EigenPrism: Inference for High-Dimensional Signal-to-Noise Ratios

Lucas Janson, Rina Foygel Barber, Emmanuel Candès

Abstract

Consider the following three important problems in statistical inference, namely, constructing confidence intervals for (1) the error of a high-dimensional ($p > n$) regression estimator, (2) the linear regression noise level, and (3) the genetic signal-to-noise ratio of a continuous-valued trait (related to the heritability). All three problems turn out to be closely related to the little-studied problem of performing inference on the ℓ_2-norm of the coefficient vector in high-dimensional linear regression. We derive a novel procedure for this, which is asymptotically correct and produces valid confidence intervals in finite samples as well. The procedure, called *EigenPrism*, is computationally fast and makes no assumptions on coefficient sparsity or knowledge of the noise level. We investigate the width of the EigenPrism confidence intervals, including a comparison with a Bayesian setting in which our interval is just 5% wider than the Bayes credible interval. We are then able to unify the three aforementioned problems by showing that the EigenPrism procedure with only minor modifications is able to make important contributions to all three. We also investigate the robustness of coverage and find that the method applies in practice and in finite samples much more widely than in the regime covered by the theory. Finally, we apply EigenPrism to a genetic dataset to estimate the genetic signal-to-noise ratio for a number of continuous phenotypes.

1 Introduction

1.1 Problem Statement

Throughout this paper we will assume the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{y}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Denote the i^{th} row and j^{th} column of $\boldsymbol{X}$ by $\boldsymbol{x}_i$ and $\boldsymbol{X}_j$, respectively.

Our goal is to construct a two-sided confidence interval (CI) for the signal squared magnitude $\theta^2 := \|\boldsymbol{\beta}\|_2^2$ (or equivalently just θ). Explicitly, for a given significance level $\alpha \in (0, 1)$, we want to produce statistics L_α and U_α, computed from the data, obeying

$$\begin{aligned} \mathbb{P}\left(\theta^2 < L_\alpha\right) &\leq \alpha/2, \\ \mathbb{P}\left(\theta^2 > U_\alpha\right) &\leq \alpha/2. \end{aligned} \tag{1.2}$$

In words, we want to be able to make the following statement: "with $100(1-\alpha)\%$ confidence, θ^2 lies between L_α and U_α."

1.2 Motivation

This problem can be motivated first from a high level as an approach to performing inference on $\boldsymbol{\beta}$ in high dimensions. Since $p > n$, we cannot hope to perform inference on the individual elements of

$\boldsymbol{\beta}$ directly (without further assumptions, such as sparsity), but there is hope for the one-dimensional parameter θ. Although θ is not often considered a parameter of inference in regression problems, it turns out to be closely related to a number of well-studied problems.

Suppose one has an estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$. Perhaps the most important question to be asked is: how close $\hat{\boldsymbol{\beta}}$ is to $\boldsymbol{\beta}$? This question can be answered statistically by estimating and/or constructing a CI for the error of that estimate, namely, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$. This is a fundamental statistical problem arising in many applications. Consider, for example, a compressed sensing (CS) experiment in which a doctor performs an MRI on a patient. In MRI, the image is observed not in the spatial domain, but in the frequency domain. If as many observations as pixels are made, the result is the Fourier transform (with some added noise) of the image, from which the original spatial pixels can be inferred. CS theory suggests that one can instead use a number of observations (rows of the Fourier matrix) that is a fraction of the number of pixels, and still get very good recovery of the original image using perhaps sophisticated ℓ_1 methods (Candès et al., 2006). However, for a specific instance, there is no good way to estimate how "good" the recovery is. This can be important if the doctor is looking for a specific feature on the MRI, such as a small tumor, and needs to know if what he or she sees on the reconstructed image is accurate. In the authors' experience, this is the most common question asked by end-users of CS algorithms. Put another way, when the Nyquist sampling theorem is violated, there is always a possibility of missing some of the signal, so what reassurances can we make about the quality of the reconstruction?

The estimation of the noise level σ^2 in a linear model is another important statistical problem. Consider, for example, performing inference on individual coefficients in the linear model. When $n > p$, OLS theory provides an answer that depends on σ^2 or at least an estimate of it. Indeed, one can find in almost any introductory statistics textbook both estimation and inference results for σ^2 in the case of $n > p$. However much recent work has investigated the problem of performing inference on individual coefficients in the high-dimensional setting of $n \leq p$ (Berk et al., 2013; Lockhart et al., 2014; Taylor et al., 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014; Lee et al., 2015), and they all require knowledge of σ^2. Unfortunately very few such results exist for the high-dimensional setting of $n \leq p$. Beyond regression coefficient inference, σ^2 can be useful for benchmarking prediction accuracy and for performing model selection, for instance using AIC, BIC, or the Lasso. It also may be of independent interest to know σ^2, for instance to understand the variance decomposition of $\boldsymbol{y}$.

A third topic is the study of genetic heritability (Visscher et al., 2008), which can be characterized by the following question: what fraction of variance in a trait (such as height) is explained by our genes, as opposed to our environment? Colloquially, this can be considered a way of quantifying the nature versus nurture debate.

It turns out that all three of these problems can be solved by connection with our original problem of estimating and constructing CIs for θ^2. Indeed, in the MRI example, the doctor may split the collected observations into two independent subsamples, $(\boldsymbol{y}^{(0)}, \boldsymbol{X}^{(0)})$ and $(\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)})$, and construct an estimator $\hat{\boldsymbol{\beta}}$ from just $(\boldsymbol{y}^{(0)}, \boldsymbol{X}^{(0)})$. Then the vector $\tilde{\boldsymbol{y}} := \boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\hat{\boldsymbol{\beta}}$ follows a linear model,

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}^{(1)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}, \tag{1.3}$$

so that inference on θ in this linear model corresponds exactly to inference on the ℓ_2 regression error of $\hat{\boldsymbol{\beta}}$. Note that since the analysis is *conditional* on $\hat{\boldsymbol{\beta}}$, there is no restriction on how $\hat{\boldsymbol{\beta}}$ is computed from $(\boldsymbol{y}^{(0)}, \boldsymbol{X}^{(0)})$, and so the method applies to *any* coefficient estimation technique. We defer the connection between inference for θ^2 and inference for σ^2 and genetic variance decomposition to Section 3.

1.3 Main Result

Although we will ultimately argue that our method applies more broadly, we will begin with the following distributional assumptions,

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} N(0, \boldsymbol{I}_p), \qquad \varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \tag{1.4}$$

with $\boldsymbol{X}$ independent of $\boldsymbol{\varepsilon}$. Note that $p > n$ ensures the design matrix will have a nontrivial null space, and thus conditional on $\boldsymbol{X}$, the linear model (1.1) (including θ) is unidentifiable (since any vector in the null space of $\boldsymbol{X}$ can be added to $\boldsymbol{\beta}$ without changing the data-generating process). This necessitates a random design framework. The assumption of independence on the rows of the design matrix is often satisfied in realistic settings when observations are drawn independently from a population. However, the independence (and multivariate Gaussianity) of the columns is rather stringent and just a starting point; we will explain how it can be relaxed in Section 4.1. We are treating the coefficient vector $\boldsymbol{\beta}$ as fixed, not random.

Under these assumptions, we will develop in Section 2 an estimator that is unbiased for θ^2, is almost exactly normally distributed, and has an estimable tight bound on its variance. None of these properties, including estimability of the variance, require knowledge of the noise level σ^2 or any assumption, such as sparsity, on the structure of the coefficient vector $\boldsymbol{\beta}$. From these results, it is easy to generate valid CIs for θ^2 (or θ), and we will show that such CIs are nearly as short as they can be, and provide nominal coverage in finite samples under a variety of circumstances (even beyond the assumptions made here).

1.4 Related Work

When $n > p$, ordinary least squares (OLS) theory gives us inference for $\boldsymbol{\beta}$ and thus also for θ. When $n \leq p$, the problem of estimating θ^2 has been studied in Dicker (2014). Dicker (2014) uses the method of moments on two statistics to estimate θ^2 and σ^2 without assumptions on $\boldsymbol{\beta}$. Dicker (2014) also derives asymptotic distributional results, but does not explore the estimation of the parameters of the asymptotic distributions, nor the coverage of any CI derived from it. The main contribution of our work is to provide tight, *estimable* CIs which achieve nominal coverage even in finite samples.

Inference for high-dimensional regression error, noise level, and genetic variance decomposition are each individually well-studied, so we review some relevant works here. To begin with, many authors have studied high-dimensional regression error for specific coefficient estimators, such as the Lasso (Tibshirani, 1996), often providing conditions under which this regression error asymptotes to 0 (see for example Bayati et al. (2013); Knight and Fu (2000)). To our knowledge the only author who has considered inference for a general estimator is Ward (2009), who does so using the Johnson–Lindenstrauss Lemma and assuming *no noise*, that is, $\varepsilon_i \equiv 0$ in the linear model (1.1). Thus the problem studied there is quite different from that addressed here, as we allow for noise in the linear model. Furthermore, because the Johnson–Lindenstrauss Lemma is not distribution-specific, it is conservative and thus Ward's bounds are in general conservative, while we will show that in most cases our CIs will be quite tight.

There has also been a lot of recent interest in estimating the noise level σ^2 in high-dimensional regression problems. Fan et al. (2012) introduced a refitted cross validation method that estimates σ^2 assuming sparsity and a model selection procedure that misses none of the correct variables. Sun and Zhang (2012) introduced the scaled Lasso for estimating σ^2 using an iterative procedure that includes the Lasso. Städler et al. (2010) also use an ℓ_1 penalty to estimate the noise level, but in a finite mixture of regressions model. Bayati et al. (2013) use the Lasso and Stein's unbiased risk

estimate to produce an estimator for σ^2. All of these works prove consistency of their estimators, but under conditions on the sparsity of the coefficient vector. Under the same sparsity conditions, Fan et al. (2012) and Sun and Zhang (2012) also provide asymptotic distributional results for their estimators, allowing for the construction of asymptotic CIs. What distinguishes our treatment of this problem from the existing literature is that our estimator and CI for σ^2 make *no* assumptions on the sparsity or structure of $\boldsymbol{\beta}$.

An unpublished paper (Owen, 2012) estimates θ^2 using a type of method of moments, with the goal of estimating genetic heritability by way of a variance decomposition. Although Owen gives conditions for consistency of his esimator, no inference is discussed, and he points out that the work is only valid for estimating heritability if the SNPs are assumed to be independent. In general, heritability is a well-studied subject in genetics, with especially accurate estimates coming from studies comparing a trait within and between twins (e.g. Silventoinen et al. (2003)). However, in order to better understand the genetic basis of such traits, some authors have tried to directly predict a trait from genetic information. Since most forms of genetic information, such as SNP data, are much higher-dimensional than the number of samples that can be obtained, the main approaches are either to try and find a small number of important variables through genome-wide association studies (e.g. Weedon et al. (2008)) before modeling, to estimate the kinships among subjects and use maximum likelihood, assuming independence among SNPs and random effects, on the trait covariances among subjects to estimate the (narrow-sense) heritability (e.g. Yang et al. (2010); Golan and Rosset (2011)), or to assume random effects and use maximum likelihood to estimate the signal-to-noise ratio in a linear model (e.g. Kang et al. (2008); Bonnet et al. (2014); Owen (2014)). However, attempts to explain heritability by genetic prediction have fallen quite short of the estimates from twin studies, leading to the famous conundrum of *missing heritability* (Manolio et al., 2009). Our main contribution to this field will be to consistently estimate and provide inference for the signal-to-noise ratio in a linear model, which is related to the heritability, without assumptions on the coefficient vector (such as sparsity or random effects), knowledge of the noise variance, or feature independence. This contribution may be especially valuable given the increased popularity of the rare variants hypothesis (Pritchard, 2001) for missing heritability, which conjectures that the effects of genetic variation on a trait may not be strong and sparse, but instead distributed and weak (and their corresponding mutations rare).

We note that neuroscientists have also done work estimating a signal-to-noise ratio, namely the explainable variance in functional MRI. That problem is made especially challenging due to correlations in the noise, making it different from the i.i.d. noise setting considered in this paper. For this related problem, Benjamini and Yu (2013) are able to construct an unbiased estimator in the random effects framework by permuting the measurement vector in such a way as to leave the noise covariance structure unchanged.

2 Constructing a Confidence Interval for θ^2

In this section we develop a novel method for constructing a valid CI for θ^2. This method does not require σ^2 to be known. However, for pedagogical reasons, we begin with the simpler situation in which σ^2 is known, which may arise in many signal or image processing applications.

2.1 Known σ^2

Consider a sample of size n from the linear model (1.1). Then

$$y_i \overset{i.i.d.}{\sim} N(0, \theta^2 + \sigma^2),$$

which implies

$$\frac{\|\boldsymbol{y}\|_2^2}{\theta^2+\sigma^2} \sim \chi_n^2. \tag{2.1}$$

Denote the τ^{th} quantile of the χ_n^2 distribution by $Q_\tau^{(n)}$. Then when σ^2 is known, a valid CI can be obtained by setting

$$L_\alpha = \frac{\|\boldsymbol{y}\|_2^2}{Q_{1-\alpha/2}^{(n)}} - \sigma^2, \qquad U_\alpha = \frac{\|\boldsymbol{y}\|_2^2}{Q_{\alpha/2}^{(n)}} - \sigma^2,$$

that is, (1.2) is satisfied under this choice of L_α, U_α. Note that the method of Ward (2009) also assumes σ^2 is known, and equal to zero, so we may consider comparing it to the above. In particular we want to emphasize that Ward (2009)'s inference method is conservative due to the generality of the Johnson–Lindenstrauss lemma, while $[L_\alpha, U_\alpha]$ contitutes an *exact* $100(1-\alpha)\%$ CI. The same procedure can be generalized using the bootstrap on the unbiased estimator

$$T_1 := \|\boldsymbol{y}\|_2^2/n - \sigma^2 \,. \tag{2.2}$$

See Appendix A for details.

