(W) = \lambda_{\min}(W)$ as the eigenvalues of W, and subsequently $\text{Tr}(W) := \sum_{i=1}^{d} \lambda_i(W)$ as the trace of the matrix. For a matrix W, we denote the rank of the matrix as $\text{rank}(W)$. The identity matrix in $\mathbb{R}^{d \times d}$ is denoted by I_d and the zero vector in $\mathbb{R}^d$ is denoted as $\mathbf{0}_d$. The symbols $|| \cdot ||$ and $|| \cdot ||_\infty$ denotes the ℓ_2 and ℓ_∞ norm for a vector/matrix, respectively. For two symmetric $d \times d$ matrices U and V, we write $U \succ V$ and $U \succeq V$ when $U - V$ is positive definite and positive semi-definite, respectively.

We utilize the following asymptotic notation in n. We use the standard Bachmann-Landau notations $o(\cdot), O(\cdot)$ for deterministic sequences. For deterministic sequences, $\rightarrow$ denotes the limit of the sequence as n tends to infinity. For two deterministic sequences $\{a_n\}_{n \geq 1}, \{b_n\}_{n \geq 1}$, we say a_n is asymptotic to b_n denoted by $a_n \asymp b_n$, if $\frac{a_n}{b_n} \rightarrow 1$. The symbols $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{d}$ denote convergence in probability and in distribution, respectively. The symbol $\overset{d}{=}$ denotes equivalence in distribution.

For a sequence of random variables $\{Y_n\}_{n \geq 1}$ and a deterministic positive sequence $\{a_n\}_{n \geq 1}$, we write $Y_n = o_{\mathbb{P}}(a_n)$ when $\frac{Y_n}{a_n} \xrightarrow{\mathbb{P}} 0$, and $Y_n = O_p(a_n)$ when $\lim_{K \to \infty} \lim_{n \to \infty} \mathbb{P}\left(\frac{|Y_n|}{a_n} \leq K\right) = 1$.

1.4. *Organization*

The following sections contain details of the results summarized above and are organized as follows. We present the mathematical setup and the technical background required for the analysis in Section 2. Then, we outline the construction of the proposed estimators in Section 3. Subsequently, we present our main theoretical analyses in Section 4. Finally, we validate the performance of estimators through a suite of simulations with synthetic and real-life data, along with a novel data exploration, in Section 5.

2. SETUP

We consider observing data in a *target population* of n individuals as $(y_i, X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$, where $y_i \in \mathbb{R}$ is an outcome of interest, $X_i \in \mathbb{R}^p$ contains baseline covariates. In this setup, we are interested in the precise quantification of the association between X on y under a constrained linear regression model as follows, where

$$
\begin{aligned}
y_i &= X_i^T \beta^* + \epsilon_i, \ \epsilon_i \text{ independent of } X_i, \\
A_j^T \beta^* &= c_j, \ j = 1, \ldots, q \\
\mathbb{E}[\epsilon_i] &= 0, \text{Var}(\epsilon_i) =: \sigma^2 < \infty, \exists \, \delta > 0, \ \mathbb{E}\left[\epsilon_i^{2+\delta}\right] < \infty
\end{aligned} \tag{1}
$$

We include discussions of generalized linear models in the Supplementary Materials. Under this model, β^* will be referred to as *effect vector*, which, in this model, belongs to the affine space defined by known vectors $A_j \in \mathbb{R}^p$ and constants $c_j \in \mathbb{R}$ for $j = 1, \ldots, q$. For subsequent discussions, we shall refer to $X := [X_1 \ldots X_n]^T$ as the *design matrix*, $y := [y_1 \ldots y_n]^T$ the *outcome vector*, $A := [A_1 \ldots A_q]^T$ the *constraint matrix*, and $c := [c_1 \ldots c_q]^T$ the *constraint vector*.

Although we operate under generic known constraints, in our real data applications and motivations, they arise from borrowing information from an external *reference population*. We briefly elaborate on this to explain the nature of the linear constraints that arise in our data setting. Specifically, suppose we have access to a reference population where one has $(Y, y_r, \tilde{X}) \sim \mathbb{P}_r$ with $Y \in \mathbb{R}^q$ contains $q < p$ auxiliary outcomes, y_r is the outcome of interest in the reference population, and $\tilde{X}$ contains baseline covariates. We work under a regression model in the reference population designating $y_r = \tilde{X}^T \beta^* + \tilde{\epsilon}$, $\mathbb{E}[\tilde{\epsilon}_i] = 0$ and $Y_j = \tilde{X}^T \beta_j + \tilde{\epsilon}_j$, $\mathbb{E}[\tilde{\epsilon}_j] = 0$, $j = 1, \ldots, q$, with a common variance-covariance matrix profile of covariates under both the reference and the target population (i.e. $\Sigma := \mathbb{E}\left[\tilde{X}\tilde{X}^T\right] = \mathbb{E}\left[XX^T\right]$). Drawing motivation from genetic association studies, we will connect the target to the reference population by defining constraint matrices and vectors as $A_j = \beta_j^T \Sigma, c_i = A_j \beta^*$, $j = 1, \ldots, q$. The precise reason behind these definitions can be understood from the background of genetic correlation analysis that we provide in the Supplementary Materials. For now, especially for our main data analysis (see Section 5.2 for details), the relevance of such target and reference populations and their implied connections through constraints on parameter space can be roughly described as follows. Suppose we focus on genetic variants known as single-nucleotide polymorphisms (SNPs), particularly those responsible for expression levels of certain proteins, known as protein quantitative trait loci (pQTLs). To draw information on this relationship, suppose we have a reference and target population as above with genotypic and metabolic data, with the reference population additionally possessing
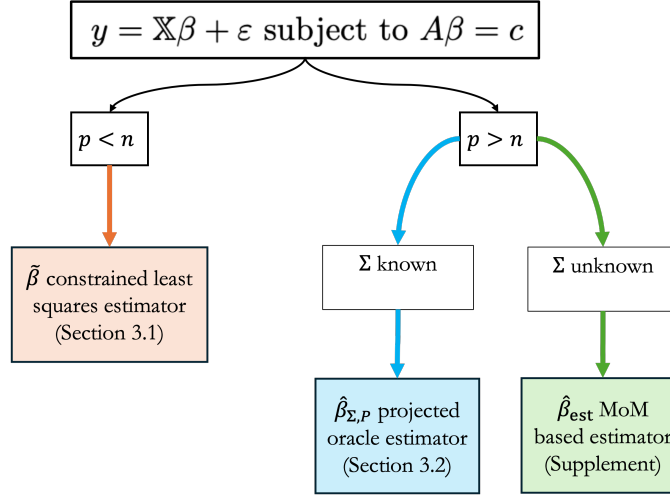
Fig. 1: Schematic describing different estimators described in this paper

protein quantification data. Mathematically, one can consider q proteins for our analysis of p SNPs on a specific metabolic trait in a reference population of individuals with genotype data $\tilde{\mathbf{X}}$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$ the additive allelic effect vectors corresponding to each protein, $\boldsymbol{\beta}^*$ the effect vector for the metabolic trait, $Y_1, \ldots, Y_q$ the protein expression outcomes in the reference population, and $\boldsymbol{y}_r$ the metabolic trait outcomes in the reference population. Similarly, in the target population, suppose $\boldsymbol{y}_t$ are the metabolic trait outcomes in the target population, arising from the same corresponding additive allelic effect vector $\boldsymbol{\beta}^*$. The dispersion matrix Σ is the linkage disequilibrium (LD) matrix, which contains information on the non-random association of alleles at different loci in a population. In this framework, the above described constraints arise naturally from genetic correlation perspectives between Y_j's and y's. We elaborate on this more in subsequent works.

Since the general construction of our methodology will not depend on the precise manner in which the constraints were obtained, we will not refer to the reference population in the theoretical parlance of the paper.

3. METHODS

In this section, we describe the construction of our estimators under scenarios based on different regimes of the dimension p and knowledge of Σ. Throughout we will assume a proportional asymptotic regime where $p/n \to \alpha \in (0, \infty)$ and divide our discussions according to moderately high dimensions ($\alpha < 1$) and high dimensions ($\alpha \geq 1$). In some regimes, we are interested in the optimal estimation of the vector $\boldsymbol{\beta}$, while in others, we provide frameworks for efficient estimation of functionals of $\boldsymbol{\beta}$. We will explicitly specify the targets of estimation in each section.

3.1. *Moderately High Dimension ($\alpha < 1$)*

We can incorporate the constraint $A\boldsymbol{\beta}^* = \boldsymbol{c}$ into ordinary least squares optimization problem whose unique solution exists with high probability under suitable assumptions on X to get an optimal estimator for $\boldsymbol{\beta}$. A feasible solution which will be termed as the *standard projected estimator* is defined as

$$\hat{\boldsymbol{\beta}}_{\mathcal{P}} := \mathcal{P}_{A^\perp}\hat{\boldsymbol{\beta}}_{\mathrm{LS}} + A^T \left(AA^T\right)^{-1} \boldsymbol{c}$$

where $\mathcal{P}_{A^\perp} := \mathcal{P}_\perp(A) = I_p - A^T \left(AA^T\right)^{-1} A$ and $\hat{\beta}_{\text{LS}}$ is the *ordinary least squares* (OLS) estimator of β^*. Here, the least squares solution is projected into a fixed affine space defined by A, agnostic of the data generating mechanism. This estimator can be improved under more general covariance structures on X by utilizing random projections that preserve the constraint. We derive an optimal procedure through the Lagrangian optimization based *constrained least squares* (CLS) estimator (Amemiya, 1985). With the sample estimate of the covariance matrix $\hat{\Sigma}_n = \frac{1}{n} X^T X$, we define the CLS estimator as

$$\tilde{\beta} := C_{A^\perp,inv}\hat{\beta}_{\text{LS}} + \hat{\Sigma}_{n,inv}^{-1} A^T \left(A\hat{\Sigma}_{n,inv}^{-1} A^T\right)^{-1} c$$

where $C_{A^\perp,inv} = I_p - \hat{\Sigma}_{n,inv}^{-1} A^T \left(A\hat{\Sigma}_{n,inv}^{-1} A^T\right)^{-1} A$ and

$$\hat{\Sigma}_{n,inv} = \begin{cases} \hat{\Sigma}_n & \text{if } \hat{\Sigma}_n \text{ is invertible} \\ I_p & \text{otherwise} \end{cases}$$

The matrix $C_{A^\perp,inv}$ is the row space projector on the eigenbasis of $\hat{\Sigma}_{n,inv}$. In our analysis, we consider sub-Gaussian distributions for the independent rows of X, which means that $\hat{\Sigma}_n$ is invertible with probability 1 (Eaton & Perlman, 1973; Vershynin, 2010). In the case where we have discrete data or the probability of invertibility is not large enough for consistent estimation, we will need to be precise in our definition of this estimator. For notational convenience, under the assumptions on the rows of X, we will replace $\hat{\Sigma}_{n,inv}$ with $\hat{\Sigma}_n$ and $C_{A^\perp,inv}$ with $C_{A^\perp}$, implicitly considering invertibility. We study the theoretical and numerical properties of these estimators in Section 4 and Section 5 respectively.

In the case where $q \ll p$ or $q = \mathcal{O}(1)$, which will be relevant to our data application, we do not enjoy the direct benefits provided by the main results of this paper. However, we do gain some benefits in the second-order, which we demonstrate in Sections 4.1 and 5, providing evidence that it is still worthwhile incorporating constraints, even if the number of constraints does not grow with the number of covariates.

We also include a discussion of the construction of estimators for generalized linear models in this moderately high dimensional case in the Supplementary Materials.

3.2. *High Dimension ($\alpha \geq 1$)*

In high dimensions, our approach will depend on the knowledge we possess of the underlying dispersion matrix Σ. This is because in an ill-specified regime, without oracle information about Σ, estimation of any particular coordinate of β^* will be biased. We present a simple estimator that demonstrates desirable properties, leaving the discussion of comparison between estimators to future works. This estimator will utilize Σ as a stand-in for the information matrix in a traditional least squares estimator. Following nomenclature given in Frostig & Heller (2022); Knight & Duan (2024), we define an *oracle* estimator $\hat{\beta}_\Sigma := \frac{1}{n}\Sigma^{-1} X^T y$ as our high-dimensional estimator of β^*, which is the estimator for β that arises from solving a method-of-moments fixed point equation. We define the *projected oracle estimator* as follows

$$\hat{\beta}_{\Sigma,\mathcal{P}} := \frac{1}{n}\mathcal{P}_{A^\perp}\Sigma^{-1} X^T y + A^T \left(AA^T\right)^{-1} c$$

incorporating the constraint information into the estimation procedure. We show in Section 4 that such an estimator attains a notion of asymptotic normality, allowing us to use this estimator for testing purposes.

Under the scenario where we cannot utilize Σ directly, without any additional imposition of structure, the notion of "optimality" is more muddied. The choice of an "optimal" estimator in high dimensions is subjective and depends on problem-specific requirements. For example, projections of regularized estimators for $\boldsymbol{\beta}$ optimize specific objective functions, but suffer from bias in the estimation of coordinates. We relegate the discussion of optimality to future work, and only focus on the estimation of functionals of $\boldsymbol{\beta}$. We do so by constructing estimators of β_j^* incorporating the constraint information using the procedures given in Kong & Valiant (2018) and subsequently Chen et al. (2024). The procedure will involve approximating functions of Σ^{-1} using Chebyshev polynomials (DeVore & Lorentz, 1993), and using higher-order U-statistics to estimate these approximations. We provide this algorithmic procedure in the Supplementary Materials, along with simulations that validate the performance of these estimators.

4. Properties of Estimators

To discuss the theoretical properties of the estimators introduced in Section 3, we will work with the following set of assumptions:

Assumption 1. (Dimensionality and Asymptotics)
$p, q, n \to \infty$ such that $\frac{p}{n} \to \alpha \in [0, \infty)$ (*aspect ratio*) and $\frac{q}{p} \to \gamma \in [0, 1)$ (*constraint ratio*)

Assumption 2. (Random Design)

(a) (i) $X_i \stackrel{d}{=} \Sigma^{1/2} Z_i$, where entries of Z are mean zero, unit variance and have finite $(8 + c)$th moment where $c > 0$ is fixed and arbitrary.

  (ii) $\exists C > 0$ such that $\frac{1}{C} \leq \lambda_k(\Sigma) \leq C$ for $k = 1, \ldots, p$. Additionally, the support of μ_∞ lies within $\left(\frac{1}{C}, C\right)$, where μ_∞ denotes the limiting empirical spectral distribution of Σ.

(b) (i) For any 1-Lipschitz function f of X_i, $\exists C$ such that we have the following concentration inequality:

$$\mathbb{P}(|f(X_i) - \mathbb{E}[f(X_i)]| > t) \leq C \exp\left\{-\frac{t^2}{C}\right\}$$

  (ii) $\exists C' > 0$ such that $\frac{1}{C'} \leq \frac{p}{n} \leq C'$ for all $n > 0$

  (iii) $\exists C'' > 0$ such that $\frac{1}{C''} \leq \lambda_k(\Sigma) \leq C''$ for all $k = 1, \ldots, p$. Additionally, the support of μ_∞ lies within $\left(\frac{1}{C''}, C''\right)$, where μ_∞ denotes the limiting empirical spectral distribution of Σ.

Assumption 3. (Gaussian Error) $y_i \sim \mathcal{N}\left(X_i^T \boldsymbol{\beta}^*, \sigma^2\right)$ for $i = 1, \ldots, n$

Remark: The stipulations on X for Assumption 2(a) are stronger than those for Assumption 2(b), but independent control of each X requires the aspect ratio to be bounded away from 0, which is not required in Assumption 2(a). Also, the assumptions on the limiting empirical spectral distribution of Σ guarantee the existence of subsequent trace variables. Additionally, from Bai & Zhou (2008), we have the convergence of the empirical spectral measure of $\hat{\Sigma}_n$ to a free multiplicative convolution of a Marchenko-Pastur law MP(α) and μ_∞, which is necessary for many subsequent convergence in probability results. These assumptions arise from prior work focused on the construction of deterministic equivalents for sample covariance matrices Hachem et al. (2007); Louart & Couillet (2021); Chouard (2022). In subsequent results, we will only refer

to the random design assumptions as Assumption 2, as either assumption will be sufficient, and we discuss these conditions further in the Supplementary Material.

4.1. *Moderately High Dimension*

For the CLS estimator $\tilde{\beta}$, defined through the matrices $\hat{\Sigma}_n$ and $C_{A^\perp}$, we have the following theorem describing the exact minimax error of estimation under a linear constraint, and the asymptotics of the same error.

THEOREM 1. *For the setup given in (1), under Assumption 3, the minimax risk of estimation of β^* in squared error loss is*

$$\inf_{\hat{\beta}} \sup_{\beta^*} \mathbb{E} \left\| \hat{\beta} - \beta^* \right\|^2 = \frac{\sigma^2}{n} \mathbb{E} \left[\mathrm{Tr} \left(C_{A^\perp} \hat{\Sigma}_n^{-1} \right) \right]$$

where

$$\hat{\Sigma}_n = \frac{X^T X}{n}, C_{A^\perp} = I - \hat{\Sigma}_n^{-1} \left(A^T \left(A \hat{\Sigma}_n^{-1} A^T \right)^{-1} A \right)$$

Additionally, the estimator

$$\tilde{\beta} = C_{A^\perp} \hat{\Sigma}_n^{-1} \left(\frac{X^T y}{n} \right) + \hat{\Sigma}_n^{-1} A^T (A \hat{\Sigma}_n^{-1} A^T)^{-1} c$$

achieves the minimax risk

We would also like to categorize the asymptotics of such an error formulation, which is put forth in the following corollary.

COROLLARY 1. *For the setup given in (1), under Assumptions 1,2,3, the constrained least squares estimator $\tilde{\beta}$ achieves the following asymptotic squared error loss*

$$\lim_{n \to \infty} \mathbb{E} \left\| \tilde{\beta} - \beta^* \right\|^2 = \frac{\sigma^2}{(1 - (1 - \gamma)\alpha)} \lim_{n \to \infty} \frac{1}{n} \mathrm{Tr} \left(\Sigma^{-1} - \Sigma^{-1} \left(A^T \left(A \Sigma^{-1} A^T \right)^{-1} A \right) \Sigma^{-1} \right)$$

Under an isotropic ($\Sigma = I_p$) generation process, this reduces to

$$\lim_{n \to \infty} \mathbb{E} \left\| \tilde{\beta} - \beta^* \right\|^2 = \frac{\sigma^2 (1 - \gamma)\alpha}{1 - (1 - \gamma)\alpha}$$

Remark: We assume the existence of the limit in the first equation. This is a reasonable assumption based on the random design restrictions, and the second statement of the corollary demonstrates it in the isotropic case.

