Robust Inference Under Heteroskedasticity via the Hadamard Estimator

Zhixiang Zhang, Yachong Yang, Weijie J. Su, and Edgar Dobriban*

Abstract

Drawing statistical inferences from large datasets in a model-robust way is an important problem in statistics and data science. In this paper, we propose methods that are robust to large and unequal noise in different observational units (i.e., heteroskedasticity) for statistical inference in linear regression. We leverage the *Hadamard estimator*, which is unbiased for the variances of ordinary least-squares regression. This is in contrast to the popular White's sandwich estimator, which can be substantially biased in high dimensions. We propose to estimate the signal strength, noise level, signal-to-noise ratio, and mean squared error via the Hadamard estimator. We also develop inference on the coordinates of linear regression parameter and its quadratic functionals. We propose a new degrees of freedom adjustment that gives more accurate confidence intervals than variants of White's sandwich estimator. Moreover, we provide conditions ensuring the estimator is well-defined, by studying a new random matrix ensemble in which the entries of a random orthogonal projection matrix are squared. We also show approximate normality of the Hadamard estimators. Our work provides improved statistical theory and methods for linear regression in high dimensions.

Keywords: linear regression, heteroskedasticity, high dimensions, White's sandwich estimator, MacKinnon-White estimator, Hadamard product

Contents

*Author affiliations: Department of Mathematics, University of Macau (ZZ), Department of Statistics and Data Science, University of Pennsylvania (ED, WJS, YY). E-mail addresses: zhixzhang@um.edu.mo, yachong@wharton.upenn.edu, suw@wharton.upenn.edu, dobriban@wharton.upenn.edu.

1 Introduction

The linear regression model

$$Y = X\beta + \varepsilon \tag{1}$$

is widely used in many areas. The goal is to understand the dependence of an outcome variable Y on some p covariates $x = (x_1, \ldots, x_p)^\top$. We observe n such data points, arranging their outcomes into the $n \times 1$ vector Y, and their covariates into the $n \times p$ matrix X. We assume that Y depends linearly on X, via some unknown $p \times 1$ parameter vector β. The noise vector ε consists of independent random variables.

A fundamental practical problem is that the structure of noise ε affects the accuracy of inferences about the regression coefficient β. If the noise level in an observation is very high, that observation contributes little useful information. Such an observation could bias our inferences, and we should discard or down-weight it. The practical meaning of large noise is that our model underfits the specific observation. However, we usually do not know the noise level of each observation. Therefore, we must design procedures that adapt to unknown noise levels, for instance by constructing preliminary estimators of the noise. This problem of unknown and unequal noise levels, i.e., *heteroskedasticity*, has long been recognized as a central problem in many applied areas, especially in finance and econometrics.

In applied data analysis, and especially in the fields mentioned above, it is a common practice to use the ordinary least-squares (OLS) estimator $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ as the estimator of the unknown regression coefficients, despite the potential of heteroskedasticity. The OLS estimator is still unbiased, and has other desirable properties—such as consistency—under mild conditions. For statistical inference about β, the common practice is to use heteroskedasticity-robust confidence intervals.

Specifically, in the classical low-dimensional case when the dimension p is fixed and the sample size n grows, the OLS estimator is asymptotically normal with asymptotic covariance matrix $C_\infty = \lim_{n\to\infty} nC$, with

$$C = \mathrm{Cov}(\hat{\beta}) = (X^\top X)^{-1} X^\top \Sigma X (X^\top X)^{-1}. \tag{2}$$

Here the covariance matrix of the noise is a diagonal matrix $\mathrm{Cov}(\varepsilon) = \Sigma$. To form confidence intervals for individual components of β, we need to estimate diagonal entries of C. White (1980), in one of highest cited papers in econometrics, studied the following plug-in estimator of C, which simply estimates the unknown noise variances by the squared residuals:

$$\widehat{C}_{\mathrm{W}} = (X^\top X)^{-1} X^\top \operatorname{diag}(\widehat{\varepsilon})^2 X (X^\top X)^{-1}. \tag{3}$$

Here $\widehat{\varepsilon} = Y - X\hat{\beta}$ is the vector containing the residuals from the OLS fit. This is also known as the *sandwich estimator*, the *Huber-White*, or the *Eicker-Huber-White* estimator. White showed that this estimator is consistent for the true covariance matrix of $\hat{\beta}$, when the sample size grows to infinity, $n \to \infty$, with fixed dimension p. Earlier closely related work was done by Eicker (1967); Huber (1967). In theory, these works considered more general problems, but White's estimator was explicit and directly applicable to the central problem of inference in OLS. This may explain why White's work has achieved such a large practical impact, with more than 34,000 citations at the time of writing.

However, it was quickly realized that White's estimator is *substantially biased* when the sample size n is not too large—for instance when we only have twice as many samples as the

dimension. This is a problem, because it can lead to incorrect statistical inferences. MacKinnon and White (1985) proposed a bias-correction that is unbiased under homoskedasticity. However, the question of forming confidence intervals has remained challenging. Despite the unbiasedness of the MacKinnon-White estimate in special cases, confidence intervals based on it have below-nominal probability of covering the true parameters in low dimensions (see e.g., Kauermann and Carroll, 2001). It is not clear if this continues to hold in the high-dimensional case. In fact in our simulations we observe that these confidence intervals (CIs) can be anti-conservative in high dimensions. Thus, constructing accurate CIs in high dimensions remains a challenging open problem.

In this paper, we propose to construct confidence intervals via a variance estimator that is unbiased *even under heteroskedasticity*. Since the estimator (described later), is based on Hadamard products, we call it the *Hadamard estimator*. This remarkable estimator has been discovered several times (Hartley et al., 1969; Chew, 1970; Cattaneo et al., 2018), and the later works do not always appear to be aware of the earlier ones. The estimator does not appear to be widely known by researchers in finance and econometrics, and does not appear in standard econometrics textbooks such as Greene (2003), or in recent review papers such as Imbens and Kolesar (2016). We came upon the Hadamard estimator in 2017 while studying the bias of White's estimator, and were surprised to find out about how early it was discovered. We emphasize that the papers above did not study many of the important properties of this estimator. For instance, it is not even clear based on these works under what conditions this estimator exists.