2.2 Unknown σ^2

2.2.1 Theory

Consider again the linear model (1.1) with assumptions (1.4), in particular that $\boldsymbol{X}$ has i.i.d. standard Gaussian elements. Recall that we assume $n < p$, and let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ be a singular value decomposition (SVD) of $\boldsymbol{X}$, so that $\boldsymbol{U}$ is $n \times n$ orthonormal, $\boldsymbol{D}$ is $n \times n$ diagonal with non-negative, non-increasing diagonal entries, and $\boldsymbol{V}$ is $p \times n$ orthonormal. Let $\boldsymbol{z} = \boldsymbol{U}^\top \boldsymbol{y}$, and denote the diagonal vector of $\boldsymbol{D}$ by $\boldsymbol{d}$. We emphasize that the singular values in $\boldsymbol{D}$ are arranged along the diagonal in decreasing order, so that $d_1 \geq d_2 \geq \cdots d_n \geq 0$. Then

$$\boldsymbol{z} = \boldsymbol{D}(\boldsymbol{V}^\top \boldsymbol{\beta}) + \boldsymbol{U}^\top \boldsymbol{\varepsilon},$$

and note that

$$\begin{aligned}
\mathbb{E}\left[z_i^2 | \boldsymbol{d}\right] &= \mathbb{E}\left[\left(d_i \boldsymbol{V}_i^\top \boldsymbol{\beta} + \boldsymbol{U}_i^\top \boldsymbol{\varepsilon}\right)^2 \middle| \boldsymbol{d}\right], \\
&= d_i^2 \mathbb{E}\left[\left(\boldsymbol{V}_i^\top \boldsymbol{\beta}\right)^2 \middle| \boldsymbol{d}\right] + d_i \mathbb{E}\left[\boldsymbol{V}_i^\top \boldsymbol{\beta} \boldsymbol{U}_i^\top \boldsymbol{\varepsilon} \middle| \boldsymbol{d}\right] + \mathbb{E}\left[\left(\boldsymbol{U}_i^\top \boldsymbol{\varepsilon}\right)^2 \middle| \boldsymbol{d}\right], \\
&= d_i^2 \theta^2/p + \sigma^2,
\end{aligned}$$

where the third equality follows from the fact that in our model the columns of $\boldsymbol{V}$ are uniformly and independently distributed on the unit sphere, and independent of $\boldsymbol{d}$.

To give some intuition for what follows, assume n is even and consider the expectation, conditional on $\boldsymbol{d}$, of the difference between the sum of squares of the first half of the entries of $\boldsymbol{z}$ and the sum of squares of the second half of the entries of $\boldsymbol{z}$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n/2} z_i^2 - \sum_{i=n/2+1}^{n} z_i^2 \middle| \boldsymbol{d}\right] &= \left(\sum_{i=1}^{n/2} d_i^2 \theta^2/p + \frac{n}{2}\sigma^2\right) - \left(\sum_{i=n/2+1}^{n} d_i^2 \theta^2/p + \frac{n}{2}\sigma^2\right), \\
&= \frac{\theta^2}{p} \sum_{i=1}^{n/2} \left(d_i^2 - d_{i+n/2}^2\right).
\end{aligned}$$

Note that the terms containing σ^2 in the first line cancel out, but because the singular values d_i of $\boldsymbol{X}$ are in decreasing order, a term proportional to θ^2 remains. We generalize this idea below.

Let $\lambda_i = d_i^2/p$ be the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top/p$, let $\boldsymbol{w} \in \mathbb{R}^n$ be a vector of weights (which need not be nonnegative), and consider the statistic $S = \sum_{i=1}^n w_i z_i^2$. We can compute its expectation, conditional on $\boldsymbol{d}$, as

$$\begin{aligned}\mathbb{E}\left[S\middle|\ \boldsymbol{d}\right] &= \mathbb{E}\left[\sum_{i=1}^n w_i z_i^2 \middle| \boldsymbol{d}\right] \\ &= \sum_{i=1}^n w_i\Big(\lambda_i\theta^2 + \sigma^2\Big) \\ &= \theta^2\sum_{i=1}^n w_i\lambda_i + \sigma^2\sum_{i=1}^n w_i.\end{aligned} \tag{2.3}$$

Based on this calculation, constraining $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i\lambda_i = 1$ makes S an unbiased estimator of θ^2 (even conditionally on $\boldsymbol{d}$). We can also compute its conditional variance (see Appendix B for a detailed computation),

$$\mathrm{Var}(S|\boldsymbol{d}) = 2\sigma^4\sum_{i=1}^n w_i^2 + 4\sigma^2\theta^2\sum_{i=1}^n w_i^2\lambda_i + 2\theta^4\left(\frac{p}{p+2}\sum_{i=1}^n w_i^2\lambda_i^2 - \frac{\left(\sum_{i=1}^n w_i\lambda_i\right)^2}{p+2}\right), \tag{2.4}$$

which, under the aforementioned constraint $\sum_{i=1}^n w_i\lambda_i = 1$ can be rewritten as

$$\begin{aligned}&= 2\sigma^4\sum_{i=1}^n w_i^2 + 4\sigma^2\theta^2\sum_{i=1}^n w_i^2\lambda_i + 2\theta^4\left(\frac{p}{p+2}\sum_{i=1}^n w_i^2\lambda_i^2 - \frac{1}{p+2}\right), \\ &\le 2\sum_{i=1}^n w_i^2\left(\lambda_i\theta^2 + \sigma^2\right)^2, \\ &= 2\left(\theta^2+\sigma^2\right)^2\sum_{i=1}^n w_i^2\left(\lambda_i\rho + 1 - \rho\right)^2,\end{aligned} \tag{2.5}$$

where

$$\rho = \frac{\theta^2}{\theta^2+\sigma^2} \tag{2.6}$$

is the fraction of the variance of the y_i accounted for by the signal (recall that $\mathrm{Var}(y_i) = \theta^2 + \sigma^2$). The inequality will be quite tight when p is large and $\frac{1}{p\sum_{i=1}^n w_i^2\lambda_i^2} \ll 1$. By noting that this variance bound, as a function of ρ, is a quadratic equation with positive leading coefficient, it follows that it is maximized either at $\rho = 0$ or at $\rho = 1$. This leads to one more upper-bound,

$$\mathrm{Var}(S|\boldsymbol{d}) \le 2\left(\theta^2+\sigma^2\right)^2 \cdot \max\left\{\sum_{i=1}^n w_i^2, \sum_{i=1}^n w_i^2\lambda_i^2\right\}. \tag{2.7}$$

The above equation has two striking features. The first is that it depends on θ^2 and σ^2 only through the sum $\theta^2 + \sigma^2$, for which we have an excellent estimator given by $||\boldsymbol{y}||_2^2/n$. The second feature is that it separates into the product of two terms: one term that does not depend on $\boldsymbol{w}$, and a second term that is known (in that it contains nothing that needs to be estimated) and (strictly) *convex* in $\boldsymbol{w}$. Thus we can use convex optimization to find the vector $\boldsymbol{w}$ that minimizes

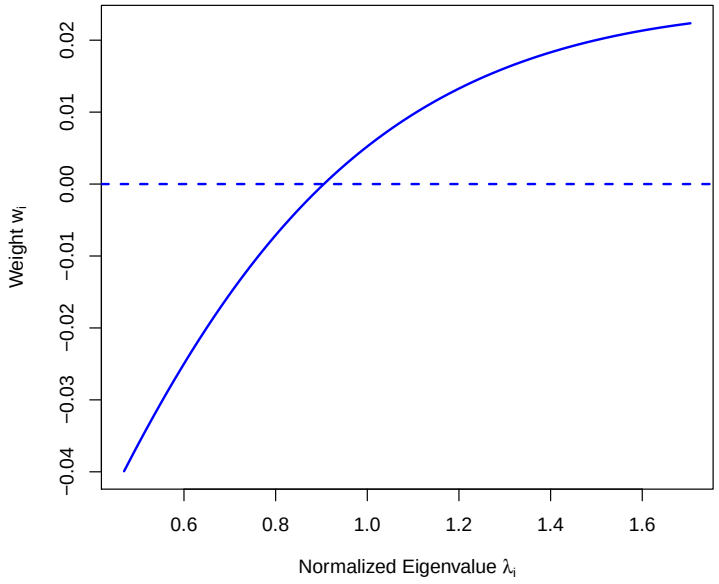


Figure 1: Plot of weights w_i as a function of normalized eigenvalues λ_i for $n = 200$ and $p = 2000$.

the upper-bound (2.7) on the variance subject to the two linear equality constraints mentioned earlier, $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i\lambda_i = 1$, which ensure that S remains unbiased for θ^2. Figure 1 shows an example of such an optimized weight vector when $n = 200$ and $p = 2000$. Note that instead of just giving some positive weight to large λ_i's and some negative weight to small λ_i's, the optimal weighting is a smooth function of the λ_i. This makes sense, as the z_i's with large associated λ_i have a larger signal-to-noise ratio, and should be given greater weight. Denote the statistic S constructed using these constrained-optimal weights by T_2. Explicitly, let $\boldsymbol{w}^*$ be the solution to the following convex optimization program $\mathcal{P}_1$:

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\arg\min}\ \max\left\{\sum_{i=1}^n w_i^2, \sum_{i=1}^n w_i^2\lambda_i^2\right\} \quad \text{such that } \sum_{i=1}^n w_i = 0,\ \sum_{i=1}^n w_i\lambda_i = 1 \tag{2.8}$$

and denote by $\mathrm{val}(\mathcal{P}_1)$ the minimized objective function value. Then the statistic for our main procedure in this paper, which we call the *EigenPrism* procedure, is the following,

$$\begin{aligned} T_2 &:= \sum_{i=1}^n w_i^* z_i^2, \\ \mathbb{E}\left[T_2 |\ \boldsymbol{d}\right] &= \theta^2, \\ \mathrm{SD}(T_2|\boldsymbol{d}) &\lesssim \sqrt{2\,\mathrm{val}(\mathcal{P}_1)}\,\frac{\|\boldsymbol{y}\|_2^2}{n}, \end{aligned} \tag{2.9}$$

where the only approximation in the variance is the replacement of $\theta^2+\sigma^2$ by its estimator $\|\boldsymbol{y}\|_2^2/n$.

With these calculations in place, we now define our $(1-\alpha)$-confidence interval for θ^2, by assuming that T_2 follows an approximately normal distribution (discussed later on). We construct lower and

upper endpoints

$$L_\alpha := \max\left\{T_2 - z^\star_{1-\alpha/2} \cdot \sqrt{2\,\mathrm{val}(\mathcal{P}_1)}\,\frac{\|\boldsymbol{y}\|_2^2}{n}, 0\right\}\ , \quad U_\alpha := T_2 + z^\star_{1-\alpha/2} \cdot \sqrt{2\,\mathrm{val}(\mathcal{P}_1)}\,\frac{\|\boldsymbol{y}\|_2^2}{n}\ ,$$

where the value of L_α is clipped at zero since it holds trivially that $\theta^2 > 0$, and where $z^\star_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

Remark. The idea of constructing the z_i's as contrasts has been used in the heritability literature before, e.g. Kang et al. (2008); Bonnet et al. (2014); Owen (2014), but in a strict random effects framework. In particular, when the entries of $\boldsymbol{\beta}$ are i.i.d. Gaussian, the z_i's become independent. With independent z_i's whose distribution depends only on the signal (θ^2) and noise (σ^2) parameters, the authors are able to apply maximum likelihood estimation, with associated asymptotic inference results for the signal, noise, or signal-to-noise ratio (we note that Bonnet et al. (2014) generalize such estimators somewhat to the case of a Bernoulli-Gaussian random effects model). The crucial difference between our work and theirs is that we make no assumptions (e.g., Gaussianity, sparsity) on the coefficient vector, and thus not only are the z_i's not independent in our setting, but their dependence (and thus the full likelihood) is a function of the products $\beta_i\beta_j$, and thus a maximum likelihood approach in this setting would still be overparameterized.

Next, we discuss the coverage and width properties of this constructed confidence interval.

2.2.2 Coverage

Now that we are equipped with an unbiased estimator and a computable variance (upper-bound), and have constructed a confidence interval (CI) using a normal approximation, there are two main questions to answer in order to determine whether these CIs will exhibit the desired coverage properties. In particular, we would like to know if substituting $\theta^2 + \sigma^2$ with $\|\boldsymbol{y}\|_2^2/n$ substantially affects the variance formula, and we would like to know if T_2 is approximately normally distributed (so that we can construct arbitrary CIs from just the second moment). For the first question, since $\|\boldsymbol{y}\|_2^2$ is a rescaled χ^2_n random variable, we can compute its coefficient of variation,

$$\frac{\mathrm{SD}(\|\boldsymbol{y}\|_2^2/n)}{\theta^2+\sigma^2} = \frac{(\theta^2+\sigma^2)\cdot \mathrm{SD}(\chi^2_n/n)}{\theta^2+\sigma^2} = \sqrt{\frac{2}{n}}. \tag{2.10}$$

Thus for even a modest high-dimensional sample size of $n = 1000$, the standard deviation of $\|\boldsymbol{y}\|_2^2/n$ is less than 5% of its expected value. For the second question, we may hope that T_2 is close to normal by a central limit theorem (CLT) argument, since it is a (albeit weighted) mean of (nearly) independent random variables z_i^2. Ultimately we defer to the simulation results of Section 3.1 to show that for problems of reasonable size ($\min\{n,p\} \gtrsim 100$), CIs constructed as if T_2 were *exactly* normal with variance *exactly* given by Equation (2.9) *never* result in below-nominal coverage.

2.2.3 Width

Once we have confirmed that our CIs provide the proper coverage, the next topic of interest is their widths. It is not hard to obtain a closed-form asymptotic upper-bound for $\mathrm{Var}(T_2)$ (the details are worked out in Appendix C). In particular, letting Y_γ denote a random variable with Marčenko–Pastur (MP) distribution with parameter γ (Marčenko and Pastur, 1967), and M_γ denote the median of Y_γ, define the constants,

$$\begin{aligned} A_\gamma &= \mathbb{E}\left[Y_\gamma \cdot (\mathbb{1}_{Y_\gamma \geq M_\gamma} - \mathbb{1}_{Y_\gamma < M_\gamma})\right], \\ B_\gamma &= \mathbb{E}\left[Y_\gamma^2\right]. \end{aligned} \tag{2.11}$$

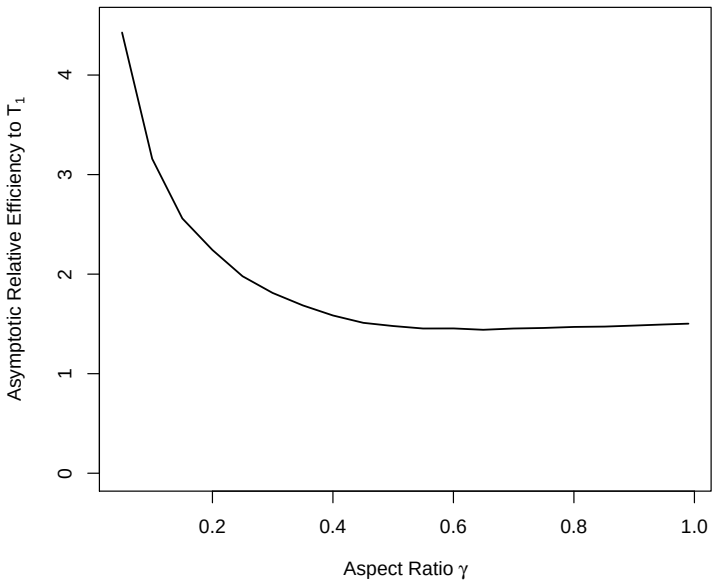


Figure 2: Estimate of the asymptotic relative efficiency of T_2 to T_1.