This result describes the precise "difficulty" of recovering the true effect vector. In the context of mean-squared error, we provide a justification for projection, even in the case where $q \ll p$. Let us denote $G_n := \|\hat{\beta}_{LS} - \beta^*\|^2 - \|\hat{\beta}_{\mathcal{P}} - \beta^*\|^2$ as the "gain" in squared error loss of projecting.

PROPOSITION 1. *Under Assumption 1 and 2, with an isotropic ($\Sigma = I_p$) generation process, we have*

$$\mathbb{E}[G_n] = \frac{q\sigma^2}{n(1 - \alpha)} (1 + o(1))$$

Additionally

(a) *For $q \to \infty$, we have*

$$\frac{nG_n}{q} \xrightarrow{\mathbb{P}} \frac{\sigma^2}{1-\alpha}$$

(b) Additionally, under Assumption 3, for $q = O(1)$, we have

$$\frac{nG_n}{\sigma^2} \overset{d}{=} \sum_{i=1}^{q} w_i \chi_1^{(i)}$$

where $\chi_1^{(i)}$ are i.i.d. central chi-squared random variables with degrees of freedom 1 and $w_1 \geq w_2 \geq \ldots \geq w_q$ are the non-zero eigenvalues of $nX\hat{\Sigma}_n^{-1}\mathcal{P}_A\hat{\Sigma}_n^{-1}X^T$.

This proposition, in tandem with Theorem 1, demonstrates the benefit of utilizing even a fixed number of constraints over the OLS, which is relevant to our data example in Section 5.2. We additionally verify this benefit in Section 5. Our next result is about the asymptotic distribution of coordinates of $\tilde{\beta}$ at a parametric rate.

PROPOSITION 2. *Under Assumptions 1 and 2, we have*

$$\sqrt{n}\left(\tilde{\beta}_j - \beta_j^*\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2 s_{C,j}}{1 - (1-\gamma)\alpha}\right)$$

where $s_{C,j} = \lim\limits_{n\to\infty} e_j^T \left(\Sigma^{-1} - \Sigma^{-1}\left(A^T\left(A\Sigma^{-1}A^T\right)^{-1}A\right)\Sigma^{-1}\right)e_j$

Remark: We assume the existence of the limit $s_{C,j}$, similar to the remark on Corollary 1. Also, since $\frac{\sigma^2}{1-\alpha}\Sigma^{-1} \succeq \frac{\sigma^2}{1-(1-\gamma)\alpha}\left(\Sigma^{-1} - \Sigma^{-1}\left(A^T\left(A\Sigma^{-1}A^T\right)^{-1}A\right)\Sigma^{-1}\right)$ for any fixed p, where $\frac{\sigma^2}{1-\alpha}\Sigma^{-1}$ is the variance matrix of the OLS estimator, the total of the variances of each coordinate of the CLS estimator is lower than that of the OLS estimator. We cannot make any direct comparison of the asymptotic variance for a given coordinate.

The asymptotic variance is a function of Σ^{-1}, which requires careful estimation while constructing confidence intervals. We will use a corrected jackknife estimator for the variance, elucidated in (Karoui & Purdom, 2016), for our numerical experiments. We have the following result, which follows from Theorem 4.1 in (Karoui & Purdom, 2016):

PROPOSITION 3. *Under Assumptions 1 and 2, given that every row of X_i is drawn from a multivariate normal distribution with mean $\mathbf{0}$ and variance Σ, we have*

$$\frac{\mathbb{E}\left[\widehat{\mathrm{Var}}_j\right]}{\mathrm{Var}(\tilde{\beta}_j)} \to \frac{1}{(1 - (1-\gamma)\alpha)}$$

where $\widehat{\mathrm{Var}}_j = \frac{n-1}{n}\sum_{i=1}^{n}\left(\tilde{\beta}_{(i),j} - \tilde{\beta}_j\right)^2$, $\tilde{\beta}_{(i)}, \tilde{\beta}$ are the CLS estimators with the ith observation removed and for the whole dataset respectively. This holds for any contrast $v^T\tilde{\beta}$ with $||v||_2^2 = 1$

Using Propositions 1 and 2, we can conduct inference on any constrast of our regression vector in this proportional regime.

4.2. *High Dimension*

We demonstrate that the projected oracle estimator has a notion of asymptotic normality, akin to the moderately-high dimension estimator, allowing for inference in this regression problem in high dimensions.

PROPOSITION 4. *Given the setup in (1), under Assumptions 1,2,3, we have*

$$\sqrt{n} \left(\hat{\beta}_{\Sigma,\mathcal{P},j} - \beta_j^* \right) \xrightarrow{d} \mathcal{N} \left(0, \left(\beta_j^* \right)^2 + \left(\sigma^2 + \ell_{\beta^*,\Sigma} \right) s_{\mathcal{P},j} \right)$$

where $\ell_{\beta^*,\Sigma} = \lim\limits_{n \to \infty} ||\Sigma^{1/2} \beta^*||^2$ *and* $s_{\mathcal{P},j} = \lim\limits_{n \to \infty} e_j^T \left(\mathcal{P}_{A^\perp} \Sigma^{-1} \mathcal{P}_{A^\perp} \right) e_j$

Remark: We assume the existence of the limit $s_{\mathcal{P},j}$, similar to Proposition 2. Also, since $\Sigma^{-1} \succeq \mathcal{P}_{A^\perp} \Sigma^{-1} \mathcal{P}_{A^\perp}$ for any projection matrix $\mathcal{P}_{A^\perp}$ in fixed p, the total variance of each coordinate of this projected oracle estimator is less than that of a standard oracle estimator.

Note that the asymptotic variance is a function of the signal strength and the magnitude of the regression coordinate. Additional information on this signal strength along with variance stabilization will allow us to invert the variance form to construct a confidence interval for β_j^*. We demonstrate the efficacy of this projected oracle estimator in utilizing the constraints over a range of constraint ratios in Section 5.

5.   EXPERIMENTAL EFFICACY

We look to evaluate the performance of our estimators through a set of simulations. We will first demonstrate the properties detailed in the previous section through simulations using synthetic data, and we will apply the proposed estimators to genetic datasets for the purpose of genetic variant discovery.

5.1.   *Synthetic Simulations*

For the purposes of our simulations, we refer to the formulation of the reference population from Section 2. We use a reference population to construct our constraints, which we use for our target population, and all subsequent analyses are conditional on this reference population. We fix our sample size $n = 200$ and $N = 1000$, with 1000 iterations for repeated experiments, and vary the aspect and constraint ratio as described below.

(**s1**) $p = (10, 20, \ldots, 190), q = \lfloor \frac{p}{2} \rfloor + 1$ (Moderately high dimension, varying aspect ratio)
(**s2**) $p = 100, q = (0, 5, \ldots, 100)$ (Moderately high dimension, varying constraint ratio)
(**s3**) $p = 300, q = (0, 10, \ldots, 300)$ (High dimension, varying constraint ratio)

Additionally, we consider two cases of the dispersion matrix; (**m1**) Σ isotropic, generated as $\Sigma = I_p$, (**m2**) Σ anisotropic, generated as $\Sigma = 0.5 I_p + 0.5 J_p$. Moreover, we generate $X_i \sim \mathcal{N}_p(0, \Sigma)$ for i in $1, \ldots, n$ and $\tilde{X}_{\tilde{i}} \sim \mathcal{N}_p(0, \Sigma)$ for $\tilde{i}$ in $1, \ldots, N$, and the sample covariance matrices $\hat{\Sigma}_n = \frac{X^T X}{n}$ and $\tilde{\Sigma}_N = \frac{\tilde{X}^T \tilde{X}}{N}$. We use $B = [I_q : 0_{p-q}], \beta_j^* \sim \mathcal{N}(5, 5)$ for j in $1, \ldots, p$, and define $A = B \hat{\Sigma}_N$ and $c = A \beta^*$. Finally, for the outcome, we consider $y = X \beta^* + \epsilon, \epsilon_i \sim \mathcal{N}(0, 1)$ for i in $1, \ldots, n$. Below we highlight a few results surrounding these setups and defer additional results to the Supplementary Material.

Fig. 2 [**s2,m2**] shows comparisons of mean squared error (MSE) between the OLS estimator, a standard projection estimator, and the CLS estimator, described in the moderately high dimension part of Section 3. The error of the OLS estimator $\hat{\beta}_{LS}$ is constant since it is constraint-agnostic, while the other estimators have errors that tend to zero as the constraint level increases. As verified theoretically, the CLS estimator $\tilde{\beta}$ has uniformly lower error than the projected estimator $\hat{\beta}_{\mathcal{P}}$. Next, Fig. 3 [**s2,m1**] shows the out-of-sample prediction error for a subset of values for q, on 100 new observations generated under a similar setup. We see that as the constraint ratio increases,
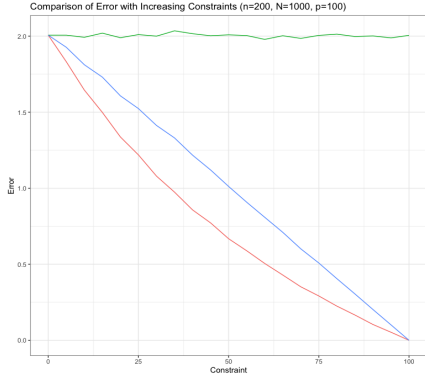
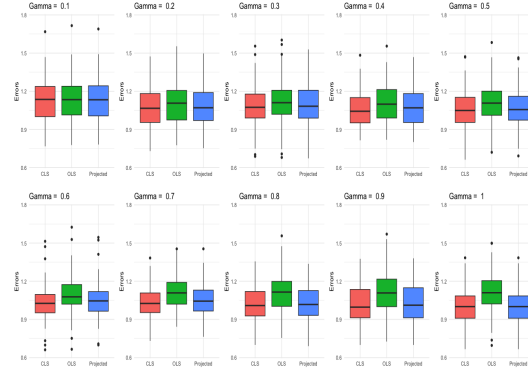Fig. 2: Error comparison $(\hat{\boldsymbol{\beta}}_{LS}, \hat{\boldsymbol{\beta}}_{\mathcal{P}}, \tilde{\boldsymbol{\beta}})$ [s2,m2]

Fig. 3: Prediction error $(\hat{\boldsymbol{\beta}}_{LS}, \hat{\boldsymbol{\beta}}_{\mathcal{P}}, \tilde{\boldsymbol{\beta}})$ [s2,m1]
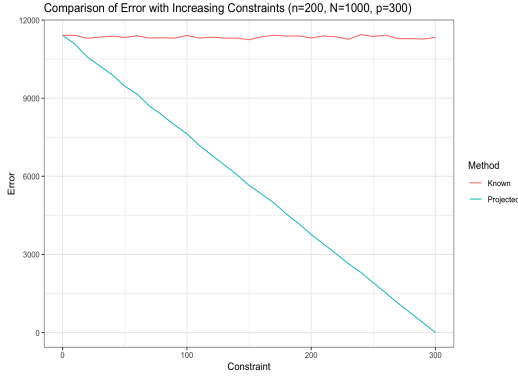
Fig. 4: Error comparison $(\hat{\boldsymbol{\beta}}_{\Sigma}$ and $\hat{\boldsymbol{\beta}}_{\Sigma,\mathcal{P}})$ [s3,m1]
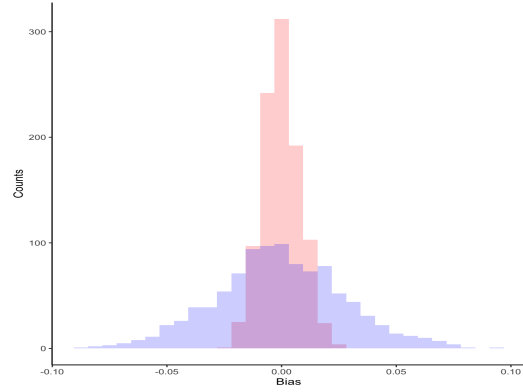
Fig. 5: Fluctuations of error for a coordinate of the CLS (red) and OLS (blue) estimators with $n = 3000, p = 300, q = 10$ over 1000 iterations

the predictive errors of the projected and CLS estimators decrease, while the predictive error of the OLS estimator remains steady.

Fig. 4 [s3,m1] shows a similar conclusion as Fig. 2 but in a high dimensional case. Here, the comparison is between the oracle estimator $\hat{\boldsymbol{\beta}}_{\Sigma}$ and the projected oracle estimator $\hat{\boldsymbol{\beta}}_{\Sigma,\mathcal{P}}$ described in the high dimension part of Section 3.

Fig. 5 shows the fluctuations of a coordinate of the CLS and OLS estimators around the true value of the regression vector at this coordinate over 1000 iterations, for a dimension specification that is similar to our data. We see that the fluctuations are lesser for the CLS estimator, which means that there are still benefits to incorporating constraints in the analysis even if the number of constraints is much smaller than the number of covariates.

A suite of data-informed simulations, using the genetic datasets described in the Section 5.2, are available in the Supplementary Materials.
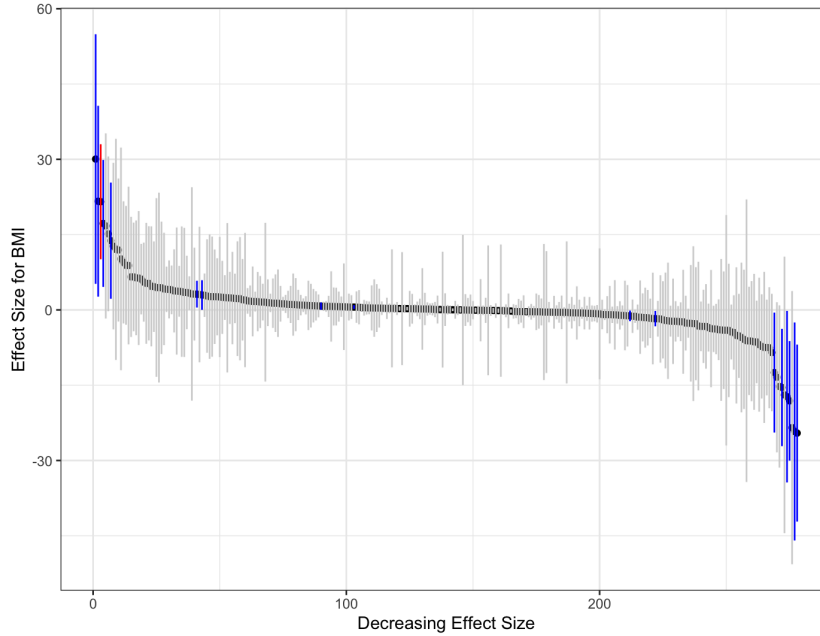
Fig. 6: Estimated coefficients ordered by effect size and 95% CIs of genetic variants in association with BMI, using $\tilde{\beta}$. Blue and red CIs are significant before and after controlling for the family-wise error rate

5.2. *Genetic Datasets*

The remaining numerical analyses will revolve around the Jackson Heart Study (JHS) (Sempos et al., 1999) and the Multi-Ethnic Study of Atherosclerosis (MESA) (Bild, 2002) datasets. In what follows, we explain how we used the JHS dataset as our reference population, to extract information for building constraints, and the MESA dataset as our target population, to perform association analysis under these constraints.

The Jackson Heart Study (JHS) is a study with the purpose of establishing a single-site cohort study to identify the risk factors for cardiovascular disease in African American men and women in Mississippi. The total sample size is 5306 individuals, with a subsample of 2050 individuals ($N = 2050$) with protein expression data, alongside metabolic traits including body-mass index (BMI), fasting insulin, and glycated hemoglobin (HbA1c). For each metabolic trait, we selected 10 proteins that are both highly heritable and have high genetic correlations to the relevant trait to define constraints (Elgart et al., 2021b; Tsai et al., 2023). For each trait, we performed GWAS of each of the proteins matched to this trait and identified genetic associations, i.e., single nucleotide polymorphisms (SNPs) associated with the proteins and their estimated effect sizes. The heritability and genetic correlations of each of these proteins, along with the processes of protein selection and selection of SNPs are detailed in the Supplementary Materials.

The Multi-Ethnic Study of Atherosclerosis (MESA) is a longitudinal cohort study established with the objective of investigating the prevalence, correlates and progression of subclinical cardiovascular disease in an ethnically diverse population of over 6500 men and women in the United States. We used a sample of 3280 MESA individuals ($n = 3280$) who participated in MESA Exam 5, and had available metabolic traits (BMI, fasting insulin, HbA1c).

BMI ($p = 278$)	Fasting Insulin ($p = 1080$)	HbA1c ($p = 539$)
rs5510 (KALLISTATIN)	rs2006232 (VITRONECTIN)	rs3735169 (TIG_2)
	rs2277668 (VITRONECTIN)	rs3767283 (CNTN_2)
	rs3218911 (IL_1_SRII)	rs3901740 (CNTN_2)
	rs6740281 (IL_1_SRII)	rs7137327 (CATF)
	rs7210719 (VITRONECTIN)	rs11240346 (CNTN_2)
		rs77920745 (CATF)
		rs80228806 (CATF)

Table 1: Significant associated genetic variants for each metabolic trait (with the associated protein) recovered after multiplicity corrections, along with number of genetic variants that reached genome-wide level of significance (p)

We applied the estimator from Section 3.1 to the MESA dataset, using the constraints derived from the JHS dataset, for BMI, fasting insulin and HbA1c. We explain the procedure focusing on BMI and show the results for all the traits. For BMI, after selecting protein-associated SNPs, there were $p = 278$ genetic variants to consider. We construct the CLS estimator, utilizing the constraint information and covariance matrix from the JHS population, and used this estimator to make confidence intervals using the distribution described in Proposition 2. Fig. 6 shows the coefficients of the CLS estimator, along with the confidence intervals. For each estimated coefficient, we calculated a p-value based on Proposition 2 and used the Holm-Bonferroni correction method for multiple hypothesis testing, controlling for the family-wise error rate at the 0.05 level (Holm, 1979). We rejected the null hypothesis of no association with BMI for one genetic variant (rs5510), which is associated with the Kallistatin protein. We apply a similar procedure for fasting insulin and HbA1c. For fasting insulin, out of $p = 1080$ genetic variants, we identified 5 genetic variants (rs2006232, rs6740281, rs7210719, rs2277668, rs3218911), and for HbA1c, we identified 7 of the $p = 539$ considered genetic variants (rs11240346, rs3901740, rs3767283, rs7137327, rs77920745, rs80228806, rs80228806). Table 1 lists the associated genetic variants for each metabolic trait, and the protein each variant was associated with in the JHS protein GWAS.