In our paper, we start by showing how to solve five important problems in the linear regression model using the Hadamard estimator: constructing confidence intervals, estimating signal-to-noise ratio (SNR), quadratic forms of regression parameter, noise level, and mean squared error (MSE) in a robust way under heteroskedasticity (Section 2.1). To use the Hadamard estimator, we need to show the fundamental result that it is well-defined (Section 2.2). We prove matching upper and lower bounds on the relation between the dimension and sample size guaranteeing that the Hadamard estimator is generically well-defined. We also prove well conditioning. For this, we study a new random matrix ensemble in which the entries of a random partial orthogonal projection matrix are squared. Specifically, we prove sharp bounds on the smallest and largest eigenvalues of this matrix. This mathematical contribution should be of independent interest.

Next, we develop a new degrees-of-freedom correction for the Hadamard estimator, which gives more accurate confidence intervals than several variants of the sandwich estimator (Section 2.3). We also establish the rate of convergence and approximate normality of the estimator (Section 4). Using the Hadamard estimator, we are able to perform statistical inference on linear combinations of β. We also propose estimators of the quadratic functionals of β and establish asymptotic normality. The asymptotic distribution results rely on joint central limit theorems for random quadratic forms and linear forms. We present a lemma (Section A.10) that generalizes several related results in the literature and provide a short, self-contained proof. This result may potentially also be of independent interest.

We also compare the Hadamard estimator with the leave-one-out estimator in Kline et al. (2020) which has received broad attention in recent years. The Hadamard estimator can be applied for a broad range of signal-to-noise ratio in regression models. In contrast, the KSS estimator can lead to negative variance estimators when the signal strength is strong, and can

result in an inflated type-I error when performing inference on the coordinates of β. In terms of computation, when considering inference for quadratic forms of β, the Hadamard estimator produces a simpler variance estimator than the KSS estimator, as the latter requires more complicated forms due to its leave-one-out-style sample splitting.

We perform numerical experiments to validate our theoretical results (Section 5). Software implementing our method, and reproducing our results, is available from the authors' GitHub page, `http://github.com/dobriban/Hadamard`.

Notation. For a positive integer n, we denote $[n] = \{1, \dots, n\}$. For two integers $a \leq b$, we write $[a : b] = \{a, a+1 \cdots, b\}$. For a vector $v \in \mathbb{R}^n$, let $\|v\| := (\sum_{i=1}^n v_i^2)^{1/2}$ be the Euclidean norm. For any matrix $A \in \mathbb{R}^{m \times n}$, let $\|A\|$ or $\|A\|_{\text{op}}$ stand for the operator norm, defined by $\|A\| := \sup_{v \in \mathbb{R}^n, v \neq 0} \|Av\|_2 / \|v\|_2$. The Frobenius norm is defined by $\|A\|_{\text{Fr}} := (\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2)^{1/2}$; and the infinity norm is $\|A\|_\infty := \max_{1 \leq i \leq m} \sum_{j=1}^n |A_{ij}|$. We say $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ if there exist constants $C_1, C_2 > 0$ such that $|a_n| \leq C_1 |b_n|$ and $|a_n| \geq C_2 |b_n|$ for all n sufficiently large, respectively. For an event A, $I_{\{A\}}$ denotes its indicator function, which equals one if A occurs and zero otherwise.

2 Main Results

2.1 Solving five problems under heteroskedasticity

Under heteroskedasticity, some fundamental estimation and inference tasks in the linear model are more challenging than under homoskedasticity. As we will see, the difficulty often arises from a lack of a good estimator of the variance of the OLS estimator. For the moment, assume that there is an unbiased estimator of the coordinate-wise variances of the OLS estimator. That is, we consider a vector $\widehat{V}$ satisfying $\mathbb{E}\,\widehat{V} = V$ under heteroskedasticity, where $V = \operatorname{diag} C = \operatorname{diag} \operatorname{Cov}(\hat{\beta})$ is defined through equation (2). To define this unbiased estimator, we collect some useful notation as follows, though the estimator itself shall be introduced in detail in Section 2.2. Let $S = (X^\top X)^{-1} X^\top$ be the matrix used in defining the ordinary least-squares estimate, and $Q = I_n - X(X^\top X)^{-1} X^\top$ be the projection into the orthocomplement of the column space of X. Here I_n is the identity matrix. Let us denote by $M \odot M$ the Hadamard—or elementwise—product of a matrix or vector M with itself.

Among others, the following *five* important applications demonstrate the usefulness of the unbiased variance estimator $\widehat{V}$.

Constructing confidence intervals. A first fundamental problem is inference for the regression coefficients. Assuming the noise ε in the linear model (1) follows a heteroskedastic normal distribution $\varepsilon \sim \mathcal{N}(0, \Sigma)$ for a diagonal covariance matrix Σ, the random variable $(\hat{\beta}_j - \beta_j)/\sqrt{V_j}$ follows the standard normal distribution. We replace the unknown variance V_j of the OLS estimator by its approximation $\widehat{V}_j$ and focus on the distribution of the following *approximate* pivotal quantity

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{V}_j}}. \tag{4}$$

The distribution of this random variable is approximated by a t distribution in Section 2.3 and this plays a pivotal role in constructing confidence intervals and conducting hypothesis testing

for the coefficients. More generally, our inference result handles any linear combination—contrast—of β, see the result in Section 4.

Estimating the SNR. Recall that $\|x\| = (\sum_i x_i^2)^{1/2}$ is the Euclidean norm of a vector $x \in \mathbb{R}^n$. The signal-to-noise ratio (SNR)

$$\mathrm{SNR} = \frac{n\|\beta\|^2}{\mathbb{E}\,\|\varepsilon\|^2} = \frac{n\|\beta\|^2}{\operatorname{tr}\Sigma}$$

of the linear model (1) is a fundamental measure that quantifies the fraction of variability explained by the covariates of an observational unit. In genetics, the SNR corresponds to heritability if the response y denotes the phenotype of a genetic trait (Visscher et al., 2008). Existing work on estimating this important ratio in linear models, however, largely focuses on the relatively simple case of homoskedasticity (see, for example, Dicker (2014); Janson et al. (2017)). Without appropriately accounting for heteroskedasticity, the estimated SNR may be unreliable.