Then in the limit as $n, p \to \infty$ and $n/p \to \gamma \in (0, 1)$,

$$\sqrt{n} \cdot \frac{\sqrt{\mathrm{Var}(T_2)}}{\theta^2 + \sigma^2} \leq \sqrt{2} \cdot \max\left\{ \frac{1}{A_\gamma}, \frac{\sqrt{B_\gamma}}{A_\gamma} \right\}. \tag{2.12}$$

We can draw a few conclusions from Equation (2.12). The most obvious is that for $n, p \to \infty$, $n/p \to \gamma \in (0, 1)$, σ^2 asymptotically bounded above and θ^2 asymptotically bounded *below*, the error of T_2, as a fraction of its estimand θ^2, converges to 0 in probability at a rate of $n^{-1/2}$. Note that we make no assumptions at all on the structure of $\boldsymbol{\beta}$, and just require that θ^2 does not asymptote at 0. The equation also lets us compute a conservative upper-bound on the asymptotic relative efficiency (ARE), defined as the asymptotic ratio of standard deviations (although it is often defined by variances elsewhere), of T_2 with respect to T_1 from Section 2.1 (see (2.2)), the latter of which uses exact knowledge of σ^2 and has standard deviation characterized in Equation (2.10). While we may not be able to formulate a closed-form expression for it in terms of expectations due to the constrained minimization functional, the standard deviation bound for T_2 in Equation (2.9) will also converge to a constant times $\mathrm{SD}(T_1)$ under the same asymptotic conditions, where the constant depends only on the MP distribution. This is because the optimal weights are a smooth function of the λ_i. Due to fast convergence to the MP distribution, we can numerically approximate this *exact* asymptotic ratio. Figure 2 shows this estimate of the ARE of T_2 to T_1 as a function of γ. Note that the standard deviation bound for T_2 in Equation (2.9), used to compute the curve in Figure 2, is still an upper-bound for the ARE of T_2 with respect to T_1, but it reflects the ratio of CI widths between the EigenPrism procedure and a CI constructed from T_1 with knowledge of σ^2. The figure demonstrates how close in width the EigenPrism procedure comes to an exact CI for T_1 which knows σ^2. In particular, for $\gamma \gtrsim 0.25$, the EigenPrism CIs are at most twice as wide as those for T_1.

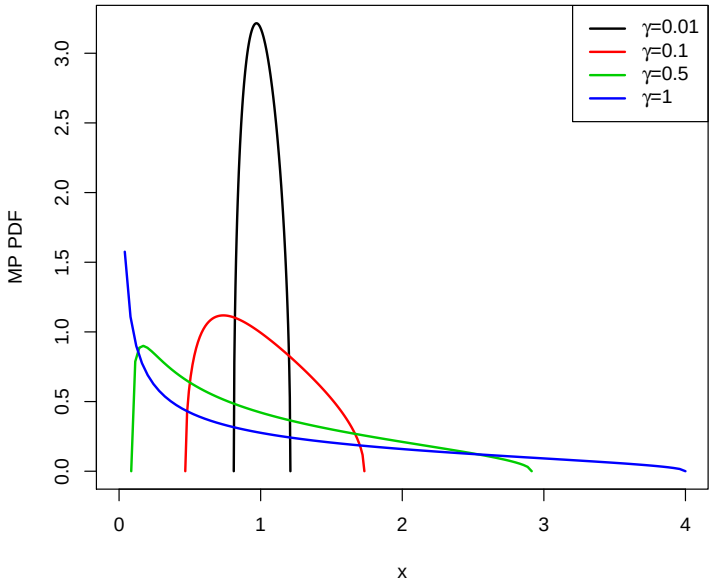


Figure 3: Probability density function (PDF) of Marčenko–Pastur distribution for various values of γ.

Another notable feature of Figure 2 is how large the ARE becomes as $\gamma \to 0$. This is a symptom of an important property of not just our procedure, but the frequentist problem as a whole. First, it is clear that if all the $\lambda_i \equiv 1$, our procedure fails, as the z_i^2 no longer provide any contrast between θ^2 and σ^2, and no linear combination of them will produce an unbiased statistic for θ^2. Intuitively, note that $\mathbb{E}\left[z_i^2\right] \propto \lambda_i \rho + (1-\rho)$, so that the problem of estimating ρ is that of estimating the slope and intercept of a regression line. But in regression, when the predictor variable assumes a constant value, as it would when $\lambda_i \equiv 1$, it becomes impossible to estimate the slope and intercept. To understand better how our procedure performs when the spread of the λ_i approaches zero, consider the case when $\lambda_1 = \cdots = \lambda_{n/2} = 1 + a$ and $\lambda_{n/2+1} = \cdots = \lambda_n = 1 - a$. In this case $\mathrm{SD}(\lambda_i) = a$, and it is easy to show that

$$\mathrm{val}(\mathcal{P}_1) = \frac{1 + a^2}{a^2 n},$$

so if $\sqrt{n}\,\mathrm{SD}(\lambda_i) \to 0$, then $a^2 n \to 0$ and so $\mathrm{val}(\mathcal{P}_1) \to \infty$.

Returning to our original model in which $\boldsymbol{X}$ is i.i.d. $N(0,1)$, the λ_i's will be approximately MP-distributed with parameter $\gamma = n/p$. Figure 3 shows visually how the width of the MP distribution depends on γ, and analytically, $\mathrm{SD}(Y_\gamma) = \sqrt{\gamma}$. We show in the following theorem (proved in Appendix D) that if the λ_i's are too close to 1 and $n \ll p$, it is impossible for *any procedure* to reliably distinguish between the case of $\rho = 0$ (pure noise) and $\rho = 1/2$ (variance equally split between signal and noise).

Theorem 1. *Let $p \geq n \geq 1$. Suppose that*

$$\boldsymbol{Z} = \theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a} + \sigma \cdot \boldsymbol{\varepsilon} \tag{2.13}$$

where $\theta > 0$, $\sigma > 0$, and a unit vector $\boldsymbol{a}$ are all fixed but unknown, $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is a known nonnegative diagonal matrix with $\sum_{i=1}^n \lambda_i := \sum_{i=1}^n D_{ii}^2/p = n$, $\boldsymbol{V} \in \mathbb{R}^{p\times n}$ is a random Haar-distributed orthonormal matrix, and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_n)$ independent of $\boldsymbol{V}$. Consider the simple scenario where we are trying to distinguish between only two possibilities, denoted by distributions P_0 and P_1:

$$P_0: \ \boldsymbol{Z} \text{ follows (2.13) with } \theta = 0 \text{ and } \sigma = 1, \quad \text{vs.} \quad P_1: \ \boldsymbol{Z} \text{ follows (2.13) with } \theta = \sigma = \frac{1}{\sqrt{2}}.$$

Then for any test $\psi : \mathbb{R}^n \to \{0, 1\}$, the power to correctly distinguish between these two distributions is bounded as

$$\mathbb{P}_{\boldsymbol{Z}\sim P_0}\{\psi(\boldsymbol{Z}) = 1\} + \mathbb{P}_{\boldsymbol{Z}\sim P_1}\{\psi(\boldsymbol{Z}) = 0\} \geq 1 - \left(\sqrt{\frac{\sum_{i=1}^n (\lambda_i - 1)^2}{8}} + \sqrt{\frac{n/p}{4\pi}}\right).$$

In other words, every test ψ has high error, with

$$\mathbb{P}(\text{Type I error}) + \mathbb{P}(\text{Type II error}) \geq 1 - \left(\sqrt{\frac{\sum_{i=1}^n (\lambda_i - 1)^2}{8}} + \sqrt{\frac{n/p}{4\pi}}\right),$$

so that if the λ_i are tightly distributed around 1 and $n \ll p$, the problem of estimating ρ, and thus θ, is extremely difficult. Note that for approximately MP-distributed λ_i with $\gamma = n/p \approx 0$, both $\sum_{i=1}^n (\lambda_i - 1)^2$ and n/p are quite small, explaining the spike in ARE in Figure 2 as $\gamma \to 0$.

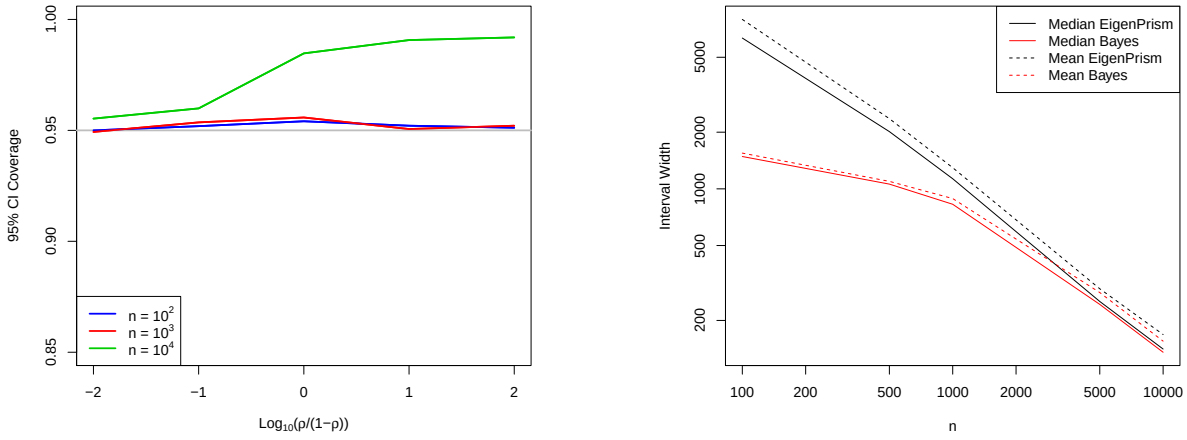
Another way to evaluate how short the EigenPrism CIs are, compared to how short they could be, is to compare to a Bayesian procedure on a Bayesian problem. This is done in Section 3.1.

2.2.4 Computation

As a procedure intended for use in high-dimensional settings, it is of interest to know how the EigenPrism procedure scales with large problem dimensions. There are essentially two parts to the procedure: the SVD, and the optimization (2.8) to choose $\boldsymbol{w}^*$. Due to the strict convexity of the optimization problem, it is extremely fast to solve (2.8) and in all of our simulations the runtime was dominated by the SVD computation. In Appendix E we include a snippet of Matlab code in the popular convex optimization language CVX (Grant and Boyd, 2014, 2008) that reformulates the optimization problem (2.8) as a second-order cone problem. Even if the optimization becomes extremely high-dimensional, note that the optimal weights $\boldsymbol{w}^*$ are a smooth function of their associated eigenvalues λ_i. Thus we can approximate $\boldsymbol{w}^*$ extremely well by subsampling the λ_i, computing a lower-resolution optimal weight vector, and then linearly interpolating to obtain the higher-resolution, high-dimensional $\boldsymbol{w}^*$. For the SVD, note that $\boldsymbol{V}$ never needs to be computed. Thus, the computation scales as $n^2 p$ with a small constant of proportionality, as the SVD of $\boldsymbol{X}\boldsymbol{X}^\top$ is all that is needed.

3 Derivative Procedures

In this section, we go into more detail about the three related problems of performing inference on estimation error of a high-dimensional regression estimator, noise level in a high-dimensional linear model, and genetic signal-to-noise ratio, including simulation results.



(a) (b)

Figure 4: (a) Coverage of 95% EigenPrism confidence intervals as a function of ρ for $p = 10^4$ and $\theta^2 + \sigma^2 = 10^4$. (b) EigenPrism confidence interval and Bayes credible interval widths as a function of n for $p = 10^4$ and $\boldsymbol{\beta}$ sampled according to the Bayesian model in Equation (F.1).

3.1 High-Dimensional Regression Error

We have already shown in Section 1.2 that the problem of inference for high-dimensional regression error is equivalent, with a change of variables, to that of inference on θ. Under assumptions (1.4), our framework even allows for selection of a subset of $\hat{\boldsymbol{\beta}}$, for instance if the doctor sees an anomaly in a region of the reconstructed image, he or she may only care about error in that region. In that case, for a subset of indices R (with corresponding complement R^c), Equation (1.3) can be rewritten as

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}_R^{(1)}(\boldsymbol{\beta}_R - \hat{\boldsymbol{\beta}}_R) + \boldsymbol{X}_{R^c}^{(1)}(\boldsymbol{\beta}_{R^c} - \hat{\boldsymbol{\beta}}_{R^c}) + \boldsymbol{\varepsilon} = \boldsymbol{X}_R^{(1)}(\boldsymbol{\beta}_R - \hat{\boldsymbol{\beta}}_R) + \tilde{\boldsymbol{\varepsilon}},$$

where $\tilde{\boldsymbol{\varepsilon}}$ is an i.i.d. Gaussian vector independent of $\boldsymbol{X}_R^{(1)}$, so that defining $\theta = \|\boldsymbol{\beta}_R - \hat{\boldsymbol{\beta}}_R\|_2$ puts this problem squarely into the EigenPrism framework, *regardless* of the fact that R may be chosen after observing $\hat{\boldsymbol{\beta}}$ (recall that $\hat{\boldsymbol{\beta}}$ was fitted on an independent subset of the data, $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)})$).

We note that the requirement that the columns of $\boldsymbol{X}$ be independent in order to perform inference on $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ cannot be relaxed. However, with a known covariance $\boldsymbol{\Sigma}$, one could instead perform inference on $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2$. Of course, inference for either $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ or $\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2$ is sufficient if the ultimate goal is to invert the CI to test a global null hypothesis on the coefficient vector.

What remains to be seen then is (1) that coverage is not lost by approximating $\theta^2 + \sigma^2$ by $\|\boldsymbol{y}\|_2^2$ and by assuming T_2 is normal, and (2) how short the resulting CIs are relative to how short they could be. To investigate (1), we fixed p at 10^4, $\theta^2 + \sigma^2 = 10^4$, varied n on a log scale between 0 and p, and varied ρ (recall Equation 2.6) between 0 and 1 by taking equally spaced values of $\log(\rho/(1-\rho))$. Note that due to rotational symmetry, the direction of $\boldsymbol{\beta}$ is irrelevant. We ran 10^4 simulations of the EigenPrism procedure to generate 95% CIs and compared coverage across the settings in Figure 4(a). Note that the CIs achieve at least nominal coverage in all cases. One setting in which we see over-coverage is when $n \approx p$ and $\rho \approx 1$. This can be explained by the variance

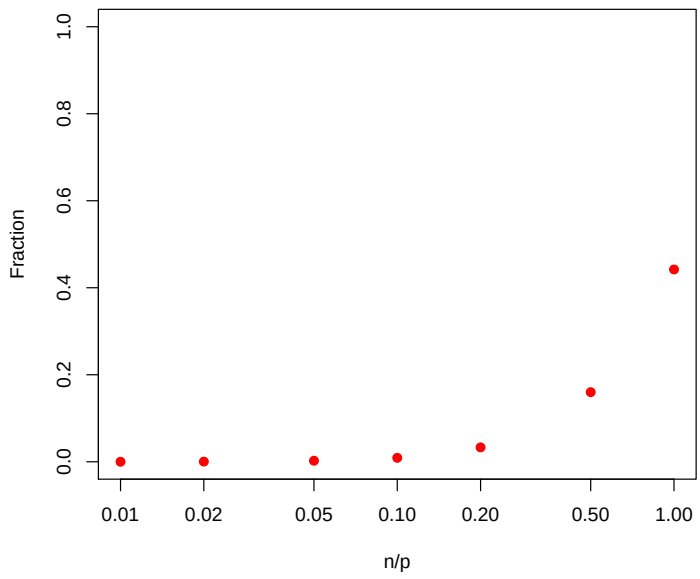


Figure 5: Plot of the fraction $\frac{1}{p\sum_{i=1}^{n}(w_i^*\lambda_i)^2}$ as a function of n/p (on the log scale) for $p = 10^4$.

upper-bound for T_2 in Equation (2.5), which is tight when p is large and $(p\sum_{i=1}^{n}(w_i^*\lambda_i)^2)^{-1} \ll 1$. Figure 5 shows that, except when $n \approx p$, we indeed have $(p\sum_{i=1}^{n}(w_i^*\lambda_i)^2)^{-1} \ll 1$.