The selected genetic variant for BMI belongs to the gene SERPINA4 (potentially controlling gene expression (Edwards et al., 2012)), which produces the protein Kallistatin, which is responsible for activity of the adrenal gland (Wang et al., 1996). This gland is responsible for steroid hormones, and thus, would have an effect on BMI (Gateva et al., 2017). In fact, rs5510 is associated with the trait "protein levels in obesity" in the GWAS Catalog (Cerezo et al., 2025). For fasting insulin, the genetic variants belong to SARM1, MAP4K4, TMEM199, SEBOX and IL1R2 genes, all genes responsible for proteins for cellular function, which could have a potential downstream effect on pancreatic tissue. For HbA1c, the genetic variants belong to CNTN2, CHPT1 and PC genes, which are all responsible for proteins related to blood cell function. This could also have a potential downstream effect on the level of glycation of the hemoglobin. While these might be feasible etiological pathways, the variants isolated are not strongly associated with the given traits, which could be the result of relatively weaker signal in the genetic data. Additionally, there are intrinsically two levels of correction for multiplicity, one at the genome-wide level and one at the constrained level, so this procedure acts conservatively for isolating variants.

6. Discussion

In this paper, we have developed methods for constrained linear regression, that arise when incorporating protein expression data into the pipeline of genetic effect estimation. We have shown the optimality of our methods through theory and simulations in moderately high dimensional cases, and provided $\sqrt{n}$ consistent estimators in the high dimensional case. These estimators were implemented on genetic datasets for the purpose of genetic discovery.

There are many natural extensions and pressing points to address with this work. In the high dimensional case, when Σ is known, a oft-used framework for estimation is through the use of convex regularization and debiasing. Indeed, under the assumptions of additional structure, such as sparsity, a regularized estimator would have better performance, and knowledge of the dispersion matrix allows the construction of an estimator which minimizes asymptotic variance (Bellec & Zhang (2021b), Bellec & Zhang (2021a)). It can be shown that the methodology of these papers can be easily extended to the framework of constraints, warranting comparative studies. In the case where Σ is not known, works such as Kong & Valiant (2018); Li & Sur (2023); Chen et al. (2024) take initial steps to the construction of potentially useful estimators, and there is scope for considering the impact of these works in a constrained setting.

In terms of direct next steps, working towards a proper characterization of estimated constraint estimators and their errors would allow for accurate implementation of these estimators on real datasets. Additionally, there is a direct connection to problems in the transfer learning literature, and utilizing the tools of this paper to investigate problems of "growing" numbers of sources could be quite elucidating, in the age of online data. We keep these goals for future research directions.

Data Availability Statement

Acknowledgements

Supplementary material

The Supplementary Material includes the following:

1. Proofs of the main results, corollaries, propositions and lemmas from the main body of the paper
2. Additional details of methodological and applicative procedures and their interpretations
3. Additional experimental results, including a suite of data-informed simulations

References

Aitchison, J. & Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics* **29**, 813–828.

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Bai, Z. & Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica* **18**, 425–442.

Bellec, P. C. & Zhang, C.-H. (2021a). De-biasing convex regularized estimators and interval estimation in linear models ArXiv:1912.11943 [math, stat].

Bellec, P. C. & Zhang, C.-H. (2021b). De-biasing the lasso with degrees-of-freedom adjustment ArXiv:1902.08885 [math, stat].

Bild, D. E. (2002). Multi-ethnic study of atherosclerosis: Objectives and design. *American Journal of Epidemiology* **156**, 871–881.

Cedric Gondro, Julius van der Werf, B. H. (2013). *Genome-Wide Association Studies and Genomic Prediction*, vol. 1019 of *Methods in Molecular Biology*. Totowa, NJ: Humana Press.

Cerezo, M., Sollis, E., Ji, Y., Lewis, E., Abid, A., Bircan, K., Hall, P., Hayhurst, J., John, S., Mosaku, A., Ramachandran, S., Foreman, A., Ibrahim, A., McLaughlin, J., Pendlington, Z., Stefancsik, R., Lambert, S. A., McMahon, A., Morales, J., Keane, T., Inouye, M., Parkinson, H. & Harris, L. W. (2025). The nhgri-ebi gwas catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research* **53**, D998–D1005.

Chen, X., Liu, L. & Mukherjee, R. (2024). Method-of-moments inference for glms and doubly robust functionals under proportional asymptotics ArXiv:2408.06103 [econ, math, stat].

Cheng, C. & Montanari, A. (2024). Dimension free ridge regression ArXiv:2210.08571 [math].

Chouard, C. (2022). Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure ArXiv:2211.13044 [math].

Clerget-Darpoux, F. & Elston, R. C. (2013). Will formal genetics become dispensable? *Human Heredity* **76**, 47–52.

Cole, J. B. & Florez, J. C. (2020). Genetics of diabetes and diabetes complications. *Nature reviews. Nephrology* **16**, 377–390.

DeVore, R. A. & Lorentz, G. G. (1993). *Constructive Approximation*. Springer Science Business Media. Google-Books-ID: cDqNW6k7˙ZwC.

Dobriban, E. & Sheng, Y. (2022). Distributed linear regression by averaging ArXiv:1810.00412 [math, stat].

Dobriban, E. & Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics* **46**, 247–279.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics* **22**, 238–270.

EATON, M. L. & PERLMAN, M. D. (1973). The non-singularity of generalized sample covariance matrices. *The Annals of Statistics* **1**, 710–717.

EDWARDS, N. C., HING, Z. A., PERRY, A., BLAISDELL, A., KOPELMAN, D. B., FATHKE, R., PLUM, W., NEWELL, J., ALLEN, C. E., S., G., SHAPIRO, A., OKUNJI, C., KOSTI, I., SHOMRON, N., GRIGORYAN, V., PRZYTYCKA, T. M., SAUNA, Z. E., SALARI, R., MANDEL-GUTFREUND, Y., KOMAR, A. A. & KIMCHI-SARFATY, C. (2012). Characterization of coding synonymous and non-synonymous variants in adamts13 using ex vivo and in silico approaches. *PLoS ONE* **7**, e38864.

ELGART, M., GOODMAN, M. O., ISASI, C., CHEN, H., DE VRIES, P. S., XU, H., MANICHAIKUL, A. W., GUO, X., FRANCESCHINI, N., PSATY, B. M., RICH, S. S., ROTTER, J. I., LLOYD-JONES, D. M., FORNAGE, M., CORREA, A., HEARD-COSTA, N. L., VASAN, R. S., HERNANDEZ, R., KAPLAN, R. C., REDLINE, S., THE TRANS-OMICS FOR PRECISION MEDICINE (TOPMED) CONSORTIUM & SOFER, T. (2021a). *Genetic and environmental correlations between complex phenotypes differ by race/ethnicity and sex.*

ELGART, M., GOODMAN, M. O., ISASI, C., CHEN, H., DE VRIES, P. S., XU, H., MANICHAIKUL, A. W., GUO, X., FRANCESCHINI, N., PSATY, B. M., RICH, S. S., ROTTER, J. I., LLOYD-JONES, D. M., FORNAGE, M., CORREA, A., HEARD-COSTA, N. L., VASAN, R. S., HERNANDEZ, R., KAPLAN, R. C., REDLINE, S., THE TRANS-OMICS FOR PRECISION MEDICINE (TOPMED) CONSORTIUM & SOFER, T. (2021b). *Genetic and environmental correlations between complex phenotypes differ by race/ethnicity and sex.*

FROSTIG, T. & HELLER, R. (2022). Inferring on joint associations from marginal associations and a reference sample ArXiv:2212.01809 [stat].

GATEVA, A., ASSYOV, Y., VELIKOVA, T. & KAMENOV, Z. (2017). Increased kallistatin levels in patients with obesity and prediabetes compared to normal glucose tolerance. *Endocrine Research* **42**, 163–168.

GYÖRFI, L. (2002). *A distribution free theory of nonparametric regression*. Springer.

HACHEM, W., LOUBATON, P. & NAJIM, J. (2007). Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability* **17**. ArXiv:math/0507172.

HAHN, S.-J., KIM, S., CHOI, Y. S., LEE, J. & KANG, J. (2022). Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: A machine learning analysis of population-based 10-year prospective cohort study. *eBioMedicine* **86**.

HAN, Q. (2022). Noisy linear inverse problems under convex constraints: Exact risk asymptotics in high dimensions ArXiv:2201.08435 [cs, math, stat].

HE, B., SHI, J., WANG, X., JIANG, H. & ZHU, H.-J. (2020). Genome-wide pqtl analysis of protein expression regulatory networks in the human liver. *BMC Biology* **18**, 97.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

HOLTZMAN, N. A. & MARTEAU, T. M. (2000). Will genetics revolutionize medicine? *The New England Journal of Medicine* **343**, 141–144.

HSU, D., KAKADE, S. M. & ZHANG, T. (2014). Random design analysis of ridge regression ArXiv:1106.2363 [math].

KAROUI, N. E. & PURDOM, E. (2016). Can we trust the bootstrap in high-dimension? ArXiv:1608.00696 [math, stat].

KHOURY, M. J., NEWILL, C. A. & CHASE, G. A. (1985). Epidemiologic evaluation of screening for risk factors: application to genetic screening. *American Journal of Public Health* **75**, 1204–1208.

KLAASSEN, C. A. J. & SUSYANTO, N. (2016). Semiparametrically efficient estimation of euclidean parameters under equality constraints ArXiv:1606.07749 [math, stat].

KNIGHT, P. & DUAN, R. (2024). Multi-task learning with summary statistics. *Advances in Neural Information Processing Systems* **36**.

KONG, W. & VALIANT, G. (2018). Estimating learnability in the sublinear data regime.

LI, S., CAI, T. T. & LI, H. (2022a). Estimation and inference with proxy data and its genetic applications ArXiv:2201.03727 [math, stat].

LI, S., CAI, T. T. & LI, H. (2022b). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**, 149–173.

LI, Y. & SUR, P. (2023). Spectrum-aware debiasing: A modern inference framework with applications to principal components regression.

LOUART, C. & COUILLET, R. (2021). Spectral properties of sample covariance matrices arising from random matrices with independent non identically distributed columns.

MOURTADA, J. (2022). Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics* **50**. ArXiv:1912.10754 [math, stat].

PATHAK, R., WAINWRIGHT, M. J. & XIAO, L. (2023). Noisy recovery from random linear observations: Sharp minimax rates under elliptical constraints ArXiv:2303.12613 [math].

PEMMASANI, S. K., ATMAKURI, S. & ACHARYA, A. (2023). Genome-wide polygenic risk score for type 2 diabetes in indian population. *Scientific Reports* **13**, 11568.

SEMPOS, C. T., BILD, D. E. & MANOLIO, T. A. (1999). Overview of the jackson heart study: A study of cardiovascular diseases in african american men and women. *The American Journal of the Medical Sciences* **317**, 142–146.

SHAPIRO, A. (1989). Asymptotic properties of statistical estimators in stochastic programming. *The Annals of Statistics* **17**, 841–858.

SHI, P., ZHANG, A. & LI, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics* **10**, 1019–1040.

SOFER, T. (2017). Confidence intervals for heritability via haseman-elston regression. *Statistical applications in genetics and molecular biology* **16**, 259.

TSAI, Y.-T., HRYTSENKO, Y., ELGART, M., TAHIR, U., CHEN, Z.-Z., WILSON, J. G., GERSZTEN, R. & SOFER, T. (2023). A parametric bootstrap approach for computing confidence intervals for genetic correlations with application to genetically-determined protein-protein networks. *medRxiv: The Preprint Server for Health Sciences* , 2023.10.24.23297474.

TSYBAKOV, A. B. (2009). *Introduction to nonparametric estimation*. Springer-Verlag New York: Springer e-books.

VAART, A. W. V. D. (1998). *Asymptotic statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge, UK; New York, NY, USA: Cambridge University Press.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

WANG, D. Z., SONG, Q., CHEN, L. M., CHAO, L. & CHAO, J. (1996). Expression and cellular localization of tissue kallikrein-kinin system in human adrenal gland. *The American Journal of Physiology* **271**, F709–716.

YU, M. (2020). *Estimation and statistical inference for high dimensional model with constrained parameter space*. Phd thesis, The University of Chicago.

YU, M., GUPTA, V. & KOLAR, M. (2019). Constrained high dimensional statistical inference ArXiv:1911.07319 [stat].

Supplementary material for Asymptotic Inference for Constrained Regression

1. Proof of Results

Proof of Theorem 1

We structure the proof into the following steps:

- Reducing the constrained problem to an equivalent unconstrained problem
- Deducing the minimax risk of the unconstrained problem
- Showing the CLS estimator $\tilde{\beta}$ matches the minimax risk

We will denote $\mathcal{N}(A)$ as the null space of A, and $\hat{\Sigma}_n = \frac{X^T X}{n}$. The loss function we will be dealing with is the squared error loss. Firstly, we will consider the constraint vector c to be equivalently equal to 0. Subsequently, we will argue that the value of this vector does not affect the minimax error.

Consider β^* such that $A\beta^* = 0$. Clearly $\beta^* \in \mathcal{N}(A)$, which is a subspace. Let $V \in \mathbb{R}^{p \times (p-q)}$ be an orthonormal basis of $\mathcal{N}(A)$. This assures the existence of a unique $\alpha^* \in \mathbb{R}^{p-q}$ such that $\beta^* = V\alpha^*$. The mapping between β^* and α^* is bijective, and α^* is unconstrained in $\mathbb{R}^{p-q}$. Now, the constraint in our minimization is equivalent to the existence of such an α^*. Thus, we can write the following:

$$\min_{\beta \in \mathcal{N}(A)} ||y - X\beta||^2 = \min_{\alpha \in \mathbb{R}^{p-q}} ||y - XV\alpha||^2$$

The second minimization is an ordinary least squares with design matrix XV.

Now, we can deduce the minimax risk of this unconstrained problem. We have the following lemma that follows from Mourtada (2022) and the analysis of the minimax risk for the OLS estimator

LEMMA 1. *For the setup in (1) with $A\beta^* = 0$ and $\beta^* = V\alpha^*$, where V is an orthonormal basis of $\mathcal{N}(A)$, we have*

$$\inf_{\hat{\alpha}} \sup_{\alpha^*} \mathbb{E} ||\hat{\alpha} - \alpha^*||^2 = \frac{\sigma^2}{n} \mathbb{E}\left[\text{Tr} \left(V^T \hat{\Sigma}_n V \right)^{-1} \right]$$

Now, let us reintroduce the constraint vector c. This converts the subspace $\mathcal{N}(A)$ into an affine space. Thus, if β satisfies $A\beta = c$, then there exists u_N, u_c such that $Au_N = 0, Au_c = c$ and $\beta = u_N + u_c$. If we limit u_c to be of a specific form $A^\dagger c$, where $AA^\dagger = I_q$, then we have a bijective correspondence between β and u_N.

$$\mathbb{E} \left\| \hat{\beta} - \beta^* \right\|^2 = \mathbb{E} \left\| (\hat{\beta} - u_c) - (\beta^* - u_c) \right\|^2 = \mathbb{E} ||V(\hat{\alpha} - \alpha^*)||^2$$

Thus, we have the same error as the case where we have a degenerate constraint vector. This, along with Lemma 1, establishes the error optimality.

We will now show that the CLS estimator achieves this minimax risk. Going back to the original optimization problem, we would like to find the estimator

$$\tilde{\beta} = \arg\min_{\beta : A\beta = c} ||y - X\beta||^2 \tag{1}$$

The following lemma shows the form of the CLS estimator given in Amemiya (1985) using the Lagrangian operator.

LEMMA 2. *Given the optimization problem in (1) with the constraint of $A\beta = c$, the solution is the estimator given by*

$$\tilde{\beta} = C_{A^\perp} \hat{\Sigma}_n^{-1} \left(\frac{X^T y}{n} \right) + \hat{\Sigma}_n^{-1} A^T (A\hat{\Sigma}_n^{-1} A^T)^{-1} c$$

where $C_{A^\perp} = I - \hat{\Sigma}_n^{-1} \left(A^T \left(A\hat{\Sigma}_n^{-1} A^T \right)^{-1} A \right)$

We can evaluate the error of this estimator directly. Without loss of generality, we assume $c = 0$, which reduces the form to $\tilde{\beta} = C_{A^\perp} \hat{\beta}_{LS}$, where $\hat{\beta}_{LS}$ is the standard OLS estimator.

$$
\begin{aligned}
\mathbb{E}\left\|\tilde{\beta} - \beta^*\right\|^2 &= \mathbb{E}\left\|C_{A^\perp}\hat{\beta}_{LS} - \beta^*\right\|^2 \\
&= \mathbb{E}\left\|C_{A^\perp}\hat{\beta}_{LS} - C_{A^\perp}\beta^*\right\|^2 \\
&= \mathbb{E}\left\|C_{A^\perp}\left(\hat{\Sigma}_n^{-1}\frac{1}{n}X^T\epsilon\right)\right\|^2 \\
&= \frac{1}{n^2}\mathbb{E}\left\|\sum_{i=1}^n C_{A^\perp}\hat{\Sigma}_n^{-1}X_i\epsilon_i\right\|^2 \\
&= \frac{1}{n^2}\mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i=1}^n C_{A^\perp}C_{A^\perp}^T\hat{\Sigma}_n^{-1}X_i\epsilon_i\right\|^2 \mid X\right]\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n \mathbb{E}\left[X_i^T\hat{\Sigma}_n^{-1}C_{A^\perp}C_{A^\perp}^T\hat{\Sigma}_n^{-1}X_i\epsilon_i^2 \mid X\right]\right] \\
&= \frac{\sigma^2}{n^2}\mathbb{E}\left[\sum_{i=1}^n X_i^T\hat{\Sigma}_n^{-1}C_{A^\perp}\hat{\Sigma}_n^{-1}X_i\right] \\
&= \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{Tr}\left(\hat{\Sigma}_n^{-1}C_{A^\perp}C_{A^\perp}^T\right)\right] \\
&= \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{Tr}\left(C_{A^\perp}\hat{\Sigma}_n^{-1}\right)\right]
\end{aligned}
$$

Now, if we show that this error is equivalent to the error given in the minimax derivation, then the estimator $\tilde{\beta}$ achieves the minimax error. We utilize the following lemma for this equivalence of errors

LEMMA 3. *Given matrices $A \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{d \times n}$ such that $BA = I_d$. Then $X = AB$ uniquely satisfies the following properties:*

- $X \in \mathbb{R}^{n \times n}$

- rank$(X) = d$
- $XA = A$
- $BX = B$

Given $AV = 0$ and Lemma 3, we have the following equations:

$$C_{A^\perp} V = V$$
$$\implies C_{A^\perp} V \left(V^T \hat{\Sigma}_n V\right)^{-1} = V \left(V^T \hat{\Sigma}_n V\right)^{-1}$$
$$V^T \hat{\Sigma}_n C_{A^\perp} = V^T \hat{\Sigma}_n$$

Also, we have $V^T \hat{\Sigma}_n V \left(V^T \hat{\Sigma}_n V\right)^{-1} = I_d$. Therefore, by the claim, we have

$$C_{A^\perp} = V \left(V^T \hat{\Sigma}_n V\right)^{-1} V^T \hat{\Sigma}_n$$
$$C_{A^\perp} \hat{\Sigma}_n^{-1} = V \left(V^T \hat{\Sigma}_n V\right)^{-1} V^T$$
$$\mathrm{Tr}\left(C_{A^\perp} \hat{\Sigma}_n^{-1}\right) = \mathrm{Tr}\left(V \left(V^T \hat{\Sigma}_n V\right)^{-1} V^T\right)$$
$$= \mathrm{Tr}\left(V^T \hat{\Sigma}_n V\right)^{-1}$$

This proves the equivalence of the errors from the previous two parts, which implies the theorem.