As an application of the estimator $\widehat{V}$, we propose to estimate the SNR using

$$\widehat{\mathrm{SNR}} = \frac{\|\hat{\beta}\|^2 - 1_p^\top \widehat{V}}{n^{-1} 1_p^\top (Q \odot Q)^{-1} (\widehat{\varepsilon} \odot \widehat{\varepsilon})}, \tag{5}$$

where recall that $\widehat{\varepsilon}$ is the vector of residuals in the linear model, and 1_p denotes a column vector with all p entries being ones. Above, $(Q \odot Q)^{-1}$ denotes the inverse of the Hadamard product $Q \odot Q$ of $Q = I_n - X(X^\top X)^{-1}X^\top$ with itself (we will later study this invertibility in detail). The numerator and denominator of the fraction in (5) are unbiased for the signal part and noise part, respectively, as we show in the next two examples. As shown in Section A.13, this estimator is ratio-consistent.

Estimating quadratic forms of β. A further fundamental problem involves estimation and inference for quadratic forms of the type $\beta^\top A\beta$ for some non-random symmetric matrix A. Applications include estimating the magnitude of the regression coefficient $\|\beta\|^2$, conducting global testing (Guo and Cheng, 2022) and performing ANOVA analyses (Kline et al., 2020). From the identity $\mathbb{E}\,\hat{\beta}^\top A\hat{\beta} = \beta^\top A\beta + \operatorname{tr}\left(A \operatorname{Cov}(\hat{\beta})\right)$, it follows that an unbiased estimator of $\beta^\top A\beta$ is $\hat{\beta}^\top A\hat{\beta} - \operatorname{tr}\left(A \operatorname{diag}(\widehat{V})\right)$. As a special case, an unbiased estimator of the squared signal magnitude is $\|\hat{\beta}\|^2 - 1_p^\top \widehat{V}$. In Section 4.2, we discuss estimation and inference for $\beta^\top A\beta$ in detail.

Estimating the total noise level. As an intermediate step in the derivation of the unbiased estimator $\widehat{V}$, we obtain the identity

$$\operatorname{diag}(\Sigma) = (Q \odot Q)^{-1}\,\mathbb{E}(\widehat{\varepsilon} \odot \widehat{\varepsilon}). \tag{6}$$

That is, the vector $\operatorname{diag}(\Sigma)$ of the entries of Σ can be written as a matrix-vector product in the appropriate way. As a consequence of this, we can use $(Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon})$ to estimate $\operatorname{diag}(\Sigma)$ in an unbiased way. In addition, we can use $1_p^\top (Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon})$ as an unbiased estimate of the total noise level $\operatorname{tr}(\Sigma) = \sum_{i=1}^n \operatorname{Var}(\varepsilon_i)$.

Estimating the MSE. An important problem concerning the least-squares method is estimating its mean squared error (MSE). Let $\mathrm{MSE} = \mathbb{E}\,\|\hat{\beta} - \beta\|^2$ be the MSE. Consider the estimator $\widehat{\mathrm{MSE}} = \sum_{i=1}^n \widehat{V}_i$. As in the part about estimating quadratic forms of β, it follows that $\widehat{\mathrm{MSE}}$ is an unbiased estimator of the MSE. Later in Section 5 we will show in simulations that this estimator is more accurate than the corresponding estimators based on White's and MacKinnon-White's covariance estimators.

2.2 The Hadamard estimator and its well-posedness

This section specifies the variance estimator $\widehat{V}$. This estimator has appeared in Hartley et al. (1969); Chew (1970); Cattaneo et al. (2018), and takes the following form of matrix-vector product

$$\widehat{V} = A(\widehat{\varepsilon} \odot \widehat{\varepsilon}),$$

where the matrix A is

$$A = (S \odot S)(Q \odot Q)^{-1}. \tag{7}$$

Here $(Q \odot Q)^{-1}$ is the usual matrix inverse of $Q \odot Q$ and recall that both $Q \odot Q$ and $\widehat{\varepsilon} \odot \widehat{\varepsilon}$ denote the Hadamard product. As such, $\widehat{V}$ is henceforth referred to as the *Hadamard estimator*. In short, this is a method of moments estimator, using linear combinations of the squared residuals.

While the Hadamard estimator enjoys a simple expression, there is little work on a fundamental question: whether this estimator *exists or not*. More precisely, in order for the Hadamard estimator to be well-defined, the matrix $Q \odot Q$ must be invertible. Without this knowledge, all five important applications in Section 2.1 would suffer from a lack of theoretical foundation. While the invertibility can be checked for a given dataset, knowing that it should hold under general conditions gives us a confidence that the method can work broadly.

As a major thrust of this paper, we provide a deep understanding of under what conditions $Q \odot Q$ should be expected to be invertible. The problem is theoretically nontrivial, because there are no general statements about the invertibility of matrices whose entries are squared values of some other matrix. In fact, $Q = I_n - X(X^\top X)^{-1}X^\top$ is an $n \times n$ *rank-deficient* projection matrix of rank $n-p < n$. Therefore, Q itself is not invertible, and it is not clear how its rank behaves when the entries are squared. However, we have the following lower bound on n for this invertibility to hold.

Proposition 2.1 (Lower bound)**.** *If the Hadamard product $Q \odot Q$ is invertible, then the sample size n must be at least*

$$n \geq p + \frac{1}{2} + \sqrt{2p + \frac{1}{4}}. \tag{8}$$

This result reveals that the Hadamard estimator simply *does not exist* if n is only slightly greater than p, (say $p = n + 1$), though the OLS estimator exists in this regime. The proof of Proposition 2.1 comes from a well-known property of the Hadamard product, that is, if a matrix B is of rank r, then the rank of $B \odot B$ is at most $r(r+1)/2$ (e.g., Horn and Johnson, 1994). For completeness, a proof of this property is given in Section A.2. Using this property, the invertibility of $Q \odot Q$ readily implies

$$n \leq \frac{(n-p)(n-p+1)}{2},$$

which is equivalent to (8).

In light of the above, it is tempting to ask whether (8) is sufficient for the existence of the Hadamard estimator. In general, this is not the case. For example, let $X = \begin{pmatrix} R \\ \mathbf{0} \end{pmatrix}$ for any orthogonal matrix $R \in \mathbb{R}^{p \times p}$. Then, $Q \odot Q$ is not invertible as Q is a diagonal matrix whose first p diagonal entries are 0 and the remaining are 1. This holds no matter how large n is compared to p. However, such design matrices X that lead to a degenerate $Q \odot Q$ are very "rare" in the sense of the following theorem. Recall that $Q = I_n - X(X^\top X)^{-1}X^\top$.