To investigate (2), we simulated the EigenPrism procedure on a Bayesian model and compared the EigenPrism widths to those obtained by computing equal-tailed Bayes credible intervals (BCI) from a Gibbs-sampled posterior. The details of the Bayesian setup are given in Appendix F, but the resulting CI widths are summarized in Figure 4(b) for $p = 10^4$ and a range of n. We plot both mean and median to show that the inflations are not skewed, but nicely centered around their means. Each point on the plot represents 1000 simulations. The salient features of this plot are that the EigenPrism CI widths decrease at a steady $\sqrt{n}$-rate, while the BCI widths start much lower and appear to asymptote around the EigenPrism CI width curve. The fact that the BCI widths are much shorter for small n can be explained by the information contained in the priors, which is important for two reasons. In any frequentist-Bayesian comparison of methods, there is always the phenomenon that small n means the data contains little information, so the prior information given to the Bayesian method makes it heavily favored over the frequentist method. However, as we saw in Section 2.2.3, the frequentist problem is fundamentally limited not just by n but by $\mathrm{SD}(\lambda_i)$ as well, and here since p is fixed, small n corresponds to small $\mathrm{SD}(\lambda_i)$ as well, adding an extra layer of challenge for the EigenPrism procedure. As n increases though, the BCIs rely more heavily on the data, and come much closer in width to the EigenPrism CIs, with the relative width increase bottoming-out at about 5% for $n = 5000$. The relative uptick in the EigenPrism CI widths for $n \approx p$ can again be explained by the upper-bound in Equation (2.5).

3.2 Inference on σ^2

We can use almost exactly the same EigenPrism procedure for σ^2 as we did for θ^2. Recall Equation (2.3),

$$\mathbb{E}\left[S|\ \boldsymbol{d}\right] = \theta^2 \sum_{i=1}^{n} w_i\lambda_i + \sigma^2 \sum_{i=1}^{n} w_i.$$

To make S unbiased for θ^2, we constrained $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i\lambda_i = 1$. However by switching these linear constraints, so that $\sum_{i=1}^n w_i = 1$ and $\sum_{i=1}^n w_i\lambda_i = 0$, we make S unbiased for σ^2. The variance formulae and upper-bounds in Equations (2.4)–(2.7) still hold, so that we can construct T_3 (and an associated CI). Let $\boldsymbol{w}^{**}$ be the solution to the following convex optimization program $\mathcal{P}_2$:

$$\underset{\boldsymbol{w}\in\mathbb{R}^n}{\arg\min}\ \max\left\{\sum_{i=1}^{n} w_i^2, \sum_{i=1}^{n} w_i^2\lambda_i^2\right\} \quad \text{such that } \sum_{i=1}^{n} w_i = 1,\ \sum_{i=1}^{n} w_i\lambda_i = 0$$

and denote by $\mathrm{val}(\mathcal{P}_2)$ the minimized objective function value. Then the EigenPrism procedure for performing inference on σ^2 reads

$$\begin{aligned} T_3 &:= \sum_{i=1}^{n} w_i^{**} z_i^2, \\ \mathbb{E}\left[T_3|\ \boldsymbol{d}\right] &= \sigma^2, \\ \mathrm{SD}(T_3|\boldsymbol{d}) &\lesssim \sqrt{2\,\mathrm{val}(\mathcal{P}_2)}\,\frac{\|\boldsymbol{y}\|_2^2}{n}, \end{aligned}$$

where again, the only approximation in the variance is the replacement of $\theta^2 + \sigma^2$ by its estimator $\|\boldsymbol{y}\|_2^2/n$. Finally, as before, we construct the lower and upper endpoints to obtain an approximate $(1-\alpha)$-CI for σ^2 via

$$L_\alpha := \max\left\{T_3 - z^\star_{1-\alpha/2}\cdot\sqrt{2\,\mathrm{val}(\mathcal{P}_2)}\,\frac{\|\boldsymbol{y}\|_2^2}{n}, 0\right\}\ , \quad U_\alpha := T_3 + z^\star_{1-\alpha/2}\cdot\sqrt{2\,\mathrm{val}(\mathcal{P}_2)}\,\frac{\|\boldsymbol{y}\|_2^2}{n}\ .$$

Note that if the columns of $\boldsymbol{X}$ have a known covariance matrix $\boldsymbol{\Sigma}$, the exact same machinery goes through by replacing $\boldsymbol{X}$ by $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ and replacing $\boldsymbol{\beta}$ by $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}$.

Turning to simulations, we aim to show that the EigenPrism CIs for σ^2 have at least nominal coverage. We take the same setup as in Figure 4(a) but instead construct 95% CIs for σ^2. Figure 6 shows the result, and as before we see that the coverage never dips below nominal levels in any of the settings. We performed a similar experiment with a Bayesian model to compare EigenPrism CI widths for σ^2 with those of equal-tailed BCIs, but found a less-desirable comparison than in the θ case. In particular, the most favorable simulations showed the EigenPrism CI approximately 30% wider than the BCI, which can likely be attributed to the more-informative prior (Inverse Gamma) on σ^2 than that on θ^2 (nearly Exponential) in the Bayesian model (F.1). Although we would have liked to try an Exponential prior for σ^2, due to a lack of conjugacy the resulting Gibbs sampler was computationally intractable. We note that except in special cases, it can be very computationally challenging to construct BCIs, especially in high dimensions.

We point out that only two other σ^2 estimators in the literature provide any inference results, namely the scaled Lasso (Sun and Zhang, 2012) and the refitted cross validation (CV) method of Fan et al. (2012). In particular, under some sparsity conditions on the coefficient vector, the authors find aymptotic normal approximations to their estimators. To compare our CIs with theirs, we compared them on the same simulations, but quickly found that scaled Lasso and refitted CV

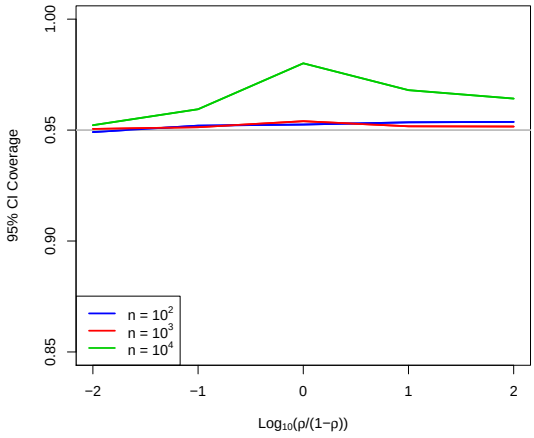


Figure 6: Coverage of 95% EigenPrism confidence intervals for σ^2 as a function of ρ for $p = 10^4$ and $\theta^2 + \sigma^2 = 10^4$. Each point represents 10^4 simulations, and the grey line denotes nominal coverage.

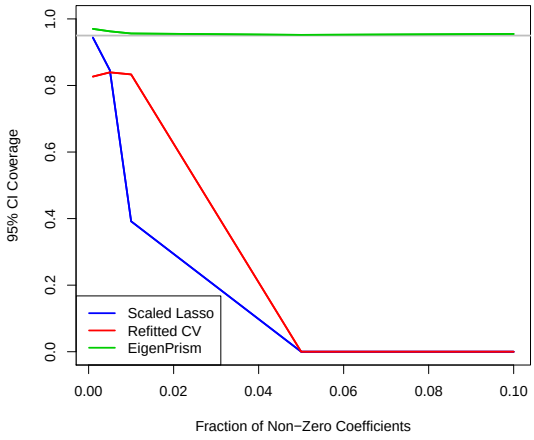


Figure 7: Coverage of scaled lasso, refitted cross validation, and EigenPrism confidence intervals when $n = 500$, $p = 1000$, $\sigma^2 = 1$, and non-zero entries of $\boldsymbol{\beta}$ equal to 1. Each point represents 10^4 simulations, and the grey line denotes nominal coverage.

CIs only achieve nominal coverage in extremely sparse settings. This is exemplified by Figure 7, which shows coverage of scaled Lasso, refitted CV, and EigenPrism CIs for a range of sparsity levels in a finite-sample setting. The scaled Lasso CIs only achieve nominal coverage when 1 out of the 1000 coefficients are non-zero, and quickly drop off to less than half of nominal coverage by 1% sparsity. The refitted CV CIs undercover by about 10% even in the sparsest settings, and also fall off further in coverage as sparsity decreases. The EigenPrism CI achieves at least nominal coverage at all sparsity levels examined.

3.3 Genetic Variance Decomposition

Consider a linear model for a centered continuous phenotype (y_i) such as height, as a function of a centered SNP array ($\boldsymbol{x}_i$). The variance can be decomposed as

$$\mathbb{E}\left[y_i^2\right] = \mathbb{E}\left[(\boldsymbol{x}_i^\top \boldsymbol{\beta})^2\right] + \sigma^2. \tag{3.1}$$

Under linkage disequilibrium, assuming column-independence is unrealistic. However, a wealth of genomic data has resulted in this column dependence possibly being estimable from outside data sets (e.g. Abecasis et al. (2012)), so we may instead take $\boldsymbol{x}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ known (we will discuss a relaxation of the normality in Section 4.1). Then Equation (3.1) reduces to

$$\mathbb{E}\left[y_i^2\right] = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 + \sigma^2,$$

which provides a formula for the linear model's signal-to-noise ratio,

$$\text{SNR} = \frac{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 + \sigma^2}.$$

The SNR is connected to the genetic heritability in that, for the simplified approximation to a linear model with additive i.i.d. noise, it quantifies what fraction of a continuous phenotype's variance can be explained by SNP data. We note that there are many different definitions of heritability, and the SNR aligns most closely with the narrow-sense, or additive, heritability, as we do not allow for interactions or dominance effects. The extent of the connection between the two definitions depends on how complete the SNP array is—if every SNP is measured, they correspond exactly.

Although until now we have been working with $\|\boldsymbol{\beta}\|_2^2$, while the SNR estimation problem seems to call for $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2$, the above problem turns out to fit right into our framework. Explicitly, the linear model can be rewritten as

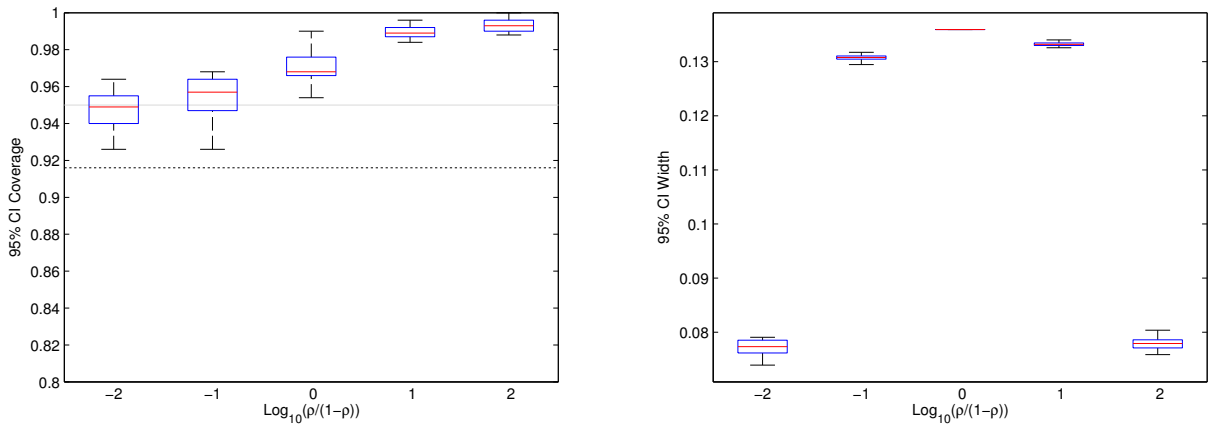
$$\boldsymbol{y} = \left(\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\right)\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\right) + \boldsymbol{\varepsilon},$$

where now the rows of $(\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2})$ are i.i.d. $N(\boldsymbol{0}, \boldsymbol{I})$, and θ^2 corresponds to the new quantity of interest: $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2$. Since $\mathbb{E}\left[\|\boldsymbol{y}\|_2^2/n\right] = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 + \sigma^2$ now, applying our methodology to $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ gives a natural estimate for SNR, namely,

$$\widehat{\text{SNR}} := \frac{T_2}{\|\boldsymbol{y}\|_2^2/n}.$$

Continuing, as we have done throughout this paper, to treat $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 + \sigma^2$ as if it is known and equal to $\|\boldsymbol{y}\|_2^2/n$, our distributional results for T_2 extend to give us an approximate confidence interval for $\widehat{\text{SNR}}$.

We turn again to simulations to demonstrate the performance of the EigenPrism procedure described above for constructing SNR CIs. One major consideration is that of course, SNP data is discrete, not Gaussian. However, we will show in Section 4.1 that the EigenPrism procedure works well empirically even under non-Gaussian marginal distributions. Here, we run experiments for $n = 10^5$, $p = 5 \times 10^5$, $\theta^2 + \sigma^2 = 10^4$, Bernoulli(0.01) design with independent columns, $\boldsymbol{\beta}$ having 10% non-zero entries, and SNR varying from nearly 0 to nearly 1. Figure 8 shows the EigenPrism CI coverage and average widths. Note that although our CIs are conservative, we never lose coverage, and at worst our 95% CI would give the SNR to within an error of $\pm 6.5\%$.



(a) (b)

Figure 8: (a) Coverage, and (b) average widths, of EigenPrism SNR 95% confidence intervals. Experiments used $n = 10^5$, $p = 5 \times 10^5$, $\theta^2 + \sigma^2 = 10^4$, Bernoulli(0.01) design, and $\boldsymbol{\beta}$ with 10% non-zero, Gaussian entries. Each boxplot summarizes the (a) coverage and (b) width for 20 different $\boldsymbol{\beta}$'s, each of which is estimated with 500 simulations. The whiskers of the boxplots extend to the maximum and minimum points. The black dotted line in (a) is the 95% confidence lower-bound for the lowest whisker in each plot assuming all CIs achieve exact coverage, and the grey line shows nominal coverage.

4 Robustness and 2-Step Procedure

In this section we follow up our investigation of the EigenPrism framework by considering its robustness to model misspecification and presenting a 2-step procedure that can improve the CI widths of the vanilla EigenPrism procedure.