To prove Lemma 3, we have the following line of argument

$$
\begin{aligned}
\mathrm{rank}(X) &= d \\
BA = I_d &\implies \mathrm{rank}(A) = d \\
&\implies X, A \text{ have the same column space} \\
\exists Y \in \mathbb{R}^{d \times n} \quad s.t. \quad & X = AY \\
\implies B = BX &= BAY = Y \\
\implies X &= AB
\end{aligned}
$$

1.1. *Minimax Optimality of CLS estimator (Proof of Lemma A1)*

We consider $A\boldsymbol{\beta}^* = 0$ and $\boldsymbol{\beta}^* = V\boldsymbol{\alpha}^*$.

We denote the ordinary least squares estimator for $\boldsymbol{\alpha}^*$ as $\hat{\alpha}_{\mathrm{LS}}$. This estimator provides an upper bound of the minimax error, which can be evaluated directly.

$$\mathbb{E}\,||\hat{\alpha}_{\mathrm{LS}} - \alpha^*||^2 = \mathbb{E}\left\|\left(V^T X^T X V\right)^{-1} V^T X^T \epsilon\right\|^2$$

$$= \mathbb{E}\left\|\left(V^T \hat{\Sigma}_n V\right)^{-1} \frac{1}{n} V^T X^T \epsilon\right\|^2$$

$$= \frac{1}{n^2} \mathbb{E}\left\|\sum_{i=1}^n \left(V^T \hat{\Sigma}_n V\right)^{-1} (XV)_i \, \epsilon_i\right\|^2$$

$$= \frac{1}{n^2} \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{i=1}^n \left(V^T \hat{\Sigma}_n V\right)^{-1} (XV)_i \, \epsilon_i\right\|^2 \,\Big|\, X\right]\right]$$

$$= \frac{1}{n^2} \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}\left[X_i^T V \left(V^T \hat{\Sigma}_n V\right)^{-2} V^T X_i \epsilon_i^2 | X\right]\right]$$

$$= \frac{\sigma^2}{n^2} \mathbb{E}\left[\sum_{i=1}^n X_i^T V \left(V^T \hat{\Sigma}_n V\right)^{-2} V^T X_i\right]$$

$$= \frac{\sigma^2}{n} \mathbb{E}\left[\mathrm{Tr}\left(V^T \hat{\Sigma}_n V\right)^{-1}\right]$$

Hence, the minimax risk of estimation of α^* is given by this quantity.

$$\inf_{\hat{\alpha}} \sup_{\alpha^*} \mathbb{E}\,||\hat{\alpha} - \alpha^*||^2 \le \sup_{\alpha^*} \mathbb{E}\,||\hat{\alpha}_{\mathrm{LS}} - \alpha^*||^2 = \frac{\sigma^2}{n} \mathbb{E}\left[\mathrm{Tr}\left(V^T \hat{\Sigma}_n V\right)^{-1}\right]$$

We will construct a lower bound for the minimax error considering the decision-theoretic problem of estimating α^* under the squared error loss function, following the procedure given in Mourtada (2022). Note that using the fact that $\alpha^* \in \mathbb{R}^{p-q}$, we can think of the decision theoretic problem as unconstrained.

We can consider a diffuse normal prior distribution on α^*, that is $\alpha^* \sim \Pi_\lambda$ where $\Pi_\lambda = \mathcal{N}_{p-q}\left(0, \sigma^2/\lambda n\right)$. The posterior for α^* will therefore follow the distribution $\Pi_\lambda\left(\cdot \mid X, y\right) = \mathcal{N}_{p-q}\left(\hat{\alpha}_{\lambda,n}, (\sigma^2/n)\left(V^T \hat{\Sigma}_n V + \lambda I_{p-q}\right)\right)$, where $\hat{\alpha}_{\lambda,n} = \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q}\right)^{-1} \frac{1}{n} V^T X^T y$ is a ridge estimator. Also, for a squared error loss, the posterior mean is the minimizer of the loss. Hence, we will investigate $\mathbb{E}\,||\,\hat{\alpha}_{\lambda,n} - \alpha^*\,||^2$

$$\mathbb{E} \parallel \hat{\alpha}_{\lambda,n} - \alpha^* \parallel^2 = \mathbb{E} \left\| \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \frac{1}{n} V^T X^T y - \alpha^* \right\|^2$$

$$= \mathbb{E} \left\| \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \, (XV)_i - \lambda \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \alpha^* \right\|^2$$

$$= \mathbb{E} \left[\mathbb{E} \left[\left\| \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \, (XV)_i - \lambda \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \alpha^* \right\|^2 \right] \mid X \right]$$

$$= \underbrace{\mathbb{E} \left\| \lambda \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \alpha^* \right\|^2}_{I} + \underbrace{\frac{\sigma^2}{n^2} \mathbb{E} \sum_{i=1}^{n} \left\| \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} (XV)_i \right\|^2}_{II}$$

$$II = \frac{\sigma^2}{n^2} \mathbb{E} \sum_{i=1}^{n} \left\| \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} (XV)_i \right\|^2$$

$$= \frac{\sigma^2}{n^2} \mathbb{E} \mathrm{Tr} \left(\left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-2} \sum_{i=1}^{n} \left(V^T X_i X_i^T V \right) \right)$$

$$= \frac{\sigma^2}{n} \mathbb{E} \mathrm{Tr} \left(\left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-2} \left(V^T \hat{\Sigma}_n V \right) \right)$$

$$I = \mathbb{E} \left\| \lambda \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \alpha^* \right\|^2$$

$$= \mathbb{E} \left[\mathbb{E}_{\Pi_\lambda} \left\| \lambda \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \alpha^* \right\|^2 \mid X \right] \quad \text{by Fubini's Theorem}$$

$$= \frac{\sigma^2}{n} \mathbb{E} \mathrm{Tr} \left(\left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-2} \right)$$

$$\implies I + II = \frac{\sigma^2}{n} \mathbb{E} \left[\mathrm{Tr} \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \right]$$

This provides a lower bound on the risk. For all $\lambda > 0$,

$$\inf_{\hat{\alpha}} \sup_{\alpha^*} \mathbb{E} \parallel \hat{\alpha} - \alpha^* \parallel^2 \geq \inf_{\alpha} \mathbb{E}_{\alpha^* \sim \Pi_\lambda} \parallel \hat{\alpha} - \alpha^* \parallel^2 = \mathbb{E}_{\alpha^* \sim \Pi_\lambda} \parallel \hat{\alpha}_{\lambda,n} - \alpha^* \parallel^2 = \frac{\sigma^2}{n} \mathbb{E} \left[\mathrm{Tr} \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \right]$$

Now, this risk is well defined for all $\lambda > 0$. Additionally, it is a decreasing function of λ and it is strictly bounded below by 0. Therefore, by monotone convergence theorem, we have $\lim_{\lambda \to 0} \frac{\sigma^2}{n} \mathbb{E} \left[\mathrm{Tr} \left(V^T \hat{\Sigma}_n V + \lambda I_{p-q} \right)^{-1} \right] = \frac{\sigma^2}{n} \mathbb{E} \left[\mathrm{Tr} \left(V^T \hat{\Sigma}_n V \right)^{-1} \right]$, and this will act as a lower bound as well. Therefore, we have

$$\inf_{\hat{\alpha}} \sup_{\alpha^*} \mathbb{E} \parallel \alpha - \alpha^* \parallel^2 = \frac{\sigma^2}{n} \mathbb{E} \left[\mathrm{Tr} \left(V^T \hat{\Sigma}_n V \right)^{-1} \right]$$

1.2. *Constrained Least Squares estimator derivation (Proof of Lemma A2)*

We have the following optimization problem

$$\min ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 \text{ subject to } A\boldsymbol{\beta} = \boldsymbol{c}$$

We label our objective function as $L(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta})$. We know, by definition, that $L(\boldsymbol{\beta})$ is minimized by $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. We can expand the objective around this minimizer as follows:

$$L(\boldsymbol{\beta}) = L(\hat{\boldsymbol{\beta}}) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T X^T X \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)$$

Denote $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$ as $\boldsymbol{\delta}$ and $A\hat{\boldsymbol{\beta}} - \boldsymbol{c} = \boldsymbol{\gamma}$. Then the optimization problem becomes

$$\min \boldsymbol{\delta}^T X^T X \boldsymbol{\delta} \text{ subject to } A\boldsymbol{\delta} = \boldsymbol{\gamma}$$

We use a Lagrangian method to optimize this by setting the new objective to

$$L(\boldsymbol{\delta}, \lambda) = \boldsymbol{\delta}^T X^T X \boldsymbol{\delta} + \lambda^T (A\boldsymbol{\delta} - \boldsymbol{\gamma})$$

Taking the derivative with respect to $\boldsymbol{\delta}$ and the multiplier parameter and setting to zero, we get

$$\frac{\partial L(\boldsymbol{\delta}, \lambda)}{\partial \lambda} = A\boldsymbol{\delta} - \boldsymbol{\gamma} := 0$$

$$\frac{\partial L(\boldsymbol{\delta}, \lambda)}{\partial \boldsymbol{\delta}} = 2\boldsymbol{\delta}^T X^T X + \lambda^T A := 0$$

$$\implies 2\boldsymbol{\delta} = -\left(X^T X\right)^{-1} A^T \lambda$$

$$\implies \lambda = -2\left(A\left(X^T X\right)^{-1} A^T\right)^{-1} \boldsymbol{\gamma} \quad \text{multiplying through by } A$$

$$\implies \boldsymbol{\delta} = \left(X^T X\right)^{-1} A^T \left(A\left(X^T X\right)^{-1} A^T\right)^{-1} \boldsymbol{\gamma}$$

Translating this solution into the original formulation, we get

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \left(X^T X\right)^{-1} A^T \left(A\left(X^T X\right)^{-1} A^T\right)^{-1} \left(A\hat{\boldsymbol{\beta}} - \boldsymbol{c}\right)$$

which is our desired estimator.

1.3. *Asymptotics of Minimax Error (Proof of Corollary 1)*

We define $D \in \mathbb{R}^{p \times p}$ as the deterministic equivalent of a random matrix $\tilde{D} \in \mathbb{R}^{p \times p}$, if, for any unit vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$ we have

$$\frac{1}{p}\mathrm{Tr}\left(D - \tilde{D}\right) = o_{\mathbb{P}}(1) \text{ and } \boldsymbol{a}^T \left(D - \tilde{D}\right) \boldsymbol{b} = o_{\mathbb{P}}(1)$$

We will be using Corollary 4.2 from Dobriban & Sheng (2022), a consequence of Rubio & Mestre (2011), which gives the notion of asymptotic equivalence for the inverse of a covariance matrix using deterministic equivalents. The corollary gives us

$$\frac{1}{p}\mathrm{Tr}\left(\hat{\Sigma}^{-1} - \frac{\Sigma^{-1}}{1 - \alpha}\right) = o_{\mathbb{P}}(1) \text{ and } \boldsymbol{a}^T \left(\hat{\Sigma}^{-1} - \frac{\Sigma^{-1}}{1 - \alpha}\right) \boldsymbol{b} = o_{\mathbb{P}}(1)$$

for unit vectors a, b.

In the given setup, we have X_i being drawn from $N_p(0, \Sigma)$, and our estimate for the covariance matrix is given by $\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$. We define $\tilde{X}_i = V^T X_i \sim N_{p-q}(0, V^T \Sigma V)$. Thus, we have

$$\frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{X}_i^T = V^T \hat{\Sigma}_n V$$

We appeal to this theory of deterministic equivalents, put forth in Hachem et al. (2007) to deal with the asymptotics associated with these sample covariance matrices and their inverses. We will be using Theorem 4.1 and Corollary 4.2 from Dobriban & Sheng (2022), which utilizes the assumptions from Hachem et al. (2007), that is X is equivalent in distribution to $\Sigma^{1/2} Z$, where entries of Z are mean zero, unit variance and have finite $8 + c$th moment where $c > 0$. Additionally, in either formulation of Assumption 2, we have $\mu_{\hat{\Sigma}_n}$ converges to $\text{MP}(\alpha) \boxtimes \mu_\infty$, where $\mu_{\hat{\Sigma}_n}$ is the empirical spectral distribution of $\hat{\Sigma}_n$, μ_∞ is the limiting empirical spectral distribution of Σ, and $\boxtimes$ denotes the free multiplicative convolution function Bai & Zhou (2008); Chouard (2022). This convergence implies the convergence of the empirical spectral measure of $M^T \hat{\Sigma}_n M$ for any sequence of deterministic matrices M. Recent works (Louart & Couillet, 2021; Chouard, 2022) have allowed us to utilize the deterministic equivalents for weaker assumptions, which we have utilized as Assumption 2.1. Namely, we assume that every row of X are i.i.d. distributed from a centered distribution, any contrast of the form $a^T X$ is subgaussian with variance proxy $a^T \Sigma a$, and Σ has bounded spectral norm. Therefore, by Corollary 4.2 from Dobriban & Sheng (2022), we have the following result, by definition of deterministic equivalent.

$$\frac{1}{n} \text{Tr}\left(V^T \hat{\Sigma}_n V\right)^{-1} = \frac{\text{Tr}\left(V^T \Sigma V\right)^{-1}}{n(1 - (1 - \gamma)\alpha)} + o_{\mathbb{P}}(1)$$

This gives a convergence in probability. In order to show a convergence of the expectation, we must show uniform integrability of $T_n = \frac{1}{n} \text{Tr}\left(V^T \hat{\Sigma}_n V\right)^{-1}$ for all n. We leave this as a claim and then prove it at the end of this section.

Also, by the equivalence proven by the theorem, we know that

$$\lim_{n \to \infty} \frac{1}{n} \text{Tr}\left(V^T \Sigma V\right)^{-1} = \lim_{n \to \infty} \frac{1}{n} \text{Tr}\left(C_\Sigma \Sigma^{-1}\right)$$

where $C_\Sigma = I_p - \Sigma^{-1} A \left(A^T \Sigma^{-1} A\right)^{-1} A^T$, which are all positive random variables.

Thus, we have

$$\lim_{n \to \infty} \mathbb{E}\left\|\tilde{\beta} - \beta^*\right\|^2 = \frac{\sigma^2}{(1 - (1 - \gamma)\alpha)} \lim_{n \to \infty} \frac{1}{n} \text{Tr}\left(\Sigma^{-1} - \Sigma^{-1}\left(A^T \left(A \Sigma^{-1} A^T\right)^{-1} A\right) \Sigma^{-1}\right)$$

We assume the existence of the limit on the right hand side. This is reasonable since the limiting spectrum of Σ is bounded away from 0 and ∞, allowing for invertibility. Plugging in $\Sigma = I_p$ for the isotropic case gives us the corresponding result.

Now, we prove uniform integrability of T_n. Since we are operating in a regime where n is bounded below by p, we will show the result for all n large enough, leaving a finite complement, which allows for uniform integrability. We will utilize the De la Vallée Poussin's condition for uniform integrability. That is, if we can show that $T_n^{1+\delta}$ is bounded for some $\delta > 0$, this is an equivalent definition of uniform integrability.

Note that for A_n a standard Wishart matrix, we can write $T_n = \text{Tr}\left(\left(V^T \Sigma V\right)^{-1/2} A_n \left(V^T \Sigma V\right)^{-1/2}\right)$, which means we can bound

$$\frac{1}{\lambda_{\max}(V^T \Sigma V)} \text{Tr}(A_n^{-1}) \leq T_n \leq \frac{1}{\lambda_{\min}(V^T \Sigma V)} \text{Tr}(A_n^{-1})$$

This shows that it is enough to show UI for $\tilde{T}_n := \text{Tr}(A_n^{-1})$, since we have constant bounds on either side. For a standard Wishart matrix, using the Schur complement, we can show that $(A_n)_{jj} \sim \chi^2_{n-p+q+1}$. For large enough n and $\delta > 0$, we have

$$\mathbb{E}\left[\left(\text{Tr}\left(A_n^{-1}\right)\right)^{1+\delta}\right] \leq (p-q)^\delta \sum_{j=1}^{p-q} \mathbb{E}\left[\left(A_n^{-1}\right)_{jj}^{1+\delta}\right]$$

$$= (p-q)^{1+\delta} \mathbb{E}\left[\left(\chi^2_{n-p+q+1}\right)^{-1-\delta}\right]$$

$$= (p-q)^{1+\delta} 2^{-1-\delta} \frac{\Gamma\left(\frac{n-p+q+1}{2} - 1 - \delta\right)}{\Gamma\left(\frac{n-p+q+1}{2}\right)}$$

The last equality holds if $\frac{n-p+q+1}{2} > 1 + \delta$

Using Gautschi's inequality (Gautschi, 1959), we can show that for a large enough constant C, we have

$$\mathbb{E}\left[\left(\text{Tr}\left(A_n^{-1}\right)\right)^{1+\delta}\right] \leq C\,(p-q)^{1+\delta}\,n^{-1-\delta} < \infty$$

This shows that $\tilde{T}_n^{1+\delta}$ is bounded for all n such that $\frac{n-p+q+1}{2} > 1 + \delta$. Note that uniform integrability holds for any finite set of $O(1)$ variables. Thus, we have shown uniform integrability of $\tilde{T}_n$ for all $n > p$, which implies uniform integrability of T_n for all $n > p$.