Theorem 1. *The set*

$$\left\{X \in \mathbb{R}^{n \times p} : Q \odot Q \text{ does not have full rank}\right\}$$

has Lebesgue measure zero in $\mathbb{R}^{np}$ *if the inequality* (8) *is satisfied.*

Therefore, the lower bound in Proposition 2.1 is sharp. Roughly speaking, $n \geq p + c\sqrt{p}$, for $c > 0$, is sufficient for the invertibility of $Q \odot Q$. The proof of this result is new in the vast literature on the Hadamard matrix product. In short, our proof uses certain algebraic properties of the determinant of $Q \odot Q$ and employs a novel induction step. Section 3 is devoted to developing the proof of Theorem 1 in detail. To be complete, Cattaneo et al. (2018) show high-probability invertibility when $p > 2n$ for Gaussian designs. Our invertibility result is more broadly applicable.

Up to now, we have conditioned on X, working in a fixed design setting. To better appreciate the theoretical contributions of our paper, we consider a random matrix X in the following corollary, which ensures that the Hadamard estimator is well-defined almost surely for popular random matrix ensembles of X such as the Wishart ensemble.

Corollary 2.2. *Under the same conditions as in Theorem 1, if* X *is sampled from a distribution that is absolutely continuous with respect to the Lebesgue measure on* $\mathbb{R}^{n \times p}$ *(put simply,* X *has a density), then* $Q \odot Q$ *is invertible almost surely.*

Although $Q \odot Q$ is invertible under very general conditions, our simulations reveal that the condition number of this matrix can be very large for p close to n due to very small eigenvalues. This is problematic, because the estimator can then amplify the error. Our next result shows that $Q \odot Q$ is well-conditioned under some conditions if $n > 2p$. We will show that this holds for certain random design matrices X.

Suppose for instance that the entries of X are i.i.d. standard normal, $X_{ij} \sim \mathcal{N}(0, 1)$. Then, each diagonal entry of $Q = I_n - X(X^\top X)^{-1}X^\top$ is relatively large, of unit order. The off-diagonal entries are of order $1/n^{1/2}$. When we square the entries, the off-diagonal entries become of order $1/n$, while the diagonal ones are still of unit order. Thus, it is possible that the matrix is *diagonally dominant*, so the diagonal entries are larger than the sum of the off-diagonal ones. This would ensure well-conditioning. We will show rigorously that this is true under some additional conditions.

Specifically, we will consider a *high-dimensional asymptotic* setting, where the dimension p and the sample size n are both large. We assume that they grow proportionally to each other, $n, p \to \infty$ with $p/n \to \gamma > 0$. This is a modern setting for high-dimensional statistics, and it has many connections to random matrix theory (see e.g., Bai and Silverstein, 2010; Paul and Aue, 2014; Yao et al., 2015).

We will provide bounds on the largest and smallest eigenvalues. We can handle correlated designs X, where each row is sampled i.i.d. from a distribution with $p \times p$ covariance matrix Γ. Let $\Gamma^{1/2}$ be the symmetric square root of Γ.

Theorem 2 (Eigenvalue bounds for the Hadamard product with a random design)**.** *Suppose the rows x_i of X are i.i.d. and have the form $x_i = \Gamma^{1/2} z_i$, where z_i have independent entries with mean zero, unit variance and uniformly bounded $(8+\delta)$-th moment. Suppose that Γ is invertible. Then, as $n, p \to \infty$ such that for $\gamma_{p,n} := p/n$, we have $\limsup \gamma_{p,n} < 1/2$, the matrix $T = Q \odot Q$ with $Q = I_n - X(X^\top X)^{-1} X^\top$ satisfies the following eigenvalue bounds for any fixed $\xi > 0$ and sufficiently large n:*

$$\lambda_{\max}(T) < 1 - \gamma_{p,n} + \xi,$$

and

$$\lambda_{\min}(T) > (1 - \gamma_{p,n})(1 - 2\gamma_{p,n}) - \xi,$$

with probability at least $1 - Cn^{-1-\delta/4}\xi^{-4-\delta/2}$ for some positive constant C not depending on ξ.

See Section A.3 for a proof. Hence, if $p/n \to \gamma < 1/2$, then almost surely

$$(1-\gamma)(1-2\gamma) \leq \liminf \lambda_{\min}(T) \leq \limsup \lambda_{\max}(T) \leq (1-\gamma).$$

Practically speaking, the above result states that the condition number of T is at most $1/(1-2\gamma)$ with high probability. Our invertibility results are stronger than those of Cattaneo et al. (2018). Specifically, we show generic invertibility in finite dimensional designs with probability one, and condition number bounds on non-Gaussian correlated designs that go beyond those considered in their work.

While Corollary 2.2 proves invertibility for continuous distributions, in practice some columns of X can be discrete. In that case, we can still obtain invertibility or condition number bounds by applying Lemma A.2 used in the proof of Theorem 2. That result which has no assumptions on the continuity of X and provides non-asymptotic bounds for the eigenvalues of the matrix.

As an illustration, we study a one-way ANOVA model $y_i = \alpha_g + \varepsilon_i$ where y_i is the observation i, α_g is the group effect for group g, and ε_i is the error term. Suppose there are p groups with sizes $n_1, \ldots, n_p$ such that $\sum_{i=1}^p n_i = n$. The design matrix X is constructed as follows: $X_{j,1} = 1$ for $j \in [1, n_1]$, $X_{j,2} = 1$ for $j \in [n_1+1, n_1+n_2]$, and so on, until $X_{j,p} = 1$ for $j \in [\sum_{i=1}^{p-1} n_i + 1, \sum_{i=1}^p n_i]$. All other entries of X are zero. This design matrix encodes the group membership for all observations. It is readily verified that $x_i^\top R_i^{-1} x_i = 1/n_g$ if observation i is in group g. If $3 \leq \min_{j\in[p]} n_j \leq \max_{j\in[p]} n_j \leq C$, then $\lambda_{\min}(T) \geq 2/9$ and $\lambda_{\max}(T) \leq C/(C+1)$.