4.1 Robustness

An important practical question is how robust the EigenPrism CI is to model misspecification. In particular, our theoretical calculations made some fairly stringent assumptions, and we explore here their relative importances. Some standard assumptions that we rely on are that the model is indeed linear and the noise is i.i.d. Gaussian and independent of the design matrix. These assumptions are all present, for instance, in OLS theory, and we assume that problems substantially deviating from satisfying them are not appropriate for our procedure. As explained in Section 1.2, the random design assumption is necessitated by the high-dimensionality ($p > n$) of our problem, and within the random design paradigm, the assumption of i.i.d. rows is still broadly applicable, for instance whenever the rows represent samples drawn independently from a population.

The not-so-standard assumption we make is that the columns of $\boldsymbol{X}$ are also independent, and all of $\boldsymbol{X}$'s entries are $N(0, 1)$ (note that each column of a real design matrix can always be standardized so that at least the first two marginal moments match this assumption). These assumptions are important because they ensure that the columns of $\boldsymbol{V}$ are uniformly distributed on the unit sphere, so that we can characterize both the expectation and variance of their inner product with $\boldsymbol{\beta}$. Although we will see that the marginal distribution of the elements of $\boldsymbol{X}$ is not very important as long as n and p are not small, in general the independence of the columns is crucial. We note that there is work in random matrix theory showing that for certain random matrices which are

not i.i.d. Gaussian, the eigenvectors are still in some sense asymptotically uniformly distributed on the unit sphere (see for example Bai et al. (2007)). This suggests that EigenPrism CIs, at least asymptotically, may work well in a broader context than shown so far.

Before explaining further, we feel it is important to recall that for two of the three inference problems this work addresses (inference for σ^2 and signal-to-noise ratio), the EigenPrism procedure extends to easily account for any known covariance matrix among the columns of $\boldsymbol{X}$. However in the vanilla example of simply constructing CIs for θ^2, correlation among the columns of $\boldsymbol{X}$ can cause serious problems. To first order, we need $\mathbb{E}\left[\|\boldsymbol{V}^\top\boldsymbol{\beta}\|_2^2 \middle| \boldsymbol{d}\right] \approx n/p$, or else T_2 wil be biased and the resulting shifted interval will have poor coverage. From a practical perspective, unless $\boldsymbol{\beta}$ is adversarially chosen, it may seem unlikely that $\boldsymbol{\beta}$ will be particularly aligned or misaligned (orthogonal) to the directions in which $\boldsymbol{X}$ varies. In particular, if we make a random effects assumption and say that the entries of $\boldsymbol{\beta}$ are i.i.d. $N(0, \tau^2)$, then the EigenPrism procedure will achieve nominal coverage. A slightly more subtle problem occurs if $\boldsymbol{\beta}$ is chosen not adversarially, but sparse in the basis of $\boldsymbol{X}$'s principal components. In this case, although T_2 is approximately unbiased, the variance estimate could be far too small, resulting again in degraded coverage.

To investigate how wrong the model has to be to make our CIs undercover, we construct EigenPrism CIs on data coming from models not satisfying our assumptions. In particular, we ran the EigenPrism procedure on design matrices with either i.i.d. entries with very different higher-order moments than a Gaussian, i.i.d. entries that were sparse, or Gaussian entries and correlated columns. Since the direction of $\boldsymbol{\beta}$ becomes relevant in all these cases, we performed experiments with both dense and sparse $\boldsymbol{\beta}$, and in each regime measured coverage for 20 different $\boldsymbol{\beta}$'s. The results of simulations with $n = 10^3$, $p = 10^4$, and $\theta^2 + \sigma^2 = 10^4$ are plotted in Figure 9. Each boxplot summarizes the coverage for 20 different $\boldsymbol{\beta}$'s, each of which is estimated with 500 simulations. The whiskers of the boxplots extend to the maximum and minimum points, and the black dotted line is the 95% confidence lower-bound for the lowest whisker in each plot assuming all CIs achieve exact coverage. As can be seen from Figures 9(a) and 9(c), when $\boldsymbol{\beta}$ is dense, the marginal moments and sparsity of the entries of $\boldsymbol{X}$ do not affect coverage. Figures 9(e) and 9(f) show that even small unaccounted-for correlations among the columns of $\boldsymbol{X}$ do not greatly affect coverage, although larger correlations, as expected, can result in serious undercoverage for certain $\boldsymbol{\beta}$'s. Figures 9(b) and 9(d) show that when $\boldsymbol{\beta}$ is sparse, coverage is much more sensitive to sparsity in $\boldsymbol{X}$, although if $\boldsymbol{X}$ is not sparse, coverage remains robust to higher-order moments of the design matrix. Figure 10 demonstrates the crucial difference when $\boldsymbol{X}$ is sparse by showing a few realizations of quantile-quantile plots comparing the distribution of the entries of $\boldsymbol{V}_1$ to a Gaussian distribution, for Bernoulli(0.1)- and Bernoulli(0.001)-marginally-distributed $\boldsymbol{X}$. The figure shows that the distribution for Bernoulli(0.1) is very nearly Gaussian, but that this is far from the case for Bernoulli(0.001), and thus it is the problem described at the end of the preceding paragraph that causes problems.

4.2 2-Step Procedure

Note that in the variance upper-bound of Equation (2.7), the unknown ρ is maximized over to remove it from the equation. This leads not only to conservative CIs, but suboptimal $\boldsymbol{w}^*$ as well, since $\boldsymbol{w}^*$ are obtained by minimizing this upper-bound, as opposed to the more accurate function of ρ. However by the end of the EigenPrism procedure, we have produced estimates of both θ and $\theta^2 + \sigma^2$, suggesting the possibility of a 2-step plug-in procedure to remove the need for the upper-bound in Equation (2.7). Explicitly, in the first step, we run the EigenPrism procedure to obtain an estimate $\hat{\rho} = T_2/(\|\boldsymbol{y}\|_2^2/n)$ of ρ. In the second step, we re-run the procedure treating $\rho = \hat{\rho}$ as known, and thus minimize the bound (2.5) to compute $\boldsymbol{w}^*$. Although the 2-step procedure

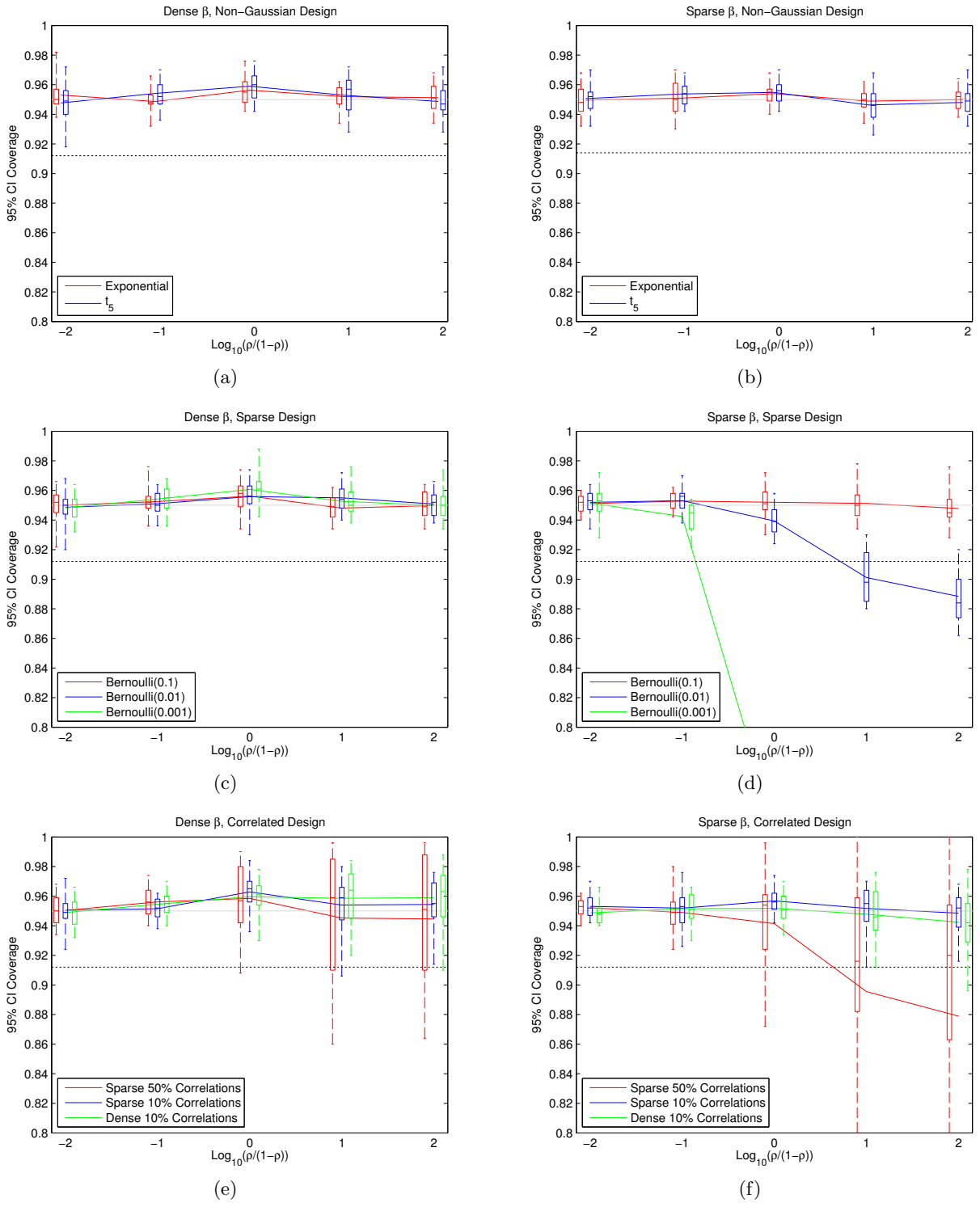
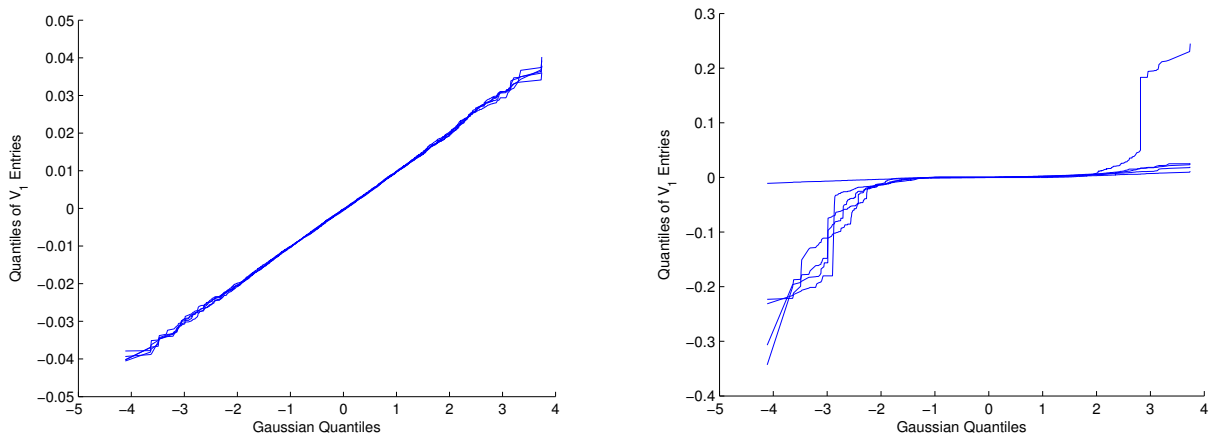
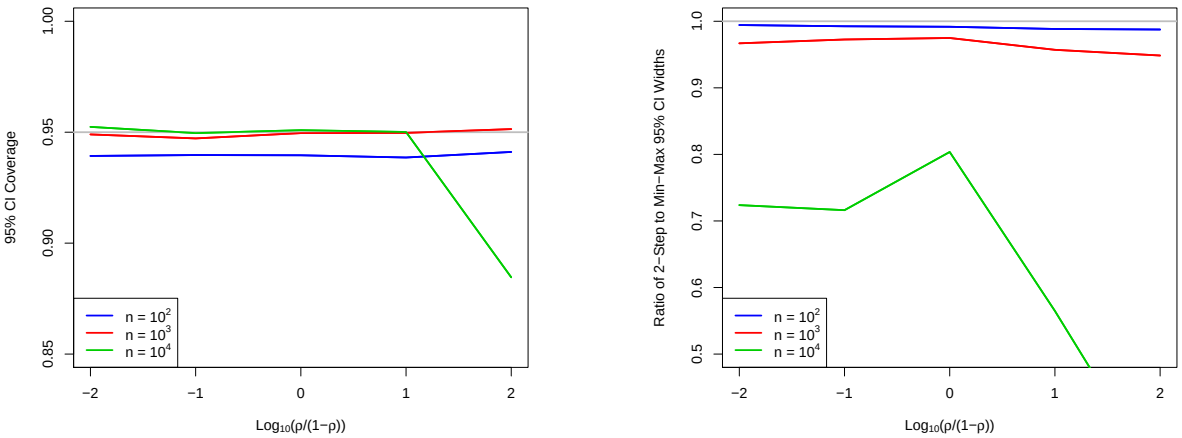


Figure 9: The first column of plots ((a), (c), (e)) generates $\beta_i \overset{i.i.d.}{\sim} N(0,1)$ and renormalizes to control θ^2, while the second column of plots ((b), (d), (f)) does the same but then sets 99% of the β_i to zero before renormalizing. The first two rows of plots ((a), (b), (c), (d)) use X_{ij} i.i.d. from some non-Gaussian distribution renormalized to have mean 0 and variance 1. The third row of plots ((e), (f)) uses marginally standard Gaussian $\boldsymbol{X}$ but with correlations among the columns; see Appendix G for detailed constructions. See text for detailed boxplot constructions and interpretation of the dashed line.



(a) (b)

Figure 10: Quantile-quantile plots measuring the Gaussianity of 5 realizations of the entries of $\boldsymbol{V}_1$ for (a) Bernoulli(0.1)-distributed $\boldsymbol{X}$ and (b) Bernoulli(0.001)-distributed $\boldsymbol{X}$.



(a) (b)

Figure 11: (a) Coverage and (b) width relative to EigenPrism of 2-step confidence intervals for $p = 10^4$ and $\theta^2 + \sigma^2 = 10^4$ across a range of ρ and n, with each point representing 10^4 simulations. Grey lines show (a) nominal coverage and (b) reference ratio of 1.

indeed produces shorter CIs than the EigenPrism procedure, it does not achieve nominal coverage with the same consistency, as shown in Figure 11.