1.4.   *"Gain" of using Constraints when $q \ll p$ (Proof of Proposition 1)*

We are analyzing the "gain"

$$G_n = ||\hat{\beta}_{LS} - \beta^*||^2 - ||\hat{\beta}_{\mathcal{P}} - \beta^*||^2$$

$$= ||\left(X^T X\right)^{-1} X^T \epsilon||^2 - ||\mathcal{P}_A^\perp \left(X^T X\right)^{-1} X^T \epsilon||^2$$

$$= ||\mathcal{P}_A \left(X^T X\right)^{-1} X^T \epsilon||^2$$

$$= \epsilon^T X \left(X^T X\right)^{-1} \mathcal{P}_A \left(X^T X\right)^{-1} X^T \epsilon$$

We have $\mathbb{E}\left[G_n \mid X\right] = \sigma^2 \text{Tr}\left(\mathcal{P}_A \left(X^T X\right)^{-1}\right) = \frac{\sigma^2}{n} \text{Tr}\left(\mathcal{P}_A \hat{\Sigma}_n^{-1}\right)$

We will use deterministic equivalents as given in Section 1.3 to evaluate this expectation.

$$\mathbb{E}\left[G_n \mid X\right] = \frac{\sigma^2}{n} \text{Tr}\left(\mathcal{P}_A \hat{\Sigma}_n^{-1}\right)$$

$$= \sigma^2 \underbrace{\frac{1}{n}\left[\text{Tr}\left(\mathcal{P}_A \hat{\Sigma}_n^{-1}\right) - \frac{1}{1-\alpha}\text{Tr}\left(\mathcal{P}_A \Sigma^{-1}\right)\right]}_{g_1} + \frac{\sigma^2}{n(1-\alpha)}\text{Tr}\left(\mathcal{P}_A \Sigma^{-1}\right)$$

Given Corollary 4.2 from Dobriban & Sheng (2022), we see that $g_1 = O_{\mathbb{P}}\left(\frac{||\mathcal{P}_A||_{op}}{n\sqrt{n}}\right)$. Considering a rescaling, we see that

$$\mathbb{E}\left[\frac{n}{q}G_n \mid X\right] = \underbrace{\frac{1}{q(1-\alpha)}\mathrm{Tr}\left(\mathcal{P}_A\Sigma^{-1}\right)}_{g_2} + O_{\mathbb{P}}\left(\frac{||\mathcal{P}_A||_{op}}{q\sqrt{n}}\right)$$

Under isotropic design, we have $g_2 = \frac{\sigma^2}{q(1-\alpha)}\mathrm{Tr}\left(\mathcal{P}_A\Sigma^{-1}\right) = \frac{\sigma^2}{q(1-\alpha)}\mathrm{Tr}\left(\mathcal{P}_A\right) = \frac{\sigma^2}{1-\alpha}$. The proof of uniform integrability of $\left\{\mathbb{E}\left[\frac{n}{q}G_n \mid X\right]\right\}_{n\geq 1}$ follows quite directly from the analysis of T_n in the previous section. Indeed, for some $\delta > 0$, we have

$$|\mathbb{E}\left[G_n \mid X\right]|^{1+\delta} = \left|\frac{\sigma^2}{n}\mathrm{Tr}\left(\mathcal{P}_A\hat{\Sigma}_n^{-1}\right)\right|^{1+\delta} \leq \sigma^{2+2\delta}\left|\frac{1}{n}\mathrm{Tr}\left(\hat{\Sigma}_n^{-1}\right)\right|^{1+\delta} = \sigma^{2+2\delta}\left|\mathrm{Tr}\left(A_n^{-1}\right)\right|^{1+\delta}$$

where A_n is a standard Wishart matrix. From the previous section, we have shown the uniform integrability of these trace objects, which in turn gives us the uniform integrability of $\left\{\mathbb{E}\left[\frac{n}{q}G_n \mid X\right]\right\}_{n\geq 1}$. Given this uniform integrability, we can assert that $\lim_{n\to\infty}\mathbb{E}\left[\frac{n}{q}G_n\right] = \frac{\sigma^2}{1-\alpha}$. Therefore, we have $\mathbb{E}\left[G_n\right] = \frac{q}{n(1-\alpha)}(1+o(1))$. We will evaluate the variance of G_n now.

$$\mathrm{Var}\left(G_n\right) = 2\sigma^4\mathrm{Tr}\left(X\left(X^TX\right)^{-1}\mathcal{P}_A\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1}\mathcal{P}_A\left(X^TX\right)^{-1}X^T\right)$$

$$= \frac{2\sigma^4}{n^2}\underbrace{\mathrm{Tr}\left(\left(\mathcal{P}_A\hat{\Sigma}_n^{-1}\right)^2\right)}_{O(q)}$$

Thus, we have

$$\mathbb{P}\left(|G_n| > \varepsilon\mathbb{E}|G_n|\right) \leq \frac{\mathrm{Var}(G_n)}{\varepsilon^2\mathbb{E}^2|G_n|} = \frac{1}{2\sigma^2\varepsilon^2 q} \to 0$$

if $q \to 0$. Therefore, $G_n \xrightarrow{\mathbb{P}} \frac{\sigma^2\gamma\alpha}{1-\alpha}$.

Now, we consider the case where $q = O(1)$. We additionally assume Assumption 3. Therefore, G_n is a quadratic form. Let us denote $Q = X\left(X^TX\right)^{-1}\mathcal{P}_A\left(X^TX\right)^{-1}X^T$. Let us diagonalize $Q = UWU^T$, where U is an orthogonal matrix and W is a diagonal matrix of eigenvalues $w_1 \geq \ldots \geq w_n$ of Q. Note that $w_{q+1} = \ldots = w_n = 0$. Since $\epsilon/\sigma \sim N_n(0, I_n)$, we have $U^T\epsilon/\sigma \sim N_n(0, I_n)$. Therefore, we have $\frac{G_n}{\sigma^2} \stackrel{d}{=} v^TWv$ where $v \sim N_n(0, I_n)$. Therefore,

$$\frac{G_n}{\sigma^2} \stackrel{d}{=} \sum_{i=1}^{q} w_i v_i^2$$

which gives us the necessary result. Also, in order to have non-vanishing eigenvalues, we scale G_n by n.

1.5. *Asymptotic Normality of $\tilde{\beta}$ (Proof of Proposition 2)*

We can establish the asymptotic normality of the CLS estimator using the independence and identical distribution of each of the errors. We operate under the assumption that the errors are independent of the draws of the design matrix. The form of the CLS estimator for the jth

coordinate is

$$\tilde{\beta}_j = \beta_j + e_j^T C_{A^\perp} \hat{\Sigma}_n^{-1} \left(\frac{1}{n} X^T \epsilon \right)$$

Rearranging, we have

$$\sqrt{n} \left(\tilde{\beta}_j - \beta_j \right) = e_j^T C_{A^\perp} \hat{\Sigma}_n^{-1} \left(\frac{1}{\sqrt{n}} X^T \epsilon \right) = \sum_{i=1}^{n} a_{n,i} \epsilon_i$$

where $a_{n,i} = \frac{1}{\sqrt{n}} e_i^T X \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j$, ϵ_i are iid with $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}\,(\epsilon_i) = \sigma^2$ and $\mathbb{E}[\epsilon^{2+\delta}] < \infty$ for δ given in (1). We can invoke the CLT if this quantity satisfies the Lyapunov condition (Billingsley, 1986). Note that $a_{n,i}$ are random variables independent of ϵ_i, therefore, we will invoke a conditional version of the Lyapunov CLT and then finally resolve the distribution of the $a_{n,i}$'s. This will provide sufficient conditions on X that would allow for asymptotic normality. Let us denote $S_n = \sum_{i=1}^{n} a_{n,i} \epsilon_i$ and $v_n^2 = \sigma^2 \sum_{i=1}^{n} a_{n,i}^2$. We have the following lemma that provides conditions on $a_{n,i}$.

LEMMA 4.*(i)*

$$e_j^T C_{A^\perp} \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j > 0$$

with probability 1

(ii)

$$\lim_{n \to \infty} \left(e_j^T C_{A^\perp} \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j - \frac{1}{1 - (1-\gamma)\alpha} e_j^T C_\Sigma \Sigma^{-1} C_\Sigma^T e_j \right) = 0$$

(iii)

$$\frac{\sum_{i=1}^{n} \left| \frac{1}{\sqrt{n}} e_i^T X \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j \right|^{2+\delta}}{\left(e_j^T C_{A^\perp} \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j \right)^{1+\delta/2}} \xrightarrow{\mathbb{P}} 0$$

for δ given in (1)

We can see that given Lemma 4, we have $v_n > 0$ w.p. 1 and v_n^2 converges to a constant (in probability) as $n \to \infty$. Let $\mathcal{A}_n$ denote the σ-field generated from $\{a_{n,i}\}$. Considering a conditional version of the scaled sum

$$\Lambda_n := \frac{1}{v_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left[|a_{n,i} \epsilon_i|^{2+\delta} \mid \mathcal{A}_n \right] = \frac{1}{v_n^{2+\delta}} \sum_{i=1}^{n} |a_{n,i}|^{2+\delta} \mathbb{E}\left[|\epsilon_i|^{2+\delta} \right]$$

$$= \underbrace{\frac{\mathbb{E}\left[|\epsilon_1|^{2+\delta} \right]}{\sigma_n^{2+\delta}}}_{<\infty} \frac{\sum_{i=1}^{n} |a_{n,i}|^{2+\delta}}{\left(\sum_{i=1}^{n} a_{n,i}^2 \right)^{1+\delta/2}}$$

$$\xrightarrow{\mathbb{P}} 0 \text{ by Lemma 4 (iii)}$$

We will use the standard Lyapunov CLT on the event that $\{\Lambda_n \le \eta\}$ for any $\eta > 0$. We have

$$\frac{S_n}{v_n} \mid \mathcal{A}_n \xrightarrow{d} N(0, 1) \text{ on } \{\Lambda_n \le \eta\}$$

by Lyapunov CLT. Since $\Lambda_n \xrightarrow{\mathbb{P}} 0$, we have uniform convergence

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_n}{v_n} \mid \mathcal{A}_n \right) - \Phi(x) \right| \xrightarrow{\mathbb{P}} 0$$

Now, for any bounded, uniformly continuous f,

$$\mathbb{E}\left[f\left(\frac{S_n}{v_n} \right) \right] = \mathbb{E}\left[\mathbb{E}\left(f\left(\frac{S_n}{v_n} \right) \mid X \right) \right] \to \mathbb{E}[f(Z)]$$

where $Z \sim N(0,1)$ and Φ is the CDF of the standard normal distribution. This gives us the asymptotic normality. Now we need to prove Lemma 4. Lemma 4 (ii) follows from Section 1.3 using deterministic equivalents. We will prove Lemma 4 (iii) and in turn show (i).

Note that we have the following inequality

$$\sum_{i=1}^{n} |a_{n,i}|^{2+\delta} \leq \max_{1 \leq i \leq n} |a_{n,i}|^{\delta} \sum_{i=1}^{n} a_{n,i}^2$$

Therefore, it is enough for us to verify

$$\left(\frac{\max_{1 \leq i \leq n} |a_{n,i}|}{\left(\sum_{j=1}^{n} a_j^2 \right)^{1/2}} \right)^{\delta} \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty$$

We have

$$\frac{\max_{1 \leq i \leq n} \left| \frac{1}{\sqrt{n}} e_i^T X \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j \right|}{\left(e_j^T C_{A^\perp} \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j \right)^{1/2}} = \frac{A_1}{A_2}$$

If we show $A_1 \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ and show that A_2 is bounded away from 0 with high probability, then we would have shown that the Lindeberg condition holds.

$$A_1 = \frac{\max_{1 \leq i \leq n} \left| e_i^T X \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j \right|}{\sqrt{n}} \leq \frac{\lambda_{\max}\left(\hat{\Sigma}_n^{-1} C_{A^\perp}^T \right)}{\sqrt{n} \min_{1 \leq i \leq n} ||e_i^T X||_1} = \frac{B_1}{B_2}$$

We use the variational form of the largest eigenvalue and the properties of the deterministic equivalent from the previous section. Define $a = \operatorname*{argmax}_{||v||=1} v^T \hat{\Sigma}_n C_{A^\perp}^T v$.

We have

$$\begin{aligned} B_1 &= a^T \hat{\Sigma}_n C_{A^\perp}^T a \\ &= \frac{a^T C_\Sigma \Sigma^{-1} a}{1 - (1 - \gamma)\alpha} + o_{\mathbb{P}}(1) \text{ from Section 1.3} \\ &\leq \frac{\lambda_{\max}\left(C_\Sigma \Sigma^{-1} \right)}{1 - (1 - \gamma)\alpha} + o_{\mathbb{P}}(1) \\ &= O(1) + o_{\mathbb{P}}(1) \end{aligned}$$

Each of the X_i are drawn independently from a multivariate normal with dispersion matrix Σ. Thus, we have $B_2 = \sqrt{n} \min_{1 \leq i \leq n} ||Z_i||_1$, where $Z_i \overset{i.i.d.}{\sim} N(0, \Sigma)$.

If $\mathbb{P}\left(\min_{1\le i\le n}||\mathbf{Z}_i||_1 > \frac{\kappa}{\sqrt{n}}\right) \to 1$ for any fixed κ, then we have $\frac{1}{B_2} \xrightarrow{\mathbb{P}} 0$

$$\mathbb{P}\left(\min_{1\le i\le n}||\mathbf{Z}_i||_1 > \frac{\kappa}{\sqrt{n}}\right) = \mathbb{P}\left(||\mathbf{Z}_1||_1 > \frac{\kappa}{\sqrt{n}}\right)^n \ge \mathbb{P}\left(|\mathbf{Z}_1^T\mathbb{1}| > \frac{\kappa}{\sqrt{n}}\right)^n$$

$$= \mathbb{P}\left(|\mathbf{Z}| > \frac{\kappa}{\sqrt{n\mathbb{1}^T\Sigma\mathbb{1}}}\right)^n = \left(1 - \mathrm{erf}\left(\frac{\kappa}{\sqrt{2n\mathbb{1}^T\Sigma\mathbb{1}}}\right)\right)^n$$

where $\mathbf{Z} \sim \mathcal{N}(0,1)$ and erf is the error function. We can see that $\mathrm{erf}\left(\frac{\kappa}{\sqrt{2n\mathbb{1}^T\Sigma\mathbb{1}}}\right) = o\left(\frac{1}{n}\right)$, since $\mathbb{1}^T\Sigma\mathbb{1} = \Omega(p)$. Thus, we have

$$\left(1 - \mathrm{erf}\left(\frac{\kappa}{\sqrt{2n\mathbb{1}^T\Sigma\mathbb{1}}}\right)\right)^n \to 1$$

which proves that $\frac{1}{B_2} \xrightarrow{\mathbb{P}} 0$

Finally, we need to find a condition such that we can bound A_2 away from zero. Relabelling $C_{A^\perp}^T e_j =: v_j$, we have

$$v_j^T\hat{\Sigma}^{-1}v_j \ge ||v_j||^2\lambda_{\min}\left(\hat{\Sigma}^{-1}\right)$$

Now, this is bounded away from 0 with high probability if $||v_j||^2$ is bounded away from 0 with high probability.

$$||v_j||^2 = e_j^T C_{A^\perp} C_{A^\perp}^T e_j$$

$$= e_j^T V\left(V^T\hat{\Sigma}_n V\right)^{-1}\left(V^T\hat{\Sigma}_n^2 V\right)\left(V^T\hat{\Sigma}_n V\right)^{-1}V^T e_j \text{ using the notation from the proof of Theorem 1}$$

$$\ge ||V^T e_j||^2\left(\lambda_{\max}\left(V^T\hat{\Sigma}_n V\right)\right)^{-2}\lambda_{\min}\left(V^T\hat{\Sigma}_n^2 V\right)$$

The second line follows from the following alternate form of the projection matrix

$$C_{A^\perp} = V\left(V^T\hat{\Sigma}_n V\right)^{-1}V^T\hat{\Sigma}_n$$

where V is an orthogonal basis of $\mathcal{N}(A)$.

By Assumption 2, $\lambda_{\max}\left(V^T\Sigma V\right) < \infty$ and $\lambda_{\min}\left(V^T\Sigma^2 V\right) > 0$. Define $a = \arg\max_{||v||=1} v^T V^T\hat{\Sigma}_n Vv$ and $b = \arg\min_{||v||=1} v^T V^T\hat{\Sigma}_n^2 Vv$. Using a similar line of argument using deterministic equivalents, we can conclude the following:

$$\lambda_{\max}\left(V^T\hat{\Sigma}_n V\right) = a^T V^T\hat{\Sigma}_n Va$$

$$= a^T V^T\Sigma Va + o_{\mathbb{P}}(1) \text{ from Section 1.3}$$

$$\le \lambda_{\max}\left(V^T\Sigma V\right) + o_{\mathbb{P}}(1)$$

$$\lambda_{\min}\left(V^T\hat{\Sigma}_n^2 V\right) = b^T V^T\hat{\Sigma}_n^2 Vb$$

$$\ge b^T V^T\Sigma^2 Vb + o_{\mathbb{P}}(1) \text{ from Section 1.3 and } \Sigma \text{ p.s.d.}$$

$$\ge \lambda_{\min}\left(V^T\Sigma^2 V\right) + o_{\mathbb{P}}(1)$$

Therefore, we can conclude that asymptotically, $\lambda_{\max}\left(V^T \Sigma V\right) < \infty$ and $\lambda_{\min}\left(V^T \Sigma^2 V\right) > 0$ with high probability. Additionally, if V is an orthonormal basis, we have $V^T e_j$ is non-zero with high probability. Therefore, we have $A_2 > 0$ with high probability, proving Lemma 4(i) and (iii).

Therefore, by Lyapunov CLT, we have

$$
\left(\frac{\sqrt{n}\left(e_j^T C_{A^\perp} \hat{\Sigma}_n^{-1}\left(\frac{1}{n}X^T \epsilon\right)\right)}{\sqrt{\sigma^2 e_j^T C_{A^\perp} \hat{\Sigma}_n^{-1} C_{A^\perp}^T e_j}} \right) \mid X \xrightarrow{d} N(0,1)
$$

Applying Lemma 4(ii) with Slutsky's Theorem and plugging in the matrices for C_Σ gives us the result.

1.6. *Jackknife Variance (Proof of Proposition 3)*

Proposition 3 gave a scaling that allows us to adjust the variance of the jackknife estimator and in turn, the lengths of the confidence intervals. We will show that this adjustment is accurate. Without loss of generality, we will study the case where $\Sigma = I_p$ and $\boldsymbol{\beta} = 0$. The proof techniques are similar to the calculations in Karoui & Purdom (2016).