2.3 Degrees-of-freedom adjustment

To obtain a confidence interval for β_j, we propose to approximate the distribution of the approximate pivot in (4) by a t-distribution. The key is to find a good approximation to the degrees of freedom. Let us denote by $V_j = \operatorname{Var} \widehat{\beta}_j$, the expected value of $\widehat{V}_j$. Suppose the

degrees of freedom of $\widehat{V}_j$ are d_j. Using the second moment properties of the $\chi^2_{d_j}$ variable, these degrees of freedom should obey that

$$\mathbb{E}\,\widehat{V}_j^2 \approx \frac{V_j^2}{d_j^2}\,\mathbb{E}\,\chi^4_{d_j} = V_j^2(1+2/d_j).$$

Consequently, we formally define

$$d_j = \frac{2}{\frac{\mathbb{E}\,\widehat{V}_j^2}{V_j^2} - 1} = \frac{2V_j^2}{\mathbb{E}\,\widehat{V}_j^2 - V_j^2}. \tag{9}$$

To proceed, we need to evaluate $\mathbb{E}[\widehat{V} \odot \widehat{V}] \in \mathbb{R}^p$. The following proposition gives a closed-form expression of this vector assuming *homoskedasticity*. Let us denote

$$E = \operatorname{diag}\left[(X^\top X)^{-1}\right] \odot \operatorname{diag}\left[(X^\top X)^{-1}\right]. \tag{10}$$

Recall that $S = (X^\top X)^{-1}X^\top$.

Proposition 2.3 (Degrees of freedom)**.** *If the noise ε has i.i.d. normal entries, we have that the vector of degrees of freedom of $\widehat{V}$, defined in equation* (9)*, has the form*

$$d = \frac{2E}{\operatorname{diag}\left[(S\odot S)1_n 1_n^\top (S\odot S)^\top\right] + 2\operatorname{diag}\left[(S\odot S)(Q\odot Q)^{-1}(S\odot S)^\top\right] - E}, \tag{11}$$

where the division is understood to be entrywise.

See Section A.6 for a proof.

We call the inference method based on approximating $(\hat{\beta}_i - \beta_i)/\hat{V}_i^{1/2}$ by a t-distribution with the degrees of freedom specified by (11) the *Hadamard-t* method. This result also leads to a useful *degrees of freedom* heuristic. If the degrees of freedom d_i are large, this suggests that inferences for β_i are based on a large amount of information. On the other hand, if the degrees of freedom are small, this suggests that the inferences are based on little information, and may thus be unstable.

In our case, the t-distribution is still a heuristic, because the numerator and denominator are not independent under heteroskedasticity. However, the degree of dependence can be bounded as follows:

$$\begin{aligned}\|\operatorname{Cov}(\hat{\beta},\widehat{\varepsilon})\|_{\mathrm{op}} &= \|S\Sigma(S^\top X^\top - I)\|_{\mathrm{op}} = \|S(\Sigma - cI)(S^\top X^\top - I)\|_{\mathrm{op}} \\ &\le \|S\|_{\mathrm{op}}\|\Sigma - cI\|_{\mathrm{op}}\|S^\top X^\top - I\|_{\mathrm{op}} \le \tfrac{|\Sigma_{\max}-\Sigma_{\min}|}{2\sigma_{\min}(X)}.\end{aligned} \tag{12}$$

In the first line, we have used that $S = (X^\top X)^{-1}X^\top$. Hence, $S(S^\top X^\top - I) = 0$. Indeed,

$$SS^\top X^\top = (X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = (X^\top X)^{-1}X^\top = S.$$

For this reason, we can add a constant times $S(S^\top X^\top - I) = 0$ in the second step. Then, we can use the inequality $\|AB\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}}\|B\|_{\mathrm{op}}$ for any two conformable matrices A, B.

In (12), we have chosen $c = (\Sigma_{\max} + \Sigma_{\min})/2$, where $\Sigma_{\max}$ and $\Sigma_{\min}$ denote the maximal and minimal entries of Σ, respectively. Moreover, we have also used that $\|S\|_{\mathrm{op}} = 1/\sigma_{\min}(X)$, while $\|S^\top X^\top - I\|_{\mathrm{op}} = \|X(X^\top X)^{-1}X^\top - I\|_{\mathrm{op}} \leq 1$.

Now, for designs X of aspect ratios $n \times p$ that are not close to 1, and with i.i.d. entries with sufficiently many moments, it is known that $\sigma_{\min}(X)$ is of the order $n^{1/2}$. This suggests that the covariance between $\hat{\beta}$ and $\widehat{\varepsilon}$ is small. Hence, this heuristic suggests that the t-approximation should be accurate. Moreover when $\widehat{V}_j - V_j \to 0$ in probability, and under the conditions in Section 4.1, we also have that the limiting distribution is standard normal.

2.4 Hadamard estimator with $p = 1$

As a simple example, consider the case of one covariate, when $p = 1$. In this case, we have $Y = X\beta + \varepsilon$, where y, X, ε are n-vectors. Assuming without loss of generality that $X^\top X = 1$, the OLS estimator takes the form $\hat{\beta} = X^\top y$. Its variance equals $V = \sum_{j=1}^n X_j^2 \Sigma_j$, where Σ_j is the variance of ε_j, and X_j are the entries of X.

The Hadamard estimator takes the form

$$\widehat{V} = \frac{\sum_{j=1}^n \frac{X_j^2}{1-2X_j^2}\widehat{\varepsilon}_j^2}{1 + \sum_{j=1}^n \frac{X_j^4}{1-2X_j^2}},$$

which is well-defined if all coordinates X_j^2 are small enough that $1 - 2X_j^2 > 0$. See section A.7 for the argument. The unbiased estimator is not always nonnegative. To ensure nonnegativity, we need $X_j^2 < 1/2$ in this case. In practice, we may enforce non-negativity by using $\max(\widehat{V}, 0)$ instead of $\widehat{V}$, but see below for a more thorough discussion.

For comparison, White's variance estimator is $\widehat{V}_{\mathrm{W}} = \sum_{j=1}^n X_j^2 \widehat{\varepsilon}_j^2$, while MacKinnon-White's variance estimator (MacKinnon and White, 1985) can be seen to take the form

$$\widehat{V}_{\mathrm{MW}} = \sum_{j=1}^n \frac{X_j^2}{1 - X_j^2}\widehat{\varepsilon}_j^2 = \sum_{j=1}^n \frac{X_j^2}{\sum_{i=1, i\neq j}^n X_i^2}\widehat{\varepsilon}_j^2.$$

We observe that each variance estimator is a weighted linear combination of the squared residuals, where the weights are some functions of the squares of the entries of the feature vector X. For White's estimator, the weights are simply the squared entries. For MacKinnon-White's variance estimator, the weights are scaled up by a factor $1/(1-X_j^2) > 1$. As we know, this ensures the estimator is unbiased under homoskedasticity. For the Hadamard estimator, the weights are scaled up more aggressively by $1/(1 - 2X_j^2) > 1$, and there is an additional normalization step. In general, these weights do not have to be larger—or smaller—than those of the other two weighting schemes.