There are two particularly surprising aspects of this plot. The first is that the 2-step procedure produces substantial gains in width even for ρ values near 0 and 1. This is surprising because the upper-bound (2.7) that is eliminated by the 2-step procedure is tight when ρ is nearly 0 or 1, however it is still not exact. The slightly loose variance upper bound turns out to have an optimizing $\boldsymbol{w}$ that is substantially different from the exact variance formula. The second surprising feature is that the width improvement is in fact *smallest* for ρ not near the endpoints 0 or 1. This can be explained by the clipping at 0. For $\rho \approx 0$, most CIs, both EigenPrism and 2-step, are cut

nearly in half by clipping, so the fractional width improvement achieved by the 2-step procedure is fully realized. For $\rho \approx 1$, both intervals are rarely clipped, and again the 2-step procedure realizes its full width improvement. However, for ρ not close to 0 or 1, many EigenPrism CIs are only slightly shrunk by clipping, so that the shorter 2-step intervals shorten the right side of the interval but leave the unclipped left side about the same, so that much less than the full width improvement is realized.

Although the 2-step procedure can provide substantial gains in width, it loses the robustness of the EigenPrism procedure, as shown in the slight undercoverage for $n = 100$ and the substantial undercoverage for large ρ and $n = p$. Therefore, in practice, we recommend use of the 2-step procedure instead of the EigenPrism procedure when $n \not\approx p$ or when the statistician is confident that ρ is not close to 1.

5 Variance Decomposition in the Northern Finland Birth Cohort

We now briefly show the result of applying EigenPrism to a dataset of SNPs and continuous phenotypes to perform inference on the SNR $= \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2/(\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 + \sigma^2)$. The data we use comes from the Northern Finland Birth Cohort 1966 (NFBC1966) (Sabatti et al., 2009; Järvelin et al., 2004), made available through the dbGaP database (accession number phs000276.v2.p1). The data consists of 5402 SNP arrays from subjects born in Northern Finland in 1966, as well as a number of phenotype variables measured when the subjects were 31 years old. After cleaning and processing the data (the details of which are provided in Appendix H), 328,934 SNPs remained. The resulting $5402 \times 328,934$ design matrix $\boldsymbol{X}$ contained approximately 58% 0's (homozygous wild type), 34% 1's (heterozygous), and 8% 2's (homozygous minor allele).

In order to use EigenPrism directly, we would need to know $\boldsymbol{\Sigma}$, as simply using $\boldsymbol{I}_p$ presents two possible problems:

(1) If $\boldsymbol{X}$ is not whitened before taking the SVD, the columns of $\boldsymbol{V}$ may be far from Haar-distributed, rendering our bias and variance computations incorrect.

(2) If $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, then the ostensible target of our procedure is $\|\boldsymbol{\beta}\|_2^2/(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)$, which may differ substantially from SNR $= \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2/(\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 + \sigma^2)$.

Unfortunately, the problem of estimating the covariance matrix of a SNP array is extremely challenging (and the subject of much current research) due to the fact that $n \ll p$, even if we use outside data, so we prefer to avoid it here. In order to simply treat the covariance matrix as diagonal, we must consider the two problems above. There is a widely-held belief that the SNP locations that are important for any given trait are relatively rare (see, for example, Yang et al. (2010); Golan and Rosset (2011)), and thus spaced far enough apart on the genome to be treated as independent. This precludes problem (2) above, since with nonzero coefficients spaced far apart, we have $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2 \approx \|\boldsymbol{\beta}\|_2^2$ (we take the columns of $\boldsymbol{X}$ to be standardized, so the diagonal of $\boldsymbol{\Sigma}$ is all ones). For problem (1), we know that far apart SNPs are very nearly independent, so we may expect that the true $\boldsymbol{\Sigma}$ is roughly diagonal, and we already showed in Section 4.1 that the EigenPrism procedure is robust to some small unaccounted-for covariances when constructing CIs for $\|\boldsymbol{\beta}\|_2^2$. To ensure that problems (1) and (2) do not cause EigenPrism to break down, we perform a series of diagnostics before applying it to the real data.

Given the approximation of $\boldsymbol{\Sigma}$ as diagonal, we first performed a series of simulations to ensure EigenPrism's accuracy was not affected. Specifically, we ran the EigenPrism procedure (with adjustments described in the paragraph below) on artificially-constructed traits, but using the same

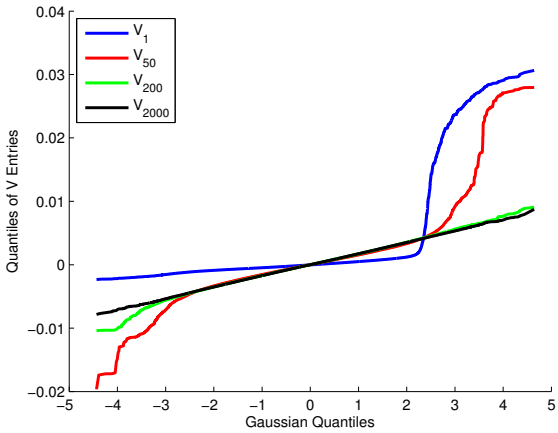


Figure 12: Distribution of the entries of some eigenvectors of the NFBC1966 design matrix.

standardized design matrix $\boldsymbol{X}$ from the NFBC1966 data set. For 20 different $\boldsymbol{\beta}$ vectors, we generated 500 independent Gaussian noise realizations and recorded the coverage of 95% EigenPrism CIs for SNR. The noise variance was 1, and the $\boldsymbol{\beta}$'s were chosen to have 300 nonzero entries with uniformly distributed positions and all nonzero entries equal to $\sqrt{0.3/((1-0.3)\cdot 300)}$ (so that SNR $= 0.3$ if $\boldsymbol{\Sigma} = \boldsymbol{I}_p$). Table 1 shows the coverage over the 20 $\boldsymbol{\beta}$'s, and they are indeed all quite close to 95%, even though this simulation was *conditional* on $\boldsymbol{X}$. Recomputing the target SNR using other estimates of $\boldsymbol{\Sigma}$, such as hard-thresholding the empirical covariance at 0.1, changed the value of SNR very little, so that coverage was largely unaffected.

Coverage	90%	91%	92%	93%	94%	95%	96%	97%	98%	99%	100%
Count	1	0	0	2	1	0	8	7	1	0	0

Table 1: Coverage of 20 SNR 95% CIs constructed for simulated traits using the NFBC1966 design matrix. Each coverage is an average over 500 simulations.

A second diagnostic was to examine the columns of $\boldsymbol{V}$ to check for Gaussianity related to the phenomenon mentioned in Section 4.1. Indeed, we find that some of the columns of $\boldsymbol{V}$ are quite non-Gaussian, as shown in Figure 12. However, this phenomenon is localized to only the columns of $\boldsymbol{V}$ corresponding to the very largest λ_i. Applying the unaltered EigenPrism procedure could cause two problems. First, if the the first columns of $\boldsymbol{V}$ are not Haar-distributed, T_2 could be biased and/or higher-variance than our theory accounts for. Second, recalling the interpretation of EigenPrism as a weighted regression of z_i^2 on λ_i, the fact that the problematic eigenvectors correspond to the largest eigenvalues means that they have high *leverage*, which exacerbates any unwanted bias or variance they create. Luckily, both problems can be remedied by running the EigenPrism SNR-estimation procedure (with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ after standardizing the columns of $\boldsymbol{X}$) with the added constraint to the optimization program in Equation (2.8) that the first entries of $\boldsymbol{w}$ are equal to zero. Explicitly, as the non-Gaussianity of the columns of $\boldsymbol{V}$ appears to dissipate after around the 100^{th} column, we set $w_1 = \cdots = w_{100} = 0$. The choice of 100 is somewhat subjective, but we tried other values and obtained very similar results. Because the resulting weights still obey the original constraints, the estimator of $\|\boldsymbol{\beta}\|_2^2$ remains unbiased and the variance upper-bounds remain valid. Although motivated by the diagnostics from Section 4.1, this adjustment has the added advantage of making the entire EigenPrism procedure completely independent of the first 100 rows of $\boldsymbol{U}$. It has been shown that the first rows of $\boldsymbol{U}$ are strongly related to the population

structure of the sample (for example, the first two principal components correspond closely with subjects' geographic origin), so constraining the first weights to be zero has the added effect of controlling for population structure (Price et al., 2006). As a final note, by subtracting off the means of each column of $\boldsymbol{X}$, we reduced $\boldsymbol{X}$'s rank by one, resulting in $\lambda_n = 0$. As this is not actually reflective of the distribution of $\boldsymbol{X}$, we also force $w_n = 0$ so that the last column of $\boldsymbol{V}$ and last row of $\boldsymbol{U}$ do not contribute to our estimate or inference.

Encouraged by the simulation results from Table 1, we proceeded to generate EigenPrism CIs for the SNRs of the 9 traits analyzed in Sabatti et al. (2009), as well as height (these 10 traits were also analyzed in Kang et al. (2010)). For each trait, transformation and subject exclusion was performed before computing SNR, following closely the procedures used in Sabatti et al. (2009); Kang et al. (2010) (see Appendix H for details). Lastly, all non-height phenotype values were adjusted for sex, pregnancy status, and oral contraceptive use, while height was only adjusted for sex. Table 2 gives

Phenotype Name	# Samples	SNR 95% CI (%)	Point Estimate (%)
Triglycerides	4644	[3.1, 29.3]	16.2
HDL cholesterol	4700	[17.1, 42.9]	30.0
LDL cholesterol	4682	[27.7, 53.6]	40.7
C-reactive protein	5290	[5.6, 28.8]	17.2
Glucose	4895	[4.0, 28.9]	16.5
Insulin	4867	[0.0, 21.5]	9.0
BMI	5122	[8.9, 32.8]	20.9
Systolic blood pressure	5280	[7.8, 31.0]	19.4
Diastolic blood pressure	5271	[7.4, 30.7]	19.0
Height	5306	[46.0, 69.1]	57.6

Table 2: CIs for heritability estimates for each of the 10 continuous phenotypes considered, along with the number of samples used for each.

the point estimate and 95% CI for the SNR of each phenotype, as well as the number of subjects used. Recall that these are CIs for the fraction of variance explained by the linear model consisting of the given array of SNPs. Still, these CIs generally agree quite well with heritability estimates in the literature (Kang et al., 2010). For instance, (Kang et al., 2010, Supplementary Information) reports two "pseudo-heritability" estimates of 73.8% and 62.5% for height, and 27.9% and 24.2% for BMI, on the same data set. This is somewhat remarkable given that they use a completely different statistical procedure with different assumptions. In particular, while other works in the heritability literature tend to treat $\boldsymbol{\beta}$ as random, EigenPrism was motivated by a simple model with $\boldsymbol{\beta}$ fixed and the rows of $\boldsymbol{X}$ random. We find this model more realistic, as true genetic effects are not in fact random, but fixed. One could argue the difference is not too important as long as the genetic effects are approximately distributed as the random effects model chosen, but such an assumption is impossible to verify in practice, as the true effects are never observed. EigenPrism's assumptions, on the other hand, are all on the design matrix, which is fully observed, leading to checks and diagnostics that can be performed to ensure the procedure will generate reasonable CIs.

6 Discussion

We have presented a framework for performing inference on the ℓ_2-norm of the coefficient vector in a linear regression model. Although the resulting confidence intervals are asymptotic, we show in extensive simulations that they achieve nominal coverage in finite samples, without making

any assumption on the structure or sparsity of the coefficient vector, or requiring knowledge of σ^2. In simulations, we are able to relax the restrictive assumptions on the distribution of the design matrix and gain an understanding of when our procedure is not appropriate. Applying this framework to performing inference on ℓ_2 regression error, noise level, and genetic signal-to-noise ratio, we develop new procedures in all three that are able to construct accurate CIs in situations not previously addressed in the literature.

This work leaves open numerous avenues for further study. We briefly introduced a 2-step procedure that provided substantially shorter CIs than the EigenPrism procedure, but had less-consistent coverage. If we could better understand that procedure or come up with diagnostics for when it would undercover, we could improve on the EigenPrism procedure. We also explored in simulation a number of model failures that our procedure was robust (or not) to, but further study could provide theoretical guarantees on the coverage of the EigenPrism procedure for a broader class of random design models. Section 4.1 also briefly alluded to improved robustness in a random effects framework, which we have not explored further here. Finally, although in this work we consider a statistic that is linear in the z_i^2, the framework and ideas of this work are not intimately tied to this restriction, and there may exist statistics that are nonlinear functions of the z_i^2 that give improved performance.

Acknowledgements

We owe a great deal of gratitude to Chiara Sabatti for her patience in explaining to us key concepts in statistical genetics and for her guidance. We also thank Art Owen for sharing his unpublished notes with us and for his constructive feedback, and Matthew Stephens and Xiang Zhu for their helpful discussions on covariance estimation of SNP data. L. J. was partially supported by NIH training grant T32GM096982. E. C. is partially supported by a Math + X Award from the Simons Foundation. The NFBC1966 Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the Broad Institute, UCLA, University of Oulu, and the National Institute for Health and Welfare in Finland. This manuscript was not prepared in collaboration with investigators of the NFBC1966 Study and does not necessarily reflect the opinions or views of the NFBC1966 Study Investigators, Broad Institute, UCLA, University of Oulu, National Institute for Health and Welfare in Finland and the NHLBI.

A Inference for θ^2 under non-Gaussian design with known variance

The method of Section 2.1 also works asymptotically under more general conditions than the Gaussianity assumptions of (1.4). Let $\boldsymbol{z} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the statement that $\boldsymbol{z}$ has some distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Consider again the linear model (1.1) but with relaxed assumptions,

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} (\boldsymbol{\mu}, \boldsymbol{I}_p - \boldsymbol{\mu}\boldsymbol{\mu}^\top), \qquad \varepsilon_i \overset{i.i.d.}{\sim} (0, \sigma^2),$$

again with σ^2 known and $\boldsymbol{X}$ independent of $\boldsymbol{\varepsilon}$. Under this model, we get that

$$y_i^2 \overset{i.i.d.}{\sim} (\theta^2 + \sigma^2, v_1)$$

and the asymptotic distribution in (2.1) in turn becomes, by the CLT,

$$\frac{1}{\sqrt{n}}\|\boldsymbol{y}\|_2^2 - \sqrt{n}\big(\theta^2 + \sigma^2\big) \xrightarrow{\mathcal{D}} N(0, v_1), \tag{A.1}$$

as $n \to \infty$, where v_1 does not depend on n but does depend on the unknown $\boldsymbol{\beta}$, and is given by

$$v_1 = \mathbb{E}\left[\varepsilon_i^4\right] + 4\sigma^2\Big(\theta^2 - \big(\boldsymbol{\mu}^\top\boldsymbol{\beta}\big)^2\Big) + 4\mathbb{E}\left[\varepsilon_i^3\right]\boldsymbol{\mu}^\top\boldsymbol{\beta} + \mathbb{E}\left[\big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\big)^4\right] - \sigma^4 - \theta^4 - \big(\boldsymbol{\mu}^\top\boldsymbol{\beta}\big)^4 + 2\theta^2\big(\boldsymbol{\mu}^\top\boldsymbol{\beta}\big)^2$$

In order to be less parametric, we can consider bootstrap confidence intervals based on the above calculations. Corresponding to (A.1) we can get an unbiased statistic,

$$\begin{aligned} T_1 &:= \frac{1}{n}\|\boldsymbol{y}\|_2^2 - \sigma^2, \\ \mathbb{E}\left[T_1\right] &= \theta^2, \\ \mathrm{SD}(T_1) &= \sqrt{v_1/n}, \end{aligned} \tag{A.2}$$

whose distribution we may hope to be close to Gaussian. T_1 can be bootstrapped (potentially with standard bias-correction and acceleration) to obtain bootstrap CIs, nonparametrically dealing with the unknown variance v_1. We ran simulations with $n = 800$, $p = 1500$, $\boldsymbol{X}$ having i.i.d. Bernoulli(0.05) entries (the columns of $\boldsymbol{X}$ were then standardized to have mean 0 and variance 1), $\theta^2 = \sigma^2 = 10$, and ε_i i.i.d. t_5 (rescaled to have variance 10). We generated a single $\boldsymbol{\beta}$ uniformly on the θ-radius sphere and ran 1000 simulations (so that $\boldsymbol{\beta}$ did not change across simulations). Bias-corrected, accelerated 95% bootstrap CIs achieved 93.8% coverage (this is within statistical uncertainty of the nominal 95%, as a 95% CI for the CI coverage is $[0.923, 0.953]$).