We have defined $\tilde{\boldsymbol{\beta}}$ as the CLS estimator, and similarly $\tilde{\boldsymbol{\beta}}_{(i)}$ as the leave-one-out CLS estimator. That is, we have $\hat{\Sigma}_{(i)} = \frac{1}{n-1}X_{(i)}X_{(i)}^T$ and thus

$$
C_{A^\perp(i)} = I_p - \hat{\Sigma}_{(i)}^{-1}A^T\left(A\hat{\Sigma}_{(i)}^{-1}A^T\right)^{-1}A
$$

We define $\hat{\epsilon}_i$ as the ith residual from the CLS estimator, and similarly $\hat{\epsilon}_{i(i)}$ as the ith residual from the leave-one-out CLS estimator, with the ith observation removed. We have the following quantification of the leave-one-out error, following from standard results involving the sample precision matrix:

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{(i)} &= C_{A^\perp}\frac{1}{n}\hat{\Sigma}_n^{-1}X^T\hat{\boldsymbol{\epsilon}} - C_{A^\perp(i)}\frac{1}{n}\hat{\Sigma}_{(i)}^{-1}X_{(i)}^T\hat{\boldsymbol{\epsilon}}_{(i)} \\
&= \left(C_{A^\perp(i)}\frac{1}{n}\hat{\Sigma}_{(i)}^{-1}X^T\hat{\boldsymbol{\epsilon}} - C_{A^\perp(i)}\frac{1}{n}\hat{\Sigma}_{(i)}^{-1}X_{(i)}\hat{\boldsymbol{\epsilon}}_{(i)}\right) + \left(C_{A^\perp}\frac{1}{n}\hat{\Sigma}_n^{-1}X^T\hat{\boldsymbol{\epsilon}} - C_{A^\perp(i)}\frac{1}{n}\hat{\Sigma}_{(i)}^{-1}X^T\hat{\boldsymbol{\epsilon}}\right) \\
&= C_{A^\perp(i)}\frac{1}{n}\hat{\Sigma}_{(i)}^{-1}X_i\hat{\epsilon}_i + \underbrace{\left(C_{A^\perp}\hat{\Sigma}_n^{-1} - C_{A^\perp(i)}\hat{\Sigma}_{(i)}^{-1}\right)X^T\hat{\boldsymbol{\epsilon}}}_{E}
\end{aligned}
$$

$$
\begin{aligned}
e_j^T E &= e_j^T\left(C_{A^\perp}\hat{\Sigma}_n^{-1} - C_{A^\perp(i)}\hat{\Sigma}_{(i)}^{-1}\right)X^T\hat{\boldsymbol{\epsilon}} \\
&= e_j^T V\left(\left(V^T\hat{\Sigma}_n V\right)^{-1} - \left(V^T\hat{\Sigma}_{(i)}V\right)^{-1}\right)V^T X^T\hat{\boldsymbol{\epsilon}} \\
&= e_j^T V\left(\left(\sum_{j\neq i}\left(V^T X_j\right)\left(V^T X_j\right)^T + \left(V^T X_i\right)\left(V^T X_i\right)^T\right)^{-1} - \left(V^T\hat{\Sigma}_{(i)}V\right)^{-1}\right)V^T X^T\hat{\boldsymbol{\epsilon}} \\
&= e_j^T V\underbrace{\left(\frac{\left(V^T\hat{\Sigma}_{(i)}V\right)^{-1}\left(V^T X_i X_i^T V\right)\left(V^T\hat{\Sigma}_{(i)}V\right)^{-1}}{1 + X_i^T V\left(V^T\hat{\Sigma}_{(i)}V\right)^{-1}V^T X_i}\right)}_{M}V^T X^T\hat{\boldsymbol{\epsilon}} \text{ by Sherman-Morrison}
\end{aligned}
$$

In order to show this object is $o_{\mathbb{P}}(1)$ for all i, we need to show that M has a uniform bound in operator norm for all i, and Proposition 2 gives us the necessary result. We have

$$1 + X_i^T V \left(V^T \hat{\Sigma}_{(i)} V \right)^{-1} V^T X_i = O_{\mathbb{P}}(p - q).$$

$$||M||_{op} \le \frac{1}{\left(1 + X_i^T V \left(V^T \hat{\Sigma}_{(i)} V \right)^{-1} V^T X_i \right)^2} \underbrace{\lambda_{\max} \left(\left(V^T \hat{\Sigma}_{(i)} V \right)^{-2} \right)^2}_{L} \underbrace{\lambda_{\max} \left(V^T X_i X_i^T V \right)^2}_{(p-q)^2} = O_{\mathbb{P}}(1)$$

$$L = \lambda_{\max} \left(\left(V^T \hat{\Sigma}_{(i)} V \right)^{-2} \right)^2 = \left(\frac{1}{\lambda_{\min} \left(V^T \hat{\Sigma}_{(i)} V \right)} \right)^4 \le \left(\frac{1}{\lambda_{\min} \left(V^T \Sigma V \right)} \frac{n - 1}{\left(\sqrt{n - 1} - \sqrt{q} - t \right)^2} \right)^4$$

with probability $1 - 2 \exp \left(-t^2 / 2 \right)$ (Vershynin, 2010). Note that this bound is independent of i, as it only depends on the Gaussianity of X_i. This gives $L = O_{\mathbb{P}}(1)$. Therefore, we have $e_j^T E = o_{\mathbb{P}}(1)$. Using a similar line of reasoning from Karoui & Purdom (2016), we can show that $\hat{\epsilon}_i = \frac{\hat{\epsilon}_{i(i)}}{1 + \frac{1}{n} X_i^T C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} X_i}$. Thus, for a given contrast v, we have

$$v^T \left(\tilde{\beta} - \tilde{\beta}_{(i)} \right) = \frac{1}{n} v^T C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} X_i \frac{\hat{\epsilon}_{i(i)}}{1 + \frac{1}{n} X_i^T C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} X_i} + o_{\mathbb{P}}(1)$$

Therefore, we have

$$n \sum_{i=1}^{n} \left[v^T \left(\tilde{\beta} - \tilde{\beta}_{(i)} \right) \right]^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\left[v^T C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} X_i \hat{\epsilon}_{i(i)} \right]^2}{\left[1 + \frac{1}{n} X_i^T C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} X_i \right]^2} + o_{\mathbb{P}}(1)$$

We can look at the asymptotics of the denominator separately.

$$1 + \frac{1}{n} X_i^T C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} X_i = 1 + \frac{1}{n} \text{Tr} \left(C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} \right) + o_{\mathbb{P}}(1)$$

$$= 1 + \frac{1}{n} \text{Tr} \left(\left(V^T \hat{\Sigma}_{(i)} V \right)^{-1} \right) + o_{\mathbb{P}}(1)$$

$$= 1 + \frac{1}{n - 1} \frac{\text{Tr} \left(\left(V^T \Sigma V \right)^{-1} \right)}{1 - (1 - \gamma)\alpha} + o_{\mathbb{P}}(1) \text{ from Section 1.3}$$

$$= 1 + \frac{(1 - \gamma)\alpha}{1 - (1 - \gamma)\alpha} + o_{\mathbb{P}}(1)$$

$$= \frac{1}{1 + (1 - \gamma)\alpha} + o_{\mathbb{P}}(1)$$

Using the deterministic equivalent, we can show this uniform control of the error terms for all i. Therefore, we can essentially factor out the denominator, and only focus on the sum of the numerators.

We will use notation from Karoui & Purdom (2016) to investigate the numerator. Let

$$T_i = v^T C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} X_i \hat{\epsilon}_{i(i)} = v^T C_{A^{\perp}(i)} \hat{\Sigma}_{(i)}^{-1} X_i \left(\epsilon_i - X_i^T \left(\tilde{\beta}_{(i)} - \beta \right) \right)$$

Using the independence of the errors, we have

$$\mathbb{E}\left[T_i^2\right] = \underbrace{\mathbb{E}\left(\epsilon_i^2\right)}_{=\sigma^2} \underbrace{\mathbb{E}\left(\left(\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} X_i\right)^2\right)}_{I} + \underbrace{\mathbb{E}\left(\left[X_i^T \left(\tilde{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}\right)\right]^2 \left[\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} X_i\right]^2\right)}_{II}$$

$$
\begin{aligned}
I &= \mathbb{E}\left(\left(\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} X_i\right)^2\right) \\
&= \mathbb{E}\left(\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} X_i X_i^T \hat{\Sigma}_{(i)}^{-1} C_{A^\perp(i)}^T \boldsymbol{v}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left[\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} X_i X_i^T \hat{\Sigma}_{(i)}^{-1} C_{A^\perp(i)}^T \boldsymbol{v} | X_{(i)}\right]\right) \\
&= \mathbb{E}\left(\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} \mathbb{E}\left[X_i X_i^T | X_{(i)}\right] \hat{\Sigma}_{(i)}^{-1} C_{A^\perp(i)}^T \boldsymbol{v}\right) \\
&= \mathbb{E}\left(\boldsymbol{v}^T \left(C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1}\right)^2 \boldsymbol{v}\right)
\end{aligned}
$$

For *II*, we use the fact that this can be expressed as the second moment of the covariance between two Gaussian vectors. That is

$$\mathbb{E}\left(\left(\boldsymbol{a}^T X_i\right)^2 \left(\boldsymbol{b}^T X_i\right)^2\right) = ||\boldsymbol{a}||_2^2 ||\boldsymbol{b}||_2^2 + 2\left(\boldsymbol{a}^T \boldsymbol{b}\right)^2$$

$$
\begin{aligned}
II &= \mathbb{E}\left(\left[X_i^T \left(\tilde{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}\right)\right]^2 \left[\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} X_i\right]^2\right) \\
&= \mathbb{E}\left\|\tilde{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}\right\|_2^2 \mathbb{E}\left\|\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1}\right\|_2^2 + 2\mathbb{E}\left(\boldsymbol{v}^T C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} \left(\tilde{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}\right)\right)^2 \\
&= \frac{\sigma^2 ||\boldsymbol{v}||_2^2 \left((1-\gamma)\alpha\right)}{\left(1 - ((1-\gamma)\alpha)\right)} I + o(1)
\end{aligned}
$$

The first term follows from Theorem 1, and the second term has an additional *n* scaling and therefore vanishes asymptotically. Putting this together, we have

$$\mathbb{E}\left[T_i^2\right] = \left(1 + \frac{(1-\gamma)\alpha}{1 - (1-\gamma)\alpha}\right) I ||\boldsymbol{v}||_2^2 \sigma^2 + o(1) = \frac{||\boldsymbol{v}||_2^2 \sigma^2 I}{(1 - (1-\gamma)\alpha)} + o(1)$$

We calculate *I* based on computations in Haff (1979); Holgersson & Pielaszkiewicz (2020). Using the formulation given in Theorem 1, we have

$$C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1} = V \underbrace{\left(V^T \hat{\Sigma}_{(i)}^{-1} V\right)^{-1}}_{\sim W(n-1, I_{p-q})} V^T$$

Based on Theorem 3.2 in Haff (1979) for the transformed Wishart matrix, we have,

$$I = \mathbb{E}\left[(V^T \boldsymbol{v})^T \left(C_{A^\perp(i)} \hat{\Sigma}_{(i)}^{-1}\right)^2 (V^T \boldsymbol{v})\right] \to \left(\frac{1}{(1 - (p-q)/n)^3} + o(1)\right) \boldsymbol{v}^T \underbrace{V V^T}_{\mathcal{P}_{A^\perp}} \boldsymbol{v}$$

Note that for the general Σ case, this form will hold, with the inner matrix would be of the form $\left(\Sigma^{-1} - \Sigma^{-1} \left(A^T \left(A\Sigma^{-1}A^T\right)^{-1} A\right) \Sigma^{-1}\right)$, and for the subsequent parts of the proof, we will use this form

Therefore, we have

$$\mathbb{E}\left[n\sum_{i=1}^{n}\left[\mathbf{v}^{T}\left(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{(i)}\right)\right]^{2}\right]=\sigma^{2}\frac{\mathbf{v}^{T}\left(\Sigma^{-1}-\Sigma^{-1}\left(A^{T}\left(A\Sigma^{-1}A^{T}\right)^{-1}A\right)\Sigma^{-1}\right)\mathbf{v}}{(1-(1-\gamma)\alpha)^{2}}+o(1)$$

Plugging in the variance of our CLS estimator from Proposition 2, we have

$$\mathbb{E}\left[\sum_{i=1}^{n}\left[\mathbf{v}^{T}\left(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{(i)}\right)\right]^{2}\right]=\left[\frac{1}{(1-(1-\gamma)\alpha)}+o(1)\right]\mathrm{Var}\left(\mathbf{v}^{T}\tilde{\boldsymbol{\beta}}\right)$$

Now, we just need to show that the centering does not affect the final scaling. We have proved
the result centered around the CLS estimator, and now, we must show the same holds around
the mean of the leave-one-out estimators. Let us define $\tilde{\boldsymbol{\beta}}_{(\cdot)}=\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{\beta}}_{(i)}$. We will consider the
quantity $n^{2}\left[\mathbf{v}^{T}\left(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{(\cdot)}\right)\right]^{2}$ and if we can show that this quantity vanishes with n, then the results
will still hold for this new centering.

Since $\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{(i)}=\frac{1}{n}C_{A^{\perp}(i)}\hat{\Sigma}_{(i)}^{-1}X_{i}\hat{\epsilon}_{i}+o_{\mathbb{P}}(1)$, we have

$$\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{(\cdot)}=\frac{1}{n^{2}}\sum_{i=1}^{n}C_{A^{\perp}(i)}\hat{\Sigma}_{(i)}^{-1}X_{i}\hat{\epsilon}_{i}+o_{\mathbb{P}}(1)$$

Therefore, we have

$$n^{2}\left[\mathbf{v}^{T}\left(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{(\cdot)}\right)\right]^{2}=\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}^{T}C_{A^{\perp}(i)}\hat{\Sigma}_{(i)}^{-1}X_{i}\left(\epsilon_{i}-X_{i}^{T}\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right)\right]^{2}+o_{\mathbb{P}}(1)$$

The ϵ_{i} term can be interpreted as a weighted mean of errors, which has mean 0 and vari-
ance going to 0 as well. Therefore, that term goes to 0 in ℓ_{2} norm. Note that $C_{A^{\perp}}\hat{\Sigma}^{-1}X_{i}=
C_{A^{\perp}(i)}\hat{\Sigma}_{(i)}^{-1}X_{i}(1-\gamma)\alpha$. We can use this to show that the second term also goes to 0. Therefore,
we have $n^{2}\left[\mathbf{v}^{T}\left(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{(\cdot)}\right)\right]^{2}\xrightarrow{\mathbb{P}}0$. This proves the proposition.

1.7. *Asymptotic Normality of $\hat{\boldsymbol{\beta}}_{\Sigma,\mathcal{P}}$ (Proof of Proposition 4)*

The high-dimensional projection estimator, when centered, can be viewed as a linear com-
bination of independent errors, similar to the formulation in the Proof of Proposition 2. Un-
der Assumption 3(b), we have $\hat{\boldsymbol{\beta}}_{\Sigma,\mathcal{P}}$ following a normal distribution. Therefore, we will show
coordinate-wise $\sqrt{n}$-consistency, and show the variance of this estimator to prove the proposition,
using the method of moments.

$\sqrt{n}$-consistency of $\hat{\beta}_{\Sigma,\mathcal{P},j}$ to β_{j}^{*}

We need to evaluate the centering of $\sqrt{n}\left(\hat{\beta}_{\Sigma,\mathcal{P},j}-\beta_{j}^{*}\right)$, that is, if we show
$\mathbb{E}\left[\sqrt{n}\left(\hat{\beta}_{\Sigma,\mathcal{P},j}-\beta_{j}^{*}\right)\right]=0$

We have defined $\hat{\boldsymbol{\beta}}_{\Sigma,\mathcal{P}}=\frac{1}{n}\mathcal{P}_{A^{\perp}}\Sigma^{-1}X^{T}y=\frac{1}{n}\mathcal{P}_{A^{\perp}}\Sigma^{-1}X^{T}X\boldsymbol{\beta}^{*}+\frac{1}{n}\mathcal{P}_{A^{\perp}}\Sigma^{-1}X^{T}\boldsymbol{\epsilon}$

$$\mathbb{E}\left[\sqrt{n}\left(\hat{\beta}_{\Sigma,\mathcal{P},j}-\beta_j^*\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\sqrt{n}\left(\hat{\beta}_{\Sigma,\mathcal{P},j}-\beta_j^*\right)\right]\mid X\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sqrt{n}\left(e_j^T\left(\frac{1}{n}\mathcal{P}_{A^\perp}\Sigma^{-1}X^TX-I_p\right)\beta^* + \frac{1}{n}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}X^T\epsilon\right)\right]\mid X\right]$$

$$= \mathbb{E}\left[\sqrt{n}\left(e_j^T\left(\frac{1}{n}\mathcal{P}_{A^\perp}\Sigma^{-1}X^TX-I_p\right)\beta^*\right)\right] \quad \text{by the Cramer-Wold device}$$

$$= \sqrt{n}\left(e_j^T\left(\frac{1}{n}\mathcal{P}_{A^\perp}\Sigma^{-1}\mathbb{E}\left[X^TX\right]-I_p\right)\beta^*\right)$$

$$= \sqrt{n}\left(e_j^T\left(\mathcal{P}_{A^\perp}-I_p\right)\beta^*\right)$$

$$= 0$$

Therefore, the projected estimator is $\sqrt{n}$-consistent.