A critical issue is that *the Hadamard estimator may not always be non-negative*. It is well known that unbiased estimators may fall outside of the parameter space (Lehmann and Casella, 1998). When $p = 1$, almost sure non-negativity is ensured when the coordinates of X are sufficiently small. It would be desirable, but seems non-obvious, to obtain such results for general dimension p.

In addition, the degrees of freedom from (11) simplifies to $d = 1 + 1/\left(\sum_{j=1}^n \frac{X_j^4}{1-2X_j^2}\right)$. This can be as large as $n-1$, for instance $d = n-1$ when all $X_i^2 = 1/n$. The degrees of freedom can only be small if the distribution of X_i^2 is very skewed.

2.5 Bias of classical estimators

As a byproduct of our analysis, we also obtain explicit formulas for the bias of the two classical estimators of the variances of the ordinary least-squares estimator, namely the White and MacKinnon-White estimators. This can in principle enable us to understand when the bias is small or large.

The estimator proposed by MacKinnon and White (1985), which we will call the *MW estimator*, is:

$$\widehat{C}_{\mathrm{MW}} = (X^\top X)^{-1}[X^\top \widehat{\Sigma}_{\mathrm{MW}} X](X^\top X)^{-1}, \tag{13}$$

where $\widehat{\Sigma}_{\mathrm{MW}} = \mathrm{diag}(Q)^{-1}\,\mathrm{diag}(\widehat{\varepsilon})^2$. This estimator is unbiased under homoskedasticity, that is, $\Sigma = \sigma^2 I_n$. It is denoted as HC2 in the paper MacKinnon and White (1985). The same estimator was also proposed by Wu (1986), equation (2.6).

Proposition 2.4 (Bias of classical estimators)**.** *Consider White's covariance estimator defined in* (3) *and MacKinnon-White's estimator defined in* (13)*. Their bias for estimating the coordinate-wise variances of the OLS estimator equals, respectively*

$$b_{\mathrm{W}} = (S \odot S)[(Q \odot Q) - I_n]\vec{\Sigma} \tag{14}$$

for White's covariance estimator, and

$$b_{\mathrm{MW}} = (S \odot S)[\mathrm{diag}(Q)^{-1}(Q \odot Q) - I_n]\vec{\Sigma} \tag{15}$$

for MacKinnon-White's estimator. Here $\vec{\Sigma}$ is the vector of diagonal entries of Σ, the covariance of the noise.

See Section A.8 for a proof. In particular, MacKinnon-White's estimator is known to be unbiased under homoskedasticity, that is when $\Sigma = I_n$ (MacKinnon and White, 1985). This can be checked easily using our explicit formula for the bias. Specifically suppose that $\Sigma = I_n$. Then, $\vec{\Sigma} = 1_n$, the vector of all ones. Therefore, $(Q \odot Q)\vec{\Sigma} = \mathrm{vec}(\|q_j\|^2)$, the vector of squared Euclidean norms of the rows of Q. Since Q is a projection matrix, $Q^2 = Q$, so $\|q_j\|^2 = Q_{jj}$. Therefore we see that

$$[\mathrm{diag}(Q)^{-1}(Q \odot Q) - I_n]\vec{\Sigma} = \mathrm{diag}(Q)^{-1}\mathrm{vec}(Q_{jj}) - 1_n = 0,$$

so that MacKinnon-White's estimator is unbiased under homoskedasticity.

In Section B.1, we conduct a more refined analysis of the bias in the MW estimator when estimating the covariance of $\hat{\beta}$ along a given direction w_p.

2.6 Some related work

There has been a lot of related work on inference in linear models under heteroskedasticity. Here we can only mention a few of the most closely related works, and refer to Imbens and Kolesar (2016) for a review. In the low-dimensional case, Bera et al. (2002) compared the Hadamard and White-type estimators and discovered that the Hadamard estimator lead to more accurate coverage, while the White estimators have better mean squared error.

As a heuristic to improve the performance of the MacKinnon-White (MW) confidence intervals in high dimensions, Bell and McCaffrey (2002) have a similar approach to ours, with a t degrees of freedom correction. Simulations in the very recent review paper by Imbens and Kolesar (2016) suggest this method is the state of the art for heteroskedasticity-consistent inference, and performs well under many settings. However, this correction is computationally more burdensome than the MW method, because it requires a separate $O(p^3)$ computation for each regression coefficient, raising the cost to $O(p^4)$. In contrast, our method has computational cost $O(p^3)$ only. In addition, the accuracy of their method typically does not increase substantially compared to the MW method. We think that this could be due to the bias of the MW method under heteroskedasticity. Kline et al. (2020) use the leave-one-out method to construct unbiased estimators for the variance components. As their priority is studying the analysis of variance, which essentially aim at inference on quadratic forms of β, we compare the Hadamard estimator with their method in detail in Section 4.2.

In this work, we have used the term "robust" informally to mean insensitivity to assumptions about the covariance of the noise. Robust statistics is a much larger field which classically studies robustness to outliers in the data distribution (e.g., Huber and Ronchetti, 2011). Recent work has focused, among many other topics, on high-dimensional regression and covariance estimation (e.g., El Karoui et al., 2013; Chen et al., 2016; Donoho and Montanari, 2016; Zhou et al., 2018; Diakonikolas et al., 2017, etc).

3 Existence of Hadamard Estimator

In this section we develop the novel proof of the existence of the Hadamard estimator. We begin by observing that Theorem 1 is equivalent to the proposition below. This is because the Lebesgue measure admits an orthogonal decomposition using the SVD.

Proposition 3.1. *Assume $r(r+1)/2 \geq n$. Denote by $\mathcal{Q}$ the set of all $n \times n$ projection matrices of rank r and let $\mathrm{d}\,\mathcal{Q}$ be the Lebesgue measure on $\mathcal{Q}$. Then, the set $\{Q \in \mathcal{Q} : \mathrm{rank}(Q \odot Q) < n\}$ has zero-$\mathrm{d}\,\mathcal{Q}$ measure.*

We take the following lemma as given for the moment.