B Calculation of variance of EigenPrism estimator

In this section we calculate the variance of the statistic $S = \sum_{i=1}^n w_i z_i^2$ when conditioning on $\boldsymbol{d}$. Here we treat $\boldsymbol{w}$ as fixed, but note that since we condition on $\boldsymbol{d}$, this includes values of $\boldsymbol{w}$ that are calculated as a function of $\boldsymbol{d}$, as in the EigenPrism method.

$$\begin{aligned} \mathrm{Var}(S|\boldsymbol{d}) &= \mathrm{Var}\left(\sum_{i=1}^n w_i z_i^2 \,\middle|\, \boldsymbol{d}\right) \\ &= \sum_{i=1}^n w_i^2 \,\mathrm{Var}\left(z_i^2 \middle|\, \boldsymbol{d}\right) + \sum_{\substack{i,j=1 \\ i\neq j}}^n w_i w_j \,\mathrm{Cov}\left(z_i^2, z_j^2 \middle|\, \boldsymbol{d}\right) \ . \end{aligned}$$

We now calculate each term. Recall that $\lambda_i := d_i^2/p$ for $i = 1, \ldots, n$. Then

$$\begin{aligned} \mathrm{Var}\left(z_i^2 \middle|\, \boldsymbol{d}\right) &= \mathbb{E}\left[z_i^4 \middle|\, \boldsymbol{d}\right] - \mathbb{E}\left[z_i^2 \middle|\, \boldsymbol{d}\right]^2 \\ &= \mathbb{E}\left[(d_i\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle + \epsilon_i)^4 \middle|\, \boldsymbol{d}\right] - \left(\lambda_i\theta^2 + \sigma^2\right)^2 \end{aligned}$$

Using $\epsilon_i \sim N(0, \sigma^2)$ (and the fact that $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{V}$),

$$= \mathbb{E}\left[(d_i\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle)^4 \middle|\, \boldsymbol{d}\right] + 6\sigma^2\mathbb{E}\left[(d_i\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle)^2 \middle|\, \boldsymbol{d}\right] + 3\sigma^4 - \left(\lambda_i\theta^2 + \sigma^2\right)^2$$

Using $\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2 \sim \theta^2 \cdot \mathsf{Beta}\left(\frac{1}{2}, \frac{p-1}{2}\right)$,

$$\begin{aligned} &= d_i^4\theta^4 \cdot \frac{1\cdot 3}{p\cdot(p+2)} + 6\sigma^2\lambda_i\theta^2 + 3\sigma^4 - \left(\lambda_i\theta^2 + \sigma^2\right)^2 \\ &= 2\lambda_i^2\theta^4\frac{p-1}{p+2} + 4\sigma^2\lambda_i\theta^2 + 2\sigma^4 \end{aligned}$$

Also, for $i \neq j$,

$$\begin{aligned}\operatorname{Cov}\left(z_i^2, z_j^2 \middle| \boldsymbol{d}\right) &= \mathbb{E}\left[z_i^2 z_j^2 \middle| \boldsymbol{d}\right] - \mathbb{E}\left[z_i^2 \middle| \boldsymbol{d}\right]\mathbb{E}\left[z_j^2 \middle| \boldsymbol{d}\right] \\ &= \mathbb{E}\left[(d_i\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle + \epsilon_i)^2 (d_j\langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle + \epsilon_j)^2 \middle| \boldsymbol{d}\right] - \left(\lambda_i\theta^2 + \sigma^2\right)\left(\lambda_j\theta^2 + \sigma^2\right)\end{aligned}$$

Using $\epsilon_i, \epsilon_j \overset{\text{iid}}{\sim} N(0, \sigma^2)$ (and the fact that $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{V}$),

$$\begin{aligned}&= \mathbb{E}\left[d_i^2 d_j^2 \langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2 \langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle^2 \middle| \boldsymbol{d}\right] + \sigma^2\mathbb{E}\left[d_i^2\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2 \middle| \boldsymbol{d}\right] + \sigma^2\mathbb{E}\left[d_j^2\langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle^2 \middle| \boldsymbol{d}\right] \\ &\qquad + \sigma^4 - \left(\lambda_i\theta^2 + \sigma^2\right)\left(\lambda_j\theta^2 + \sigma^2\right)\end{aligned}$$

Using $\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2, \langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle^2 \sim \theta^2 \cdot \mathsf{Beta}\left(\frac{1}{2}, \frac{p-1}{2}\right)$,

$$\begin{aligned}&= \mathbb{E}\left[d_i^2 d_j^2 \langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2 \langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle^2 \middle| \boldsymbol{d}\right] + \sigma^2\lambda_i\theta^2 + \sigma^2\lambda_j\theta^2 + \sigma^4 \\ &\qquad - \left(\lambda_i\theta^2 + \sigma^2\right)\left(\lambda_j\theta^2 + \sigma^2\right) \\ &= d_i^2 d_j^2 \mathbb{E}\left[\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2 \langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle^2 \middle| \boldsymbol{d}\right] - \lambda_i\lambda_j\theta^4 \\ &= d_i^2 d_j^2 \operatorname{Cov}\left(\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2, \langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle^2 \middle| \boldsymbol{d}\right)\end{aligned}$$

Using $\left(\langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2, \langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle^2, \theta^2 - \langle \boldsymbol{V}_i, \boldsymbol{\beta}\rangle^2 - \langle \boldsymbol{V}_j, \boldsymbol{\beta}\rangle^2\right) \sim \theta^2 \cdot \mathsf{Dirichlet}\left(\frac{1}{2}, \frac{1}{2}, \frac{p-2}{2}\right)$,

$$= \frac{-2}{p+2}\lambda_i\lambda_j\theta^4.$$

Then,

$$\begin{aligned}\operatorname{Var}\left(S \middle| \boldsymbol{d}\right) &= \sum_{i=1}^n w_i^2\left(2\lambda_i^2\theta^4\frac{p-1}{p+2} + 4\sigma^2\lambda_i\theta^2 + 2\sigma^4\right) + \sum_{\substack{i,j=1 \\ i\neq j}}^n w_i w_j\left(\frac{-2}{p+2}\lambda_i\lambda_j\theta^4\right) \\ &= \sum_{i=1}^n w_i^2\left(2\lambda_i^2\theta^4\frac{p}{p+2} + 4\sigma^2\lambda_i\theta^2 + 2\sigma^4\right) + \sum_{i=1}^n\sum_{j=1}^n w_i w_j\left(\frac{-2}{p+2}\lambda_i\lambda_j\theta^4\right) \\ &= 2\sigma^4\sum_{i=1}^n w_i^2 + 4\sigma^2\theta^2\sum_{i=1}^n w_i^2\lambda_i + 2\theta^4\left(\frac{p}{p+2}\sum_{i=1}^n w_i^2\lambda_i^2 - \frac{\left(\sum_{i=1}^n w_i\lambda_i\right)^2}{p+2}\right).\end{aligned}$$

C Variance upper-bound for T_2

In this section we derive the upper bound (2.12) on the variance of the statistic T_2. For simplicity we assume that n is even.

We begin by constructing a vector of weights $\tilde{\boldsymbol{w}}$:

$$\tilde{w}_i := \frac{1}{\sum_{j=1}^{n/2}\lambda_j - \sum_{j=n/2+1}^n \lambda_j} \cdot \begin{cases} +1, & \text{for } i \leq n/2, \\ -1, & \text{for } i > n/2. \end{cases}$$

Note that $\tilde{\boldsymbol{w}}$ satisfies the constraints of the optimization problem (2.8), and thus $\mathrm{Var}(T_2)$ is upper-bounded by Equation (2.7) with $\tilde{\boldsymbol{w}}$ plugged in. A second key observation is that we know from random matrix theory that for $n, p \to \infty$ and $n/p \to \gamma \in (0,1)$, the distribution of rescaled eigenvalues, λ_i, converges to the MP distribution with parameter γ.

Recalling the definitions of A_γ, B_γ given in (2.11), this implies that

$$\frac{1}{n}\sum_{i=1}^{n} \lambda_i \cdot (\mathbb{1}_{i\le n/2} - \mathbb{1}_{i>n/2}) \to A_\gamma$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \to B_\gamma \,.$$

Together with the definition of $\tilde{\boldsymbol{w}}$, these imply that as $n, p \to \infty$ with $n/p \to \gamma \in (0,1)$,

$$\begin{aligned}
n\sum_{i=1}^{n} \tilde{w}_i^2 &= \frac{n\cdot n}{\left(n\cdot\left[\frac{1}{n}\sum_{i=1}^{n}\lambda_i\cdot(\mathbb{1}_{i\le n/2}-\mathbb{1}_{i>n/2})\right]\right)^2} \to \frac{n\cdot n}{(nA_\gamma)^2} = \frac{1}{A_\gamma^2},\\
n\sum_{i=1}^{n} \tilde{w}_i^2\lambda_i^2 &= \frac{n\cdot n\cdot\left[\frac{1}{n}\sum_{i=1}^{n}\lambda_i^2\right]}{\left(n\cdot\left[\frac{1}{n}\sum_{i=1}^{n}\lambda_i\cdot(\mathbb{1}_{i\le n/2}-\mathbb{1}_{i>n/2})\right]\right)^2} \to \frac{n\cdot n\cdot B_\gamma}{(nA_\gamma)^2} = \frac{B_\gamma}{A_\gamma^2},
\end{aligned}$$

which in turn implies

$$\sqrt{n}\cdot\frac{\sqrt{\mathrm{Var}(T_2)}}{\theta^2+\sigma^2} \le \sqrt{2\cdot\max\left\{n\sum_{i=1}^{n}\tilde{w}_i^2, n\sum_{i=1}^{n}\tilde{w}_i^2\lambda_i^2\right\}} \to \sqrt{2}\cdot\max\left\{\frac{1}{A_\gamma}, \frac{\sqrt{B_\gamma}}{A_\gamma}\right\}. \tag{C.1}$$

D Proof of Theorem 1

Proof. For this proof, we use Le Cam's method (see e.g. Yu (1997, Lemma 1)), which states that

$$\mathbb{P}_{\boldsymbol{Z}\sim P_0}\{\psi(\boldsymbol{Z})=1\} + \mathbb{P}_{\boldsymbol{Z}\sim P_1}\{\psi(\boldsymbol{Z})=0\} \ge 1 - \|P_0 - P_1\|_{\mathsf{TV}} \,,$$

where $\|\cdot\|_{\mathsf{TV}}$ is the total variation norm:

$$\|P_0 - P_1\|_{\mathsf{TV}} = \sup_{\mathcal{A}\subseteq\mathbb{R}^n} \left|\mathbb{P}_{\boldsymbol{Z}\sim P_0}\{\boldsymbol{Z}\in\mathcal{A}\} - \mathbb{P}_{\boldsymbol{Z}\sim P_1}\{\boldsymbol{Z}\in\mathcal{A}\}\right| \,,$$

where the supremum is taken over Lebesgue-measurable sets.

We begin by constructing a related distribution Q_1:

$$\boldsymbol{W} = \theta\cdot\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{a}\cdot r + \sigma\cdot\boldsymbol{\varepsilon} \,, \tag{D.1}$$

where $\theta = \sigma = \frac{1}{\sqrt{2}}$, and where $r \sim \chi_p/\sqrt{p}$ is independent from $\boldsymbol{V}, \boldsymbol{\varepsilon}$. We will bound

$$\|P_0 - P_1\|_{\mathsf{TV}} \le \|P_0 - Q_1\|_{\mathsf{TV}} + \|P_1 - Q_1\|_{\mathsf{TV}} \,.$$

First, we use the fact that r concentrates tightly near 1 for the following bound:

$$\begin{aligned}
\mathbb{E}\left[\left\|\theta\cdot\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{a}\cdot r - \theta\cdot\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{a}\right\|_2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\theta\cdot\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{a}\cdot r - \theta\cdot\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{a}\right\|_2 \,\middle|\, r, \boldsymbol{V}\right]\right]\\
&= \mathbb{E}\left[\left\|\theta\cdot\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{a}\right\|_2\cdot|r-1|\right]\\
&\le \theta\cdot\sqrt{\mathbb{E}\left[\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top\boldsymbol{a}\right]}\cdot\sqrt{\mathbb{E}\left[(r-1)^2\right]}\\
&= \theta\cdot\sqrt{\mathbb{E}\left[\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top\boldsymbol{a}\right]}\cdot\frac{1}{\sqrt{p}}\sqrt{\chi_p^2 - 2\sqrt{p}\cdot\mathbb{E}\left[\chi_p\right] + p}
\end{aligned}$$

Using the fact that $\mathbb{E}\left[\chi_p^2\right] = p$ and $\mathbb{E}\left[\chi_p\right] \geq \sqrt{p} - \frac{1}{4\sqrt{p}}$, and that $\mathbb{E}\left[(\boldsymbol{V}_i^\top \boldsymbol{a})^2\right] = \frac{1}{p}$ for each $i = 1, \ldots, n$,

$$\begin{aligned}\mathbb{E}\left[\left\|\theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a} \cdot r - \theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a}\right\|_2\right] &\leq \theta \cdot \sqrt{\sum_i \frac{D_{ii}^2}{p}} \cdot \frac{1}{\sqrt{2p}} \\ &= \frac{\theta\sqrt{n}}{\sqrt{2p}}\,,\end{aligned}$$

since $\frac{1}{p}\sum_i D_{ii}^2 = n$. Next, for any measurable set $\mathcal{A} \subseteq \mathbb{R}^n$, we have

$$\begin{aligned}&\left|\mathbb{P}_{\boldsymbol{W}\sim Q_1}\left\{\boldsymbol{W} \in \mathcal{A}\right\} - \mathbb{P}_{\boldsymbol{Z}\sim P_1}\left\{\boldsymbol{Z} \in \mathcal{A}\right\}\right| \\ &= \left|\mathbb{E}\left[\mathbb{P}\left\{\boldsymbol{W} \in \mathcal{A} \mid r, \boldsymbol{V}\right\} - \mathbb{P}\left\{\boldsymbol{Z} \in \mathcal{A} \mid r, \boldsymbol{V}\right\}\right]\right| \\ &\leq \mathbb{E}\left[\left|\mathbb{P}\left\{\boldsymbol{W} \in \mathcal{A} \mid r, \boldsymbol{V}\right\} - \mathbb{P}\left\{\boldsymbol{Z} \in \mathcal{A} \mid r, \boldsymbol{V}\right\}\right|\right] \\ &= \mathbb{E}\left[\left|\mathbb{P}\left\{\theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a} \cdot r + \sigma \cdot \boldsymbol{\varepsilon} \in \mathcal{A} \;\middle|\; r, \boldsymbol{V}\right\} - \mathbb{P}\left\{\theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a} + \sigma \cdot \boldsymbol{\varepsilon} \in \mathcal{A} \;\middle|\; r, \boldsymbol{V}\right\}\right|\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\left\|N\left(\theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a} \cdot r, \sigma^2 \mathbf{I}_n\right) - N\left(\theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a}, \sigma^2 \mathbf{I}_n\right)\right\|_{\mathsf{TV}} \;\middle|\; r, \boldsymbol{V}\right]\right]\end{aligned}$$