Variance of $\hat{\beta}_{\Sigma,\mathcal{P},j}$

$$\text{Var}\left(\hat{\beta}_{\Sigma,\mathcal{P},j}\right) = \mathbb{E}\left[\text{Var}\left(\hat{\beta}_{\Sigma,\mathcal{P},j}\mid X\right)\right] + \text{Var}\left[\mathbb{E}\left(\hat{\beta}_{\Sigma,\mathcal{P},j}\mid X\right)\right]$$

$$= \underbrace{\mathbb{E}\left[\text{Var}\left(\frac{1}{n}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}X^T\epsilon\mid X\right)\right]}_{I} + \underbrace{\text{Var}\left[\frac{1}{n}e_j^T\Sigma^{-1}X^TX\beta^*\right]}_{II} \quad \text{by the Cramer-Wold device}$$

By the independence of X and ϵ, we have

$$I = \frac{\sigma^2}{n}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\mathcal{P}_{A^\perp}e_j$$

As a distinction from the analysis in the low dimensional case, the use of the inverse covariance matrix introduces variance in the form of the second term II.

$$II = \text{Var}\left[\frac{1}{n}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\left(X^TX\right)\beta^*\right]$$

$$= \mathbb{E}\left[\left(\frac{1}{n}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\left(X^TX\right)\beta^*\right)^2\right] - \left(e_j^T\mathcal{P}_{A^\perp}\beta^*\right)^2$$

$$= \mathbb{E}\left[\left(\frac{1}{n}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\left(X^TX\right)\beta^*\right)^2\right] - \left(\beta_j^*\right)^2$$

$$\mathbb{E}\left[\left(\frac{1}{n}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\left(X^TX\right)\beta^*\right)^2\right] = \mathbb{E}\left[\frac{1}{n^2}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\left(X^TX\right)\beta^*e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\left(X^TX\right)\beta^*\right]$$

$$= \frac{1}{n^2}e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\mathbb{E}\left[\left(X^TX\right)\beta^*e_j^T\mathcal{P}_{A^\perp}\Sigma^{-1}\left(X^TX\right)\right]\beta^*$$

We can evaluate this using Wishart calculations from Holgersson & Pielaszkiewicz (2020)

$$\mathbb{E}\left[\left(X^T X\right)\boldsymbol{\beta}^* \boldsymbol{e}_j^T \mathcal{P}_{A^\perp} \Sigma^{-1}\left(X^T X\right)\right] = n\mathcal{P}_{A^\perp}\boldsymbol{e}_j \boldsymbol{\beta}^{*T}\Sigma + n^2\Sigma\boldsymbol{\beta}^* \boldsymbol{e}_j^T \mathcal{P}_{A^\perp} + n\text{Tr}\left(\Sigma\boldsymbol{\beta}^* \boldsymbol{e}_j^T \mathcal{P}_{A^\perp}\Sigma^{-1}\right)\Sigma$$

$$\mathbb{E}\left[\left(\frac{1}{n}\boldsymbol{e}_j^T \mathcal{P}_{A^\perp}\Sigma^{-1}\left(X^T X\right)\boldsymbol{\beta}^*\right)^2\right] = \frac{1}{n}\boldsymbol{e}_j^T \mathcal{P}_{A^\perp}\Sigma^{-1}\mathcal{P}_{A^\perp}\boldsymbol{e}_j ||\Sigma^{1/2}\boldsymbol{\beta}^*||^2 + \left(\beta_j^*\right)^2 + \frac{1}{n}\left(\beta_j^*\right)^2$$

$$II = \frac{1}{n}\boldsymbol{e}_j^T \mathcal{P}_{A^\perp}\Sigma^{-1}\mathcal{P}_{A^\perp}\boldsymbol{e}_j ||\Sigma^{1/2}\boldsymbol{\beta}^*||^2 + \frac{1}{n}\left(\beta_j^*\right)^2$$

$$\text{Var}\left(\hat{\beta}_{\Sigma,\mathcal{P},j}\right) = \frac{1}{n}\left(\beta_j^*\right)^2 + \frac{\left(\sigma^2 + ||\Sigma^{1/2}\boldsymbol{\beta}^*||^2\right)}{n}\boldsymbol{e}_j^T \mathcal{P}_{A^\perp}\Sigma^{-1}\mathcal{P}_{A^\perp}\boldsymbol{e}_j$$

This gives us the variance in Proposition 4. Therefore, by matching moments, we see that $\sqrt{n}\left(\hat{\beta}_{\Sigma,\mathcal{P},j} - \beta_j^*\right)$ follows a normal distribution, giving the asymptotic distribution we desire.

2. Additional Information

2.1. *Biological Interpretation of Constraints*

Consider the levels of two traits $\boldsymbol{y}_1, \boldsymbol{y}_2$ observed in an individual. We can decompose these trait levels into a genetic component and a non-genetic component

$$\boldsymbol{y}_1 = \boldsymbol{g}_1 + \boldsymbol{e}_1$$

$$\boldsymbol{y}_2 = \boldsymbol{g}_2 + \boldsymbol{e}_2$$

Here $\boldsymbol{g}_1, \boldsymbol{g}_2$ are referred to as the genetic values, and $\boldsymbol{e}_1, \boldsymbol{e}_2$ are referred to as the residual values. If we consider $\boldsymbol{\beta}_i$ to be the vector of additive allelic effects for each of the SNPs for trait i, and X is the genotype vector for the individual (takes values 0,1,2), we have

$$\boldsymbol{g}_i = X\boldsymbol{\beta}_i$$

as each effect allele contributes the corresponding value to the total genetic value. Using this genetic effects, taken for the whole population, we define the genetic correlation between the two traits as

$$\rho_g = \frac{\sigma_{\boldsymbol{g}_1,\boldsymbol{g}_2}}{\sqrt{\sigma_{\boldsymbol{g}_1}^2 \sigma_{\boldsymbol{g}_2}^2}}$$

where $\sigma_{\boldsymbol{g}_1,\boldsymbol{g}_2}$ is the population level covariance of the genetic effects, and $\sigma_{\boldsymbol{g}_1}^2, \sigma_{\boldsymbol{g}_2}^2$ are the population level variances of the genetic effects. Now, if we are to take the covariance between two genetic values for the population, that have been normalized, we see that

$$\sigma_{\boldsymbol{g}_1,\boldsymbol{g}_2} = \frac{1}{N}\boldsymbol{\beta}_1^T X^T X\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 \tilde{\Sigma}_N \boldsymbol{\beta}_2$$

This gives the constraint that we are using in our analysis. Clearly, this is contingent on the allelic effects acting additively.

2.2. *Construction of Estimators in High Dimensions, Unknown Σ regime*

Our target of interest is the coordinate of a high dimensional regression vector

$$\beta_k = \mathbb{E}\left[\boldsymbol{y}^T X\right]\Sigma^{-1}\boldsymbol{e}_k$$

We use the approximation $\beta_{k,J} := \sum_{\ell=0}^{J} c_\ell \beta_k^{(\ell)}$ where $\beta_k^{(\ell)} := \mathbb{E}\left[y^T X\right] \Sigma^\ell e_k$. Here c_ℓ represent the Chebyshev polynomial coefficients, the selection of which is detailed in DeVore & Lorentz (1993); Orecchia et al. (2011); Walczyk et al. (2025), and $J \asymp (\log n)^c$ for some $c < 1$. We construct an unbiased estimator of $\beta_k^{(\ell)}$ using higher order U-statistics through terms defined as such:

$$\hat{\beta}_k^{(\ell)} := \frac{(n-(\ell+1))!}{n!} \sum_{1 \le i_1 \ne \ldots \ne i_{\ell+1} \le n} y_{i_1} X_{i_1}^T \left\{ \prod_{s=2}^{\ell+1} X_{i_s} X_{i_s}^T \right\} e_k$$

Thus, we have the following algorithm for constructing the coordinate of our regression vector:

1. Construct $\hat{\beta}_k^{(\ell)}$ for $\ell = 0, \ldots, J$ and $k = 1, \ldots, p$
2. Estimate $\hat{\beta}_k := \sum_{\ell=0}^{J} c_\ell \hat{\beta}_k^{(\ell)}$ for $k = 1, \ldots, p$ to make the vector $\hat{\beta}$
3. Project $\hat{\beta}_{est,j} := e_j^T \mathcal{P}_{A^\perp} \hat{\beta} + e_j^T A^T \left(AA^T\right) c$

This procedure provides us with an unbiased estimator for the coordinates of β. Orecchia et al. (2011) provide a bound on the error of estimation of the inverse as a function of J, and a similar computation can be extended to quantify the bias of estimating a coordinate of the effect vector using this method. However, we do not have guarantees on the rate of consistency, as mentioned in Chen et al. (2024). Additionally, the algorithm is computationally expensive. While Kong & Valiant (2018) provide a polynomial-time algorithm for evaluating these U-statistics, these estimations have to be constructed over a large grid of indices, making this computationally infeasible. We demonstrate this in a simulation setup in Section 3.

2.3.   *Construction of Estimators for GLMs*

We elucidate a procedure for constructing such estimators under the framework of generalized linear models (GLM). Note that we do not have optimality guarantees for these estimators, and look forward to working on the derivation of such estimators in the future. This methodology is based on the procedure given in Chen et al. (2024), assuming Σ, A, c are known. We refer to that paper for further discussions when Σ is not known and the next subsection for a discussion when A, c are not known.

We have a GLM setup as follows:

$$\mathbb{E}[y|X = X] = g\left(X^T \beta^*\right)$$

We would like to estimate β_j^* for all j in $1, \ldots, p$.

From Lemma 1 of Chen et al. (2024), we have

$$U := \mathbb{E}\left[y^T X\right] \Sigma^{-1} \mathbb{E}\left[X^T y\right] = f^2(\beta^{*T} \Sigma \beta^*) \beta^{*T} \Sigma \beta^*$$

$$V_j := \mathbb{E}\left[y^T X\right] \Sigma^{-1} e_j = f(\beta^{*T} \Sigma \beta^*) \beta_j^*$$

where $f(t) := \mathbb{E}\left[g'(Z)\right]$ with $Z \sim N(0, t)$ for $t \ge 0$.

Therefore, for a given link function g, there exists a continuous function h such that $Q := \beta^{*T} \Sigma \beta^* = h(U)$, which we can use for estimating β_j^*.

Thus, we have the following algorithm to construct an estimator in the GLM case, given Σ and this inverse function h and derivative function f given g:

1. Compute $\hat{U} = \frac{1}{n(n-1)} \sum_{1 \le i_1 \ne i_2 \le n} y_{i_1} y_{i_2} X_{i_1}^T \Sigma^{-1} X_{i_2}$

Protein Name	Heritability	Genetic Correlation	Genetic Covariance
ANTITHROMBIN˙III	0.4240401	-0.4778983	-0.1904203
C5A	0.5589884	0.5547244	0.2537773
CHL1	0.6229589	-0.4617222	-0.2229896
EDA	0.4689197	-0.5524264	-0.2314717
KALLISTATIN	0.5523707	-0.4978723	-0.2264162
M˙CSF˙R	0.4290930	0.4718710	0.1891356
RET	0.4126752	0.4804708	0.1888624
SCF˙SR	0.5021367	-0.4785794	-0.2075101
TGF˙B3	0.4665368	0.4866873	0.2034076
TRKC	0.6085098	-0.4903304	-0.2340436

Table 1: Significant proteins for BMI with their heritabilities and genetic correlations

2. Compute $\hat{V}_j = \frac{1}{n} \sum\limits_{1 \le i \le n} y_i X_i^T \Sigma^{-1} e_j$ for $j = 1, \ldots, p$

3. Find $\hat{Q} = h(\hat{U})$ and get $\hat{\beta}_{GLM,j} := \hat{V}_j / f(\hat{Q})$

4. Project $\hat{\beta}_{\mathcal{P},GLM} := \mathcal{P}_{A^\perp} \hat{\beta}_{GLM} + A^T \left(AA^T \right) c$

Remark: Since $(I - \mathcal{P}_{A^\perp}) V$ is positive semi-definite for any positive semi-definite matrix V, the total of the variances of each coordinate of the projected GLM estimator is lower than that of the coordinate-wise GLM estimator. Here V represents a diagonal matrix where $V_{ii} = v_i^2$, which is defined as the asymptotic variance of the coordinate-wise estimator given in Proposition 2 of Chen et al. (2024). We cannot make any direct comparison of the asymptotic variance for a given coordinate.

We implement this estimator in the case of logistic regression ($g(\cdot) = \text{expit}(\cdot)$) and demonstrate its numerical performance in Section 3.

2.4. *Details on Protein Selection*

For our analysis, we consider 10 proteins as the anchors of our analysis, for each of the glycemic traits. A concept necessary to choosing the proteins is heritability, which is the proportion of the trait variability attributable to variance in genetic factors. It is a population level statistic between 0 and 1 which will be useful for determining proteins whose expression levels are predominantly determined by genetics. We have genotypic data at pQTLs for the individuals in the JHS dataset. Elgart et al. (2021) and Sofer (2017) elucidate the procedure for estimation of genetic correlations and heritability for proteins, from individual level data in the JHS dataset. From these proteins, we choose the top 10 proteins with the greatest absolute genetic correlations, that have high heritability (> 0.4). We do this to pick proteins that will lend the most information to the analysis. The genetic correlations for these 10 proteins are extracted as c. From the GWAS of each of the selected proteins, we select genetic variants with a p-value below a threshold of 5×10^{-8}, accounting for collinearity of SNPs through pruning. We subset the effect vectors of each of the proteins at the selected SNPs to construct the matrix A. Tables 1,2,3 shows the proteins that we use for each of the glycemic traits in our analysis.

3. Additional Experimental Results

We catalogue the simulation results that have not been included in the main body of the paper.

Protein Name	Heritability	Genetic Correlation	Genetic Covariance
ANGIOPOIETIN˙2	0.4331282	0.6657746	0.1895192
ANTITHROMBIN˙III	0.4240401	-0.5810212	-0.1636489
CHL1	0.6229589	-0.4808414	-0.1641532
ERP29	0.4087223	0.5870843	0.1623425
IL˙1˙SRII	0.6033166	-0.4782250	-0.1606655
TPSB2	0.5180600	0.6821772	0.2123757
TSG˙6	0.6053944	-0.9503767	-0.3198400
TYK2	0.4019710	0.5809450	0.1593126
UB2G2	0.4204848	0.5513349	0.1546352
VITRONECTIN	0.7120431	-0.5510545	-0.2011247

Table 2: Significant proteins for fasting insulin with their heritabilities and genetic correlations

Protein Name	Heritability	Genetic Correlation	Genetic Covariance
C2	0.4469158	0.5761554	0.1576280
CATF	0.4772143	-0.6253038	-0.1767782
CNTN2	0.4540917	-0.6146444	-0.1695027
ERP29	0.4087223	0.7550191	0.1975391
PARC	0.5050623	0.5825308	0.1694229
PGCB	0.5056586	-0.7173123	-0.2087458
TIG2	0.6312174	0.7301262	0.2373932
TRKC	0.6085098	-0.5512258	-0.1759722
TYK2	0.4019710	0.8336819	0.2163110
UB2G2	0.4204848	0.8414359	0.2232940

Table 3: Significant proteins for HbA1c with their heritabilities and genetic correlations

3.1. *Synthetic Simulations*

Figures 1a and 1b [**s1**] provide credence to the asymptotic formulae of the MSE derived in Theorem 1, showing that it can be applied to small sample calculations as well. The black dots indicate the mean of the MSEs of the CLS estimator over all the iterations, and the error bars indicate the standard deviation of these MSEs. The blue line indicates the mean of theoretical asymptotic error values over all the iterations, and the blue ribbon indicates the standard deviation of these theoretical values.

Figure 2 [**s2,m1**] shows the asymptotic normality of the first coordinate of the CLS vector for a fixed $q = 50$. This demonstrates the property given in Proposition 2.

Similarly, Figure 3 [**s3,m1**] shows the asymptotic normality of the first coordinate of the CLS vector for a fixed $q = 150$. Therefore, we can work towards the high dimensional corollary of Proposition 2.

Figure 4 [**s2,m1**] shows the confidence intervals constructed from the OLS and CLS estimators around each coordinate of the true vector β^*. Figure 5 zooms in on 20 of the coordinates, which shows that the confidence intervals constructed from the CLS estimators are thinner, due to the smaller asymptotic variance. Figure 6 [**m2**] shows the bias of 1000 iterations of the U-statistics based method-of-moments estimator $\hat{\beta}_{est,j}$, along with the projected oracle estimator $\hat{\beta}_{\Sigma,\mathcal{P},j}$, on a single coordinate of the regression vector. $\hat{\beta}_{est,j}$ was computed using Chebyshev polynomials
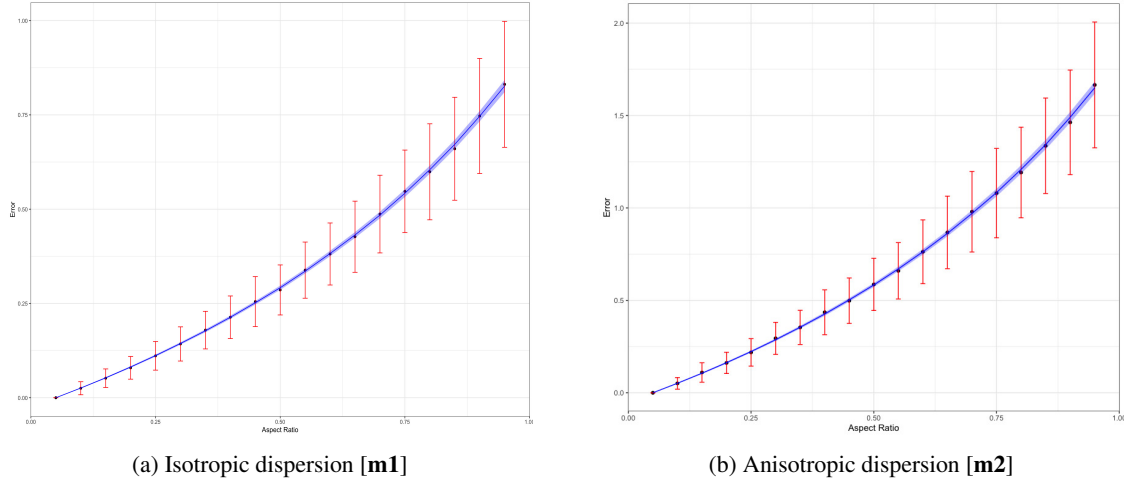
(a) Isotropic dispersion [**m1**]

(b) Anisotropic dispersion [**m2**]

Fig. 1: Adherence of MSE from simulation (black dots) to exact derived formula (blue line) for $\tilde{\beta}$ [**s1**]
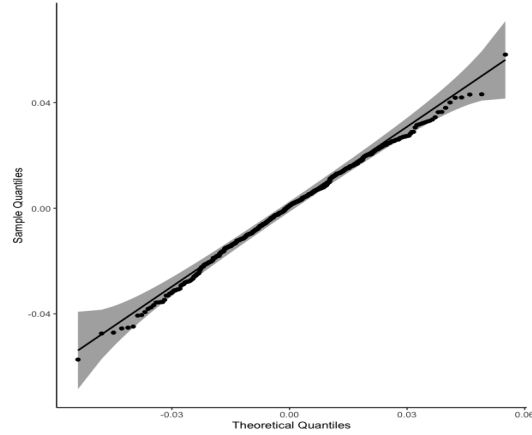

Fig. 2: QQ plot of first coordinate of CLS vector [**s2,m1**]
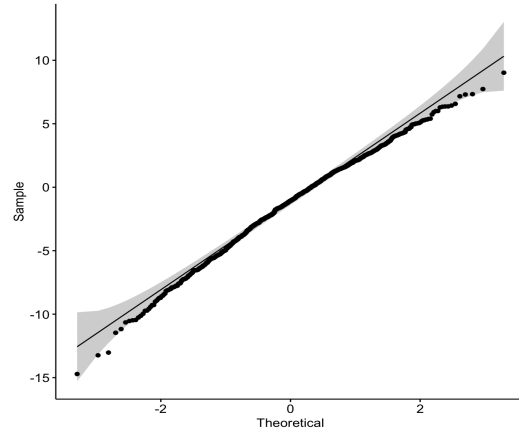

Fig. 3: QQ plot of first coordinate of CLS vector [**s3,m1**]

up to the order of $J = 3$. There is a minimal improvement as the number of constraints increase, and a variance is quite large. Additionally, the computational time increases exponentially with the value of J.