Lemma 3.2. *Under the same assumptions as Proposition 3.1, there exists a $Q^* \in \mathcal{Q}$ such that $\mathrm{rank}(Q^* \odot Q^*) = n$.*

A proof of Proposition 3.1 using Lemma 3.2 is readily given as follows.

Proof of Proposition 3.1. Let $p = n - r$. Consider the map from $\mathbb{R}^{n \times p}$ (ignoring the zero-Lebesgue measure set where X is not of rank p) to $\mathcal{Q}$:

$$X \in \mathbb{R}^{n \times p} \longrightarrow Q = I - X(X^\top X)^{-1} X^\top \in \mathcal{Q}.$$

It is easy to see that the map is a surjection and the preimage of this map for every $Q \in \mathcal{Q}$ is rotationally equivalent to each other. Hence, it suffices to show that the set of X where the Hadamard product of $I - X(X^\top X)^{-1}X^\top$ is degenerate is measure zero.

We observe that the determinant takes the form

$$\det\Big((I - X(X^\top X)^{-1}X^\top) \odot (I - X(X^\top X)^{-1}X^\top)\Big) = \frac{f_1(X)}{f_2(X)},$$

where $f_1(X)$ and $f_2(X)$ are polynomials in the np variables $X_{ij}, 1 \le i \le n, 1 \le j \le p$. As a fundamental property of polynomials, one and exactly one of the following two cases holds:

(a) The polynomial $f_1(X) \equiv 0$ for all X.

(b) The roots of $f_1(X)$ are of zero Lebesgue measure.

Lemma 3.2 falsifies case (a). Therefore, case (b) must hold. Recognizing that the set of X where the Hadamard product of $Q(X)$ is not full rank is a subset of the roots of $f_1(X)$, case (b) confirms the claim of the present lemma.

□

Now we turn to prove Lemma 3.2. For convenience, we adopt the following definition.

Definition 3.3. For a set of vectors $u_1, \ldots, u_r \in \mathbb{R}^n$, write $\mathrm{rank}^{\odot}(u_1, \ldots, u_r)$ the rank of the $r(r+1)/2$ vectors each taking the form $u_i \odot u_j$ for $1 \le i \le j \le r$.

First, we give two simple lemmas.

Lemma 3.4. *Suppose two sets of vectors $\{u_1, u_2, \ldots, u_r\}$ and $\{u'_1, u'_2, \ldots, u'_{r'}\}$ are linearly equivalent, meaning that one can be linearly represented by the other. Then,*

$$\mathrm{rank}^{\odot}(u_1, \ldots, u_r) = \mathrm{rank}^{\odot}(u'_1, \ldots, u'_{r'}).$$

Lemma 3.5. *For any matrix P that takes the form $P = u_1u_1^\top + \ldots + u_ru_r^\top$ for some vectors $u_1, \ldots, u_r$, we have*

$$\mathrm{rank}(P \odot P) = \mathrm{rank}^{\odot}(u_1, \ldots, u_r).$$

Making use of the two lemmas above, Lemma 3.2 is validated once we show the following.

Lemma 3.6. *There exist vectors $u_1, \ldots, u_r$ such that $\mathrm{rank}^{\odot}(u_1, \ldots, u_r) = n$ if $r(r+1)/2 \ge n$.*

To see this point, we apply the Gram–Schmidt orthonormalization to $u_1, \ldots, u_r$ considered in Lemma 3.6, and get orthonormal vectors $v_1, \ldots, v_r$. Write $Q^* = v_1v_1^\top + \ldots + v_rv_r^\top$, which belongs to $\mathcal{Q}$. Since $u_1, \ldots, u_r$ and $v_1, \ldots, v_r$ are linearly equivalent, Lemmas 3.4 and 3.5 reveal that

$$\mathrm{rank}(Q^* \odot Q^*) = \mathrm{rank}^{\odot}(v_1, \ldots, v_r) = \mathrm{rank}^{\odot}(u_1, \ldots, u_r) = n.$$

Now we aim to prove Lemma 3.6.

Proof of Lemma 3.6. We consider a stronger form of Lemma 3.6: for *generic* $u_1, \ldots, u_r$, any combination of n vectors from $u_i \odot u_j$ for $1 \le i \le j \le r$ have full rank. Here *generic* means that this statement does not hold only for a set of zero Lebesgue measure.

We induct on n. The statement is true for $n = 1$. Suppose it has been proven true for $n-1$. Let $\mathcal{U}$ denote an arbitrary subset of $\{(i,j) : 1 \le i \le j \le r\}$ with cardinality n. Write $P = (u_i \odot u_j)_{(i,j)\in\mathcal{U}}$.

It is sufficient to show that $\det(P)$ is generically nonzero. As earlier in the proof of Proposition 3.1, it suffices to show that $\det(P)$ is not *always* zero. Without loss of generality, let $(i_0, j_0) \in \mathcal{U}$ be the first column of P. Expressing the determinant of P in terms of its minors along the first column, we see that $\det(P)$ is an affine function of $u_{i_0}(1)u_{j_0}(1)$, with the leading coefficient being the determinant of a $(n-1)\times(n-1)$ minor matrix that results from P by removing the first row and the first column. The induction step is complete if we show that this minor matrix, denoted by $P_{1,1}$ is nonzero generically. Write $u_i^{(-1)}$ the vector in $\mathbb{R}^{n-1}$ formed by removing the first entry from u_i for $i = 1, \ldots, r$. Then, each of the $n-1$ column of $P_{1,1}$ takes the form $u_i^{(-1)} \odot u_j^{(-1)}$ for some $(i,j) \in \mathcal{U} \setminus \{(i_0, j_0)\}$. Since the induction step has been validated for $n-1$, it follows that the determinant of $P_{1,1}$ is nonzero in the generic sense.

□

To complete this section, we prove below Lemmas 3.4 and 3.5.