Using the fact that $\left\|N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n) - N(\boldsymbol{\mu}', \sigma^2 \mathbf{I}_n)\right\|_{\mathsf{TV}} \leq \frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2}{\sqrt{2\pi\sigma^2}}$ for any fixed $\boldsymbol{\mu}, \boldsymbol{\mu}', \sigma^2$,

$$\begin{aligned}&\leq \mathbb{E}\left[\mathbb{E}\left[\frac{\left\|\theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a} \cdot r - \theta \cdot \boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{a}\right\|_2}{\sqrt{2\pi\sigma^2}} \;\middle|\; r, \boldsymbol{V}\right]\right] \\ &\leq \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{\theta\sqrt{n}}{\sqrt{2p}}\,,\end{aligned}$$

where the last step uses our calculations above. Since this is true for any $\mathcal{A} \subset \mathbb{R}^n$, and since $\theta = \sigma = \frac{1}{\sqrt{2}}$ by assumption under the distribution P_1, we have

$$\|P_1 - Q_1\|_{\mathsf{TV}} \leq \sqrt{\frac{n/p}{4\pi}}\,.$$

Next, we bound $\|P_0 - Q_1\|_{\mathsf{TV}}$. By Pinsker's inequality,

$$\|P_0 - Q_1\|_{\mathsf{TV}} \leq \sqrt{\frac{1}{2}\mathsf{KL}\left(Q_1 \| P_0\right)}\,,$$

where $\mathsf{KL}\left(\cdot\|\cdot\right)$ is the Kullback-Leibler divergence. Note that the distributions P_0 and Q_1 can be reformulated as

$$P_0 : Z_i \overset{\perp\!\!\!\perp}{\sim} N(0, 1)$$

and

$$Q_1 : Z_i \overset{\perp\!\!\!\perp}{\sim} N\left(0, \frac{\lambda_i + 1}{2}\right)\,.$$

Writing $p_0(\cdot)$ and $q_1(\cdot)$ to be the densities of the distributions P_0 and Q_1, respectively, we have

$$\begin{aligned}
\mathsf{KL}\left(Q_1 \| P_0\right) &= \mathbb{E}_{\boldsymbol{Z}\sim Q_1}\left[\log\left(\frac{q_1(\boldsymbol{Z})}{p_0(\boldsymbol{Z})}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}\sim Q_1}\left[\log\left(\frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\left(\frac{\lambda_i+1}{2}\right)}} e^{-\frac{Z_i^2}{\lambda_i+1}}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{Z_i^2}{2}}}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{Z}\sim Q_1}\left[-\frac{1}{2}\sum_i\left[\log\left(\frac{\lambda_i+1}{2}\right) + Z_i^2\cdot\left(\frac{2}{\lambda_i+1}-1\right)\right]\right]
\end{aligned}$$

Since $\mathbb{E}_{\boldsymbol{Z}\sim Q_1}\left[Z_i^2\right] = \frac{\lambda_i+1}{2}$,

$$= -\frac{1}{2}\sum_i\left[\log\left(\frac{\lambda_i+1}{2}\right) + \left(1-\frac{\lambda_i+1}{2}\right)\right]$$

Using the fact that $\log(x) \geq (x-1) - 2(x-1)^2$ for all $x \geq \frac{1}{2}$,

$$\begin{aligned}
&\leq -\frac{1}{2}\sum_i\left[\frac{\lambda_i-1}{2} - 2\left(\frac{\lambda_i-1}{2}\right)^2 + \left(1-\frac{\lambda_i+1}{2}\right)\right] \\
&= \frac{1}{4}\sum_i(\lambda_i-1)^2\,.
\end{aligned}$$

Combining everything, we have

$$\|P_0 - Q_1\|_{\mathsf{TV}} \leq \sqrt{\frac{1}{2}\mathsf{KL}\left(Q_1\|P_0\right)} \leq \sqrt{\frac{1}{2}\cdot\frac{1}{4}\sum_i(\lambda_i-1)^2}\,,$$

and so

$$\|P_0 - P_1\|_{\mathsf{TV}} \leq \sqrt{\frac{1}{8}\sum_i(\lambda_i-1)^2} + \sqrt{\frac{n/p}{4\pi}}\,.$$

□

E CVX code for computing the weight vector

The following snippet of code was used with MATLAB Version 8.1 (R2013a) and CVX Version 2.1, Build 1085 on a 64-bit Linux OS. The eigenvalues λ_i are represented by the column vector `lambda`, `t` corresponds to $2\,\mathrm{val}(\mathcal{P}_1)$, and the resulting vector `w` corresponds to $\boldsymbol{w}^*$.

```
cvx_begin
variable t
variable w(n)
minimize t
subject to
```

```
    sum(w) == 0;
    sum(w .* lambda) == 1;
    norm([w; (t/2-1)/2])  <= (t/2+1)/2;
    norm([w .* lambda; (t/2-1)/2]) <= (t/2+1)/2;
cvx_end
```

F Bayesian model

The Bayesian model is given explicitly as follows (M, $\boldsymbol{Z}$, σ^2, and $\boldsymbol{\varepsilon}$ are all independent of one another):

$$\begin{aligned} M &\sim \text{Exponential}(\lambda), \\ \boldsymbol{Z} &\sim N(0, I_p), \\ \boldsymbol{\beta} &= \sqrt{M}\boldsymbol{Z}, \\ \frac{1}{\sigma^2} &\sim \text{Gamma}(A, B), \\ \boldsymbol{\varepsilon} &\sim N(0, \sigma^2 I_p), \\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \end{aligned} \tag{F.1}$$

where the values for the parameters used were $\lambda = \frac{50{,}000}{p}$ (so $\theta^2 \approx \text{Exponential}(1/2000)$), $A = 14$, $B = \frac{1}{20{,}000}$, and we have used the shape/scale parameterization of the Gamma distribution, as opposed to the shape/rate parameterization. Figure 13 shows the resulting priors for θ^2, σ^2, and $\rho = \frac{\theta^2}{\theta^2+\sigma^2}$.

Note also that, although not shown, the posteriors achieved under this setup were all unimodal, so that the equal-tailed credible intervals were very close to the minimum-length credible intervals. We used equal-tailed credible intervals to give fair comparison with the EigenPrism CIs, which are also equal-tailed. The interval widths plotted all have nominal coverage of 80%. BCI endpoints were estimated by empirical quantiles of posterior draws from a Gibbs sampler, and thus we were able to much more accurately estimate the 10th and 90th percentiles than, say, the 2.5th and 97.5th percentiles.

G Construction of correlated-column covariance matrices

Dense 10% Correlations used a covariance matrix with ones on the diagonal and 0.1's as all the other entries. The Sparse $100 \cdot P\%$ Correlations used alternating P and $-P$ as off-diagonal entries in a correlation matrix, then projected that matrix into the positive semidefinite cone and reset the diagonal entries to 1. The resulting matrix has approximately 1/4 of its entries equal to P, 1/2 of its entries equal to 0, and 1/4 of its entries equal to $-2 \times 10^{-4} \cdot (1 - P)$.

H Processing of NFBC1966 dataset

Genotype features from the original data set were removed if they met any of the following conditions:

- Not a SNP (some were, e.g., copy number variations)
- Greater than 5% of values were missing

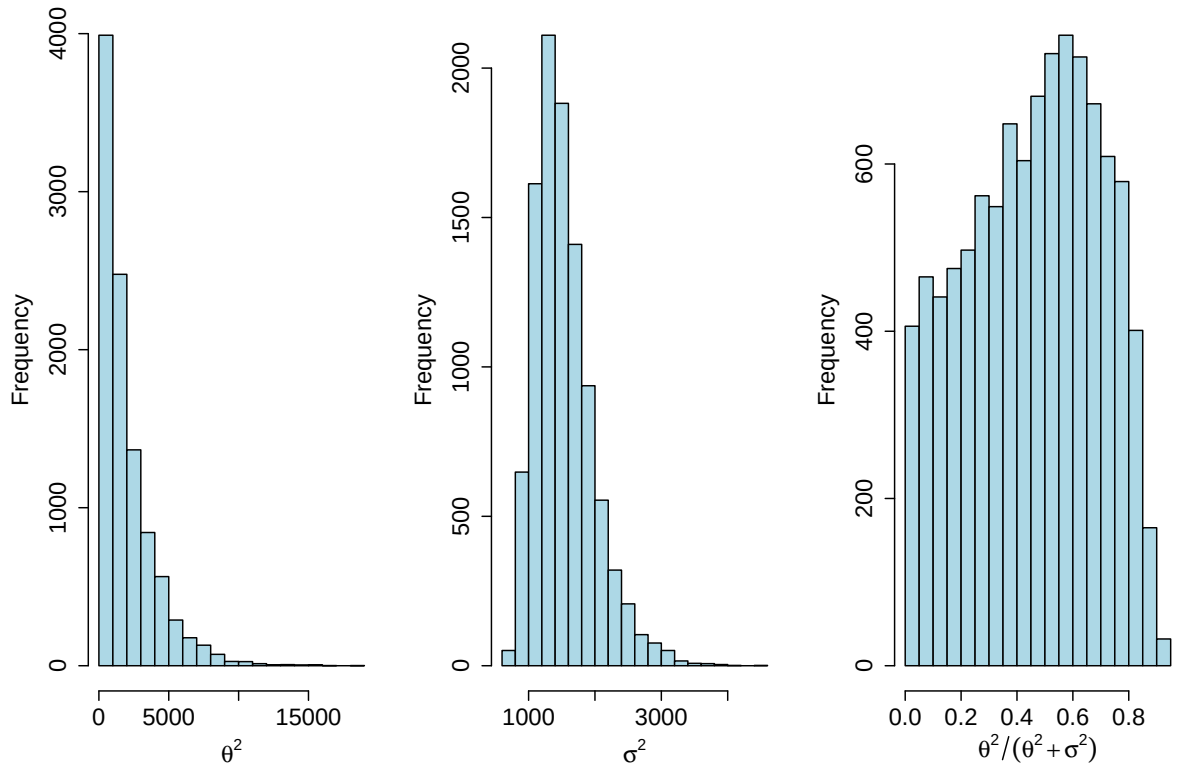


Figure 13: Priors from model (F.1).

- All nonmissing values belonged to the same nucleotide
- SNP location could not be aligned to the genome
- A χ^2 test rejected Hardy-Weinberg equilibrium at the 0.01% level
- On chromosome 23 (sex chromosome)

The remaining missing values were assumed to take the major allele value (thus were coded as 0's in the pre-centered design matrix).

For each trait, further processing was performed on the subjects. Triglycerides, BMI, insulin, and glucose were all log-transformed. C-reactive protein was also log-transformed after adding 0.002 mg/l (half the detection limit) to 0 values. Subjects were excluded from the triglycerides, HDL and LDL cholesterol, glucose, and insulin analyses if they were on diabetic medication or had not fasted before blood collection (or if either value was missing). Further subjects were excluded from the triglycerides, HDL and LDL cholesterol analyses if they were pregnant or if their respective phenotype measurement was more than three standard deviations from the mean, after correcting for sex, oral contraceptive use, and pregnancy. Subjects whose weight was not directly measured were excluded from BMI analysis. Of course any missing values in each phenotype were also excluded.

References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. a., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. a. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Bai, Z. D., Miao, B. Q., and Pan, G. M. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.*, 35(4):1532–1572.

Bayati, M., Erdogdu, M., and Montanari, A. (2013). Estimating lasso risk and noise level. *Advances in Neural Information Processing Systems*, pages 1–9.

Benjamini, Y. and Yu, B. (2013). The shuffle estimator for explainable variance in fmri experiments. *Ann. Appl. Stat.*, 7(4):2007–2033.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–837.

Bonnet, A., Gassiat, E., and Lévy-Leduc, C. (2014). Heritability estimation in high dimensional linear mixed models. *arXiv preprint arXiv:1404.3397*.

Candès, E., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509.

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society. Series B*, pages 37–65.

Golan, D. and Rosset, S. (2011). Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics (Oxford, England)*, 27(13):i317–23.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H., editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited.

Grant, M. and Boyd, S. (2014). {CVX}: Matlab Software for Disciplined Convex Programming, version 2.1. \url{http://cvxr.com/cvx}.

Järvelin, M.-R., Sovio, U., King, V., Lauren, L., Xu, B., McCarthy, M. I., Hartikainen, A.-L., Laitinen, J., Zitting, P., Rantakallio, P., and Elliott, P. (2004). Early life factors and blood pressure at age 31 years in the 1966 northern finland birth cohort. *Hypertension*, 44(6):838–846.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Lee, J., Sun, D., Sun, Y., and Taylor, J. (2015). Exact post-selection inference, with application to the lasso. *arXiv preprint arXiv:1311.6238*.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.*, 42(2):413–468.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., and Chakravarti, A. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Marčenko, V. and Pastur, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 457.

Owen, A. (2012). Quasi-regression for heritability. (March):1–13.

Owen, A. (2014). personal communication.

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

Pritchard, J. (2001). Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, (1):124–137.

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., McCarthy, M. I., Daly, M. J., Jrvelin, M.-R., Freimer, N. B., and Peltonen, L. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35–46.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., de Lange, M., Harris, J. R., Hjelmborg, J. V. B., Luciano, M., Martin, N. G., Mortensen, J., Nisticò, L., Pedersen, N. L., Skytthe, A., Spector, T. D., Stazi, M. A., Willemsen, G., and Kaprio, J. (2003). Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. *Twin Research and Human Genetics*, 6(05):399–408.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). 1-Penalization for Mixture Regression Models. *Test*, 19(2):209–256.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Taylor, J., Lockhart, R., Tibshirani, R., and Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*.

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

van de Geer, S., Bhlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics eraconcepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

Ward, R. (2009). Compressed sensing with cross validation. *Information Theory, IEEE Transactions on*, 55(12):5773–5782.

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R. B., Stevens, S., and Hall, A. S. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*, 40(5):575–583.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., and Others (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569.

Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.