Figure 8 [**s2,m1**] compares the MSEs of the coordinate-wise MOM estimator and the projected estimator in a logistic regression setup over a range of constraint ratios.

Finally, in our data exploration, we employ the CLS estimator, presuming the constraints are known. This is because we have an exact formulation of the confidence intervals for this estimator. However, Fig. 7 [**s2,m2**] shows that under a low level of constraint ($q = 5$), we can implement the CLS estimator directly, as opposed to the semiparametric estimator, with a minimal loss of accuracy. This allows us to simply use summary level data from our reference population.
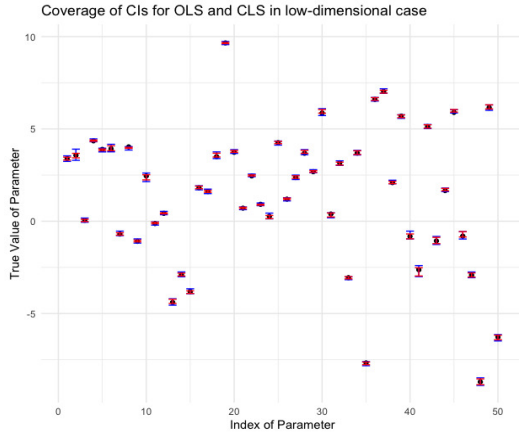
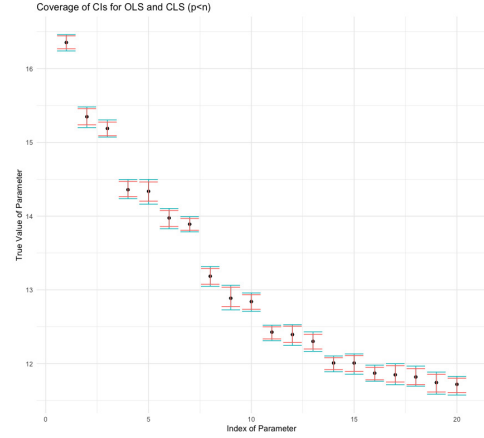Fig. 4: CIs at each coordinate for $\hat{\beta}_{LS}$ and $\tilde{\beta}$ [**s2,m1**]

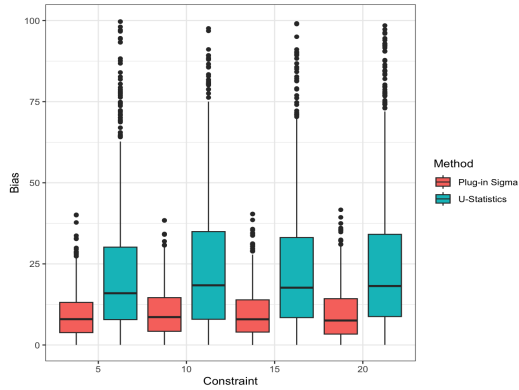Fig. 5: CIs at selected coordinates for $\hat{\beta}_{LS}$ and $\tilde{\beta}$ [**s2,m1**]

Fig. 6: Comparison of absolute bias in one coordinate ($\hat{\beta}_{\Sigma,\mathcal{P},j}$ and $\hat{\beta}_{est,j}$) with $n = 10, N = 100, p = 20$ [**m2**]
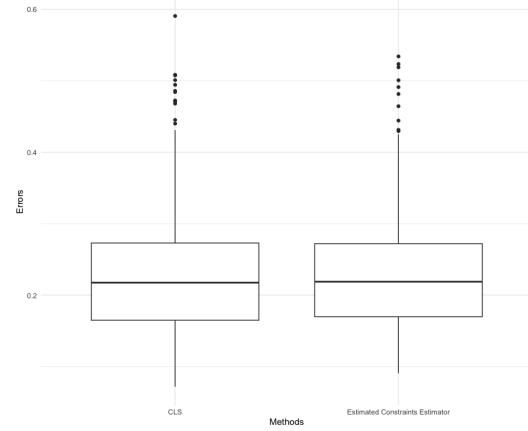

Fig. 7: Comparison of errors between $\tilde{\beta}$ and $\hat{\beta}_{semi}$ for $q = 5$ [**s2,m2**]

3.2. *Data-Informed Simulations*

We perform a series of simulations incorporating the partial data from the JHS and MESA datasets. We divide our analysis into the following cases based on the glycemic traits of interest; (**g1**) BMI, (**g2**) Fasting Insulin, (**g3**) HbA1c; and the data source used; (**d1**) A, c from JHS (p, q fixed, varying sample size), (**d2**) X, y from MESA (n, p fixed, varying constraint ratio).

Akin to Figures 2 and 4, Figure 10 [**g1,d2**] compares the MSEs of the OLS, projected and CLS estimators over a range of constraint ratios. Additional simulations in the Appendix further demonstrate the theoretical properties of our proposed estimators. We also look at the error comparison under varying sample size. Figure 9 shows the MSE over varying sample sizes. While the CLS estimator does perform better than the other two estimators, the gap is not large. This is indicative that the signal we are recovering from the JHS dataset is not very strong.

Figure 13 [**g1,d1**] shows the histogram of the first estimated coordinate over all the iterations, with the appropriate normalization, as we increase the sample size. The coordinate appears to
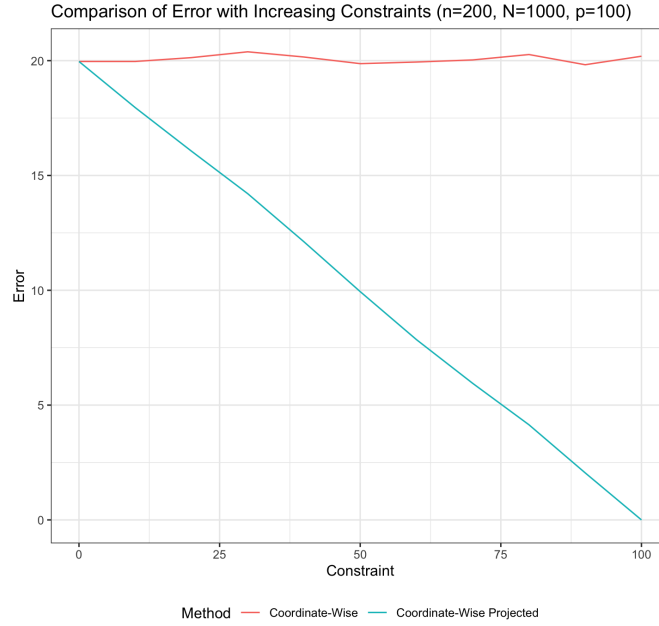
Fig. 8: Error comparison in logistic regression ($\hat{\beta}_{\mathcal{P},GLM}$ and $\hat{\beta}_{GLM}$) [**s2,m1**]
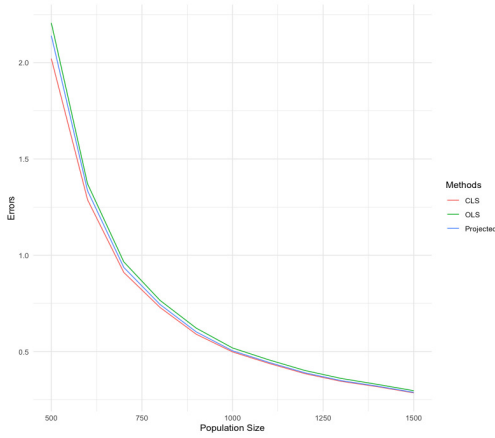

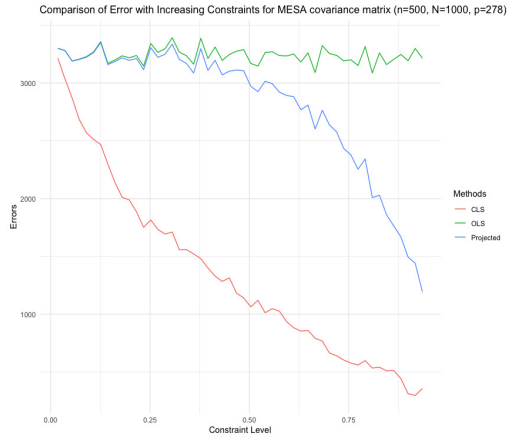Fig. 9: Comparison of MSEs between $\hat{\beta}_{LS}, \hat{\beta}_{\mathcal{P}}$ and $\tilde{\beta}$ [**g1,d1**]

Fig. 10: Error comparison ($\hat{\beta}_{LS}, \hat{\beta}\mathcal{P}, \tilde{\beta}$) [**g1,d2**]

follow a normal distribution even in the small sample case, which gives credence to the usage of asymptotic confidence intervals in the small sample case.

3.3. *Data Analysis*

For the data exploration example, we apply a similar procedure for fasting insulin and HbA1c. We tabulated the genetic variants recovered for each of the glycemic traits in Table 4. Figures 11 and 12 shows the coordinates of the CLS vector, along with the confidence intervals constructed using the naïve variance estimates, for fasting insulin and HbA1c, respectively. For fasting insulin, out of $p = 1080$ genetic variants, we reject the null hypothesis for 5 genetic variants
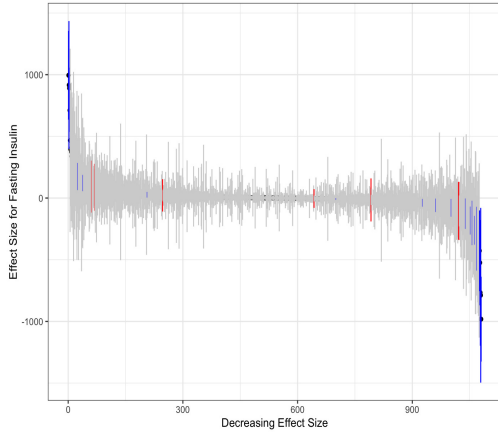
Fig. 11: Estimated coefficients ordered by effect size and 95% CIs of genetic variants in association with fasting insulin, using $\tilde{\beta}$. Blue and red CIs are significant before and after controlling for the family-wise error rate
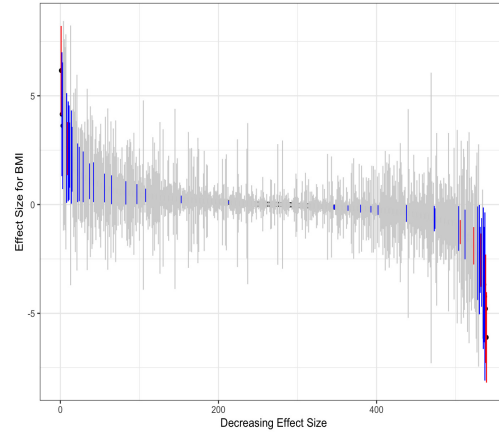
Fig. 12: Estimated coefficients ordered by effect size and 95% CIs of genetic variants in association with HbA1c, using $\tilde{\beta}$. Blue and red CIs are significant before and after controlling for the family-wise error rate
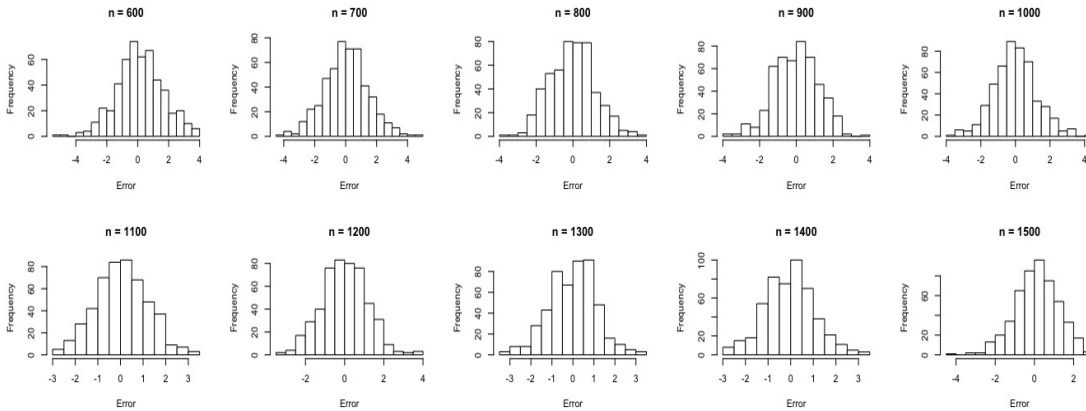

Fig. 13: Normality of coordinate over varying sample sizes [**g1,d1**]

(rs2006232, rs6740281, rs7210719, rs2277668, rs3218911), and for HbA1c, out of $p = 539$ genetic variants, we reject the null hypothesis for 6 genetic variants (rs11240346, rs3901740, rs3767283, rs7137327, rs77920745, rs80228806).

For fasting insulin, the genetic variants belong to SARM1, MAP4K4, TMEM199, SEBOX and IL1R2 genes, all genes responsible for proteins for cellular function, which could have a potential downstream effect on pancreatic tissue. For HbA1c, the genetic variants belong to CNTN2, CHPT1 and PC genes, which are all responsible for proteins related to blood cell function. This could also have a potential downstream effect on the level of glycation of the hemoglobin.

In addition to our data exploration, we perform a cross validation analysis to test the efficacy of our estimator in prediction. We split the dataset into 5 subpopulations and perform 5-fold
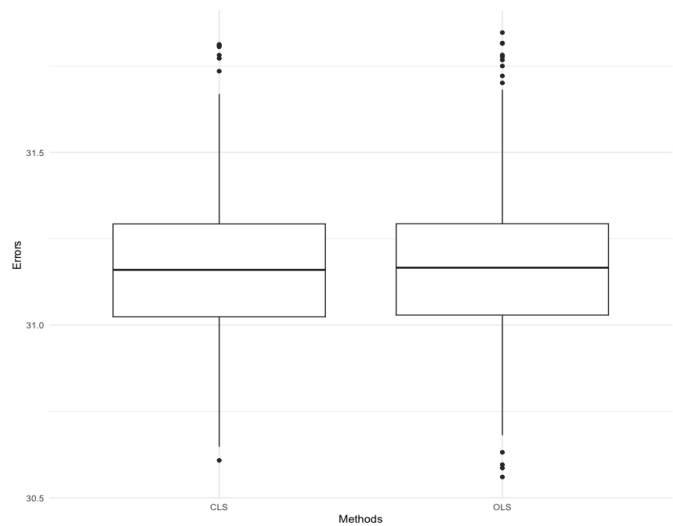
Fig. 14: Comparison of cross-validation error between $\hat{\beta}_{LS}$ and $\tilde{\beta}$

BMI ($p = 278$)	Fasting Insulin ($p = 1080$)	HbA1c ($p = 539$)
rs5510	rs2006232	rs11240346
	rs6740281	rs3901740
	rs7210719	rs3767283
	rs2277668	rs7137327
	rs3218911	rs77920745
		rs80228806

Table 4: Significant genetic variants for each glycemic traits

cross validation to get an estimate of estimation error on a real-life dataset. Figure 14 shows the mean-squared errors recovered over the iterations. The difference between the errors is negligible, most probably due to the strength of the signal from the JHS dataset. This concludes the extent of our data analysis.

REFERENCES

AMEMIYA, T. (1985). *Advanced econometrics*. Harvard university press.

BAI, Z. & ZHOU, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica* **18**, 425–442.

BILLINGSLEY, P. (1986). *Probability and Measure*. Wiley. Google-Books-ID: Q2IPAQAAMAAJ.

CHEN, X., LIU, L. & MUKHERJEE, R. (2024). Method-of-moments inference for glms and doubly robust functionals under proportional asymptotics ArXiv:2408.06103 [econ, math, stat].

CHOUARD, C. (2022). Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure ArXiv:2211.13044 [math].

DEVORE, R. A. & LORENTZ, G. G. (1993). *Constructive Approximation*. Springer Science Business Media. Google-Books-ID: cDqNW6k7˙ZwC.

DOBRIBAN, E. & SHENG, Y. (2022). Distributed linear regression by averaging ArXiv:1810.00412 [math, stat].

ELGART, M., GOODMAN, M. O., ISASI, C., CHEN, H., DE VRIES, P. S., XU, H., MANICHAIKUL, A. W., GUO, X., FRANCESCHINI, N., PSATY, B. M., RICH, S. S., ROTTER, J. I., LLOYD-JONES, D. M., FORNAGE, M., CORREA, A., HEARD-COSTA, N. L., VASAN, R. S., HERNANDEZ, R., KAPLAN, R. C., REDLINE, S., THE TRANS-OMICS FOR PRECISION MEDICINE (TOPMED) CONSORTIUM & SOFER, T. (2021). *Genetic and environmental correlations between complex*

phenotypes differ by race/ethnicity and sex.

Gautschi, W. (1959). Some elementary inequalities relating to the gamma and incomplete gamma function. *Journal of Mathematics and Physics* **38**, 77–81.

Hachem, W., Loubaton, P. & Najim, J. (2007). Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability* **17**. ArXiv:math/0507172.

Haff, L. R. (1979). An identity for the wishart distribution with applications. *Journal of Multivariate Analysis* **9**, 531–544.

Holgersson, T. & Pielaszkiewicz, J. (2020). *A Collection of Moments of the Wishart Distribution.* Cham: Springer International Publishing, p. 147–162.

Karoui, N. E. & Purdom, E. (2016). Can we trust the bootstrap in high-dimension? ArXiv:1608.00696 [math, stat].

Kong, W. & Valiant, G. (2018). Estimating learnability in the sublinear data regime.

Louart, C. & Couillet, R. (2021). Spectral properties of sample covariance matrices arising from random matrices with independent non identically distributed columns.

Mourtada, J. (2022). Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics* **50**. ArXiv:1912.10754 [math, stat].

Orecchia, L., Sachdeva, S. & Vishnoi, N. K. (2011). Approximating the exponential, the lanczos method and an (m)-time spectral algorithm for balanced separator ArXiv:1111.1491 [cs].

Rubio, F. & Mestre, X. (2011). Spectral convergence for a general class of random matrices. *Statistics Probability Letters* **81**, 592–602.

Sofer, T. (2017). Confidence intervals for heritability via haseman-elston regression. *Statistical applications in genetics and molecular biology* **16**, 259.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Walczyk, C. J., Moroz, L. V., Samotyy, V. & Cieśliński, J. L. (2025). Optimal approximation of the 1/x function using chebyshev polynomials and magic constants. *ACM Trans. Math. Softw.* **51**, 2:1–2:38.