Proof of Lemma 3.4. Since $\{u'_1, u'_2, \ldots, u'_{r'}\}$ can be linearly represented by $\{u_1, u_2, \ldots, u_r\}$, each u'_j can be written as $u'_j = \sum_{l=1}^{r} a_l^j u_l$ for constants a_l^j. Using the representation, the Hadamard product between two vectors reads

$$u'_i \odot u'_j = \left(\sum_{l=1}^{r} a_l^i u_l\right) \odot \left(\sum_{l=1}^{r} a_l^j u_l\right) = \sum_{l_1,l_2} a_{l_1}^i a_{l_2}^j u_{l_1} \odot u_{l_2}.$$

This expression for $u'_i \odot u'_j$ suggests that $u'_i \odot u'_j$ is in the linear span of $u_{l_1} \odot u_{l_2}$ for $1 \le l_1 \le l_2 \le r$. As a consequence of this, it must hold that

$$\begin{aligned}\operatorname{rank}^{\odot}(u'_1, u'_2, \ldots, u'_{r'}) &\equiv \operatorname{rank}(\{u'_i \odot u'_j : 1 \le i \le j \le r'\}) \\ &\le \operatorname{rank}(\{u_{l_1} \odot u_{l_2} : 1 \le l_1 \le l_2 \le r\}) = \operatorname{rank}^{\odot}(u_1, u_2, \ldots, u_r).\end{aligned}$$

Likewise, we have $\operatorname{rank}^{\odot}(u'_1, u'_2, \ldots, u'_{r'}) \ge \operatorname{rank}^{\odot}(u_1, u_2, \ldots, u_r)$. Taking the two inequalities together leads to an identity between the two ranks.

□

Proof of Lemma 3.5. As earlier in this section, we can write P as

$$P \odot P = \sum_{1\le i,j\le r} (u_i \odot u_j)(u_i \odot u_j)^\top.$$

Let R be an $n \times r^2$ matrix formed by the r^2 columns $u_i \odot u_j$ for $1 \le i, j \le n$. Clearly, $\operatorname{rank}(P \odot P) = \operatorname{rank}(R)$ since $P \odot P = RR^\top$. The (column) rank of R is $\operatorname{rank}^{\odot}(u_1, \ldots, u_r)$ by Definition 3.3, as $u_i \odot u_j = u_j \odot u_i$). Hence, $\operatorname{rank}(P \odot P) = \operatorname{rank}^{\odot}(u_1, \ldots, u_r)$.

□

4 Rate of Convergence

We now turn to studying the rates of convergence of estimators analyzed in this paper. For this, we need to introduce some additional notation. We use $O(\cdot)$ and $o(\cdot)$ for the standard big-O and little-o notation. For two positive sequences $(a_n)_{n\geq 1}$, $(b_n)_{n\geq 1}$, we say $a_n = \Omega(b_n)$ if there exists a universal positive constant c such that $a_n/b_n > c$. The condition number of a square matrix B is denoted by $\kappa(B)$. Convergence in probability and convergence in distribution are denoted as $\to_P$ and $\Rightarrow$, respectively.

We next give a fundamental result characterizing the sampling properties of the Hadamard estimator. This result bounds the relative error for estimating the vector of variances of all the entries of the OLS estimator. It shows that the estimation error is smaller when the aspect ratio γ is small. We write $\vec{\Sigma}$ for the vector of the diagonal elements of Σ.

Theorem 3 (Rate of convergence)**.** *Under the conditions of Theorem 2, assume in addition that the fourth moment of the entries ε_i, $i \in [n]$ is less than a constant $C \geq 3$ times the squared variance of the entries. Let V denote the vector of variances of the entries of the OLS estimator. Then, under high-dimensional asymptotics as $n, p \to \infty$ such that $\limsup \gamma_{p,n} = \limsup p/n < 1/2$, we have for any constant $c > 1$ and some constant $C' > 1$ that for all n large enough,*

$$\mathbb{P}\left(\frac{\|\widehat{V} - V\|}{\|\vec{\Sigma}\|} \geq \frac{t}{n}\right) \leq \frac{2c}{t^2}\frac{1}{\left[\sigma_{\min}(\Gamma)(1-\gamma_{p,n}^{1/2})^2(1-2\gamma_{p,n})\right]^2} + C' n^{-1-\delta/4}.$$

See Section A.9 for a proof.

4.1 Asymptotic normality

We already know that the estimator $\widehat{V}$ is unbiased for the variances of the coordinates of the OLS estimator $V = \operatorname{diag}\operatorname{Cov}(\hat{\beta})$, and in the previous section we have seen an inequality bounding the error $\|\widehat{V} - V\|$. In this section, we aim to study an estimator of $w_p^\top S\Sigma S^\top w_p$ for a sequence $(w_p)_{p\geq 1}$ of vectors $w_p \in \mathbb{R}^p$. This represents the variance of $w_p^\top \hat{\beta}$, and taking w_p to be the i-th canonical basis vectors in $\mathbb{R}^p$, for all p, it reduces to V_i. The analysis will later be further used to derive an inferential method for $w_p^\top \beta$.

We use the coordinate-wise case of estimating V_i for some $i \in [p]$, to illustrate the idea. To study the asymptotic distribution of $\widehat{V}_i = A_i^\top(\widehat{\varepsilon} \odot \widehat{\varepsilon})$, where $A_i^\top$ is the i-th row of $A = (S \odot S)(Q \odot Q)^{-1}$, we consider the noise $\varepsilon = \Sigma^{1/2}Z$ to be linear combination of a vector whose entries have zero mean, unit variance and bounded fourth moments. We can express the residuals as $\widehat{\varepsilon} = Q\Sigma^{1/2}Z$.

Thus, we see that the estimator $\widehat{V}_i$, a linear combination of squared entries of $\widehat{\varepsilon}_i$, can be written as a symmetric quadratic form in Z. In particular, if $Z \sim N(0, I_n)$, its distribution is a weighted linear combination of chi-squared random variables. For general w_p, the above discussion still applies by replacing $A_i^\top$ with $[(w_p^\top S) \odot (w_p^\top S)](Q \odot Q)^{-1}$. For a linear combination of chi-squared random variables, we expect that it is close to a normal distribution if none of the weights is too large. We obtain the approximation to the normality of the variance estimator, given in the following result, proved in Section A.11.

Theorem 4 (Approximate normality)**.** *Assume $\varepsilon = \Sigma^{1/2} Z$ where $Z = (Z_1, \ldots, Z_n)^\top$ consists of independent entries that have means zero, variances one, and fourth moments bounded by a constant. Let $\widehat{\vec{\Sigma}} = (Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon})$, and*

$$G(w_p) = \Sigma^{1/2} Q \operatorname{diag}\left\{[(w_p^\top S) \odot (w_p^\top S)](Q \odot Q)^{-1}\right\} Q \Sigma^{1/2}. \tag{16}$$

Assume that

$$\frac{\|G(w_p)\|}{\|G(w_p)\|_{\text{Fr}}} \to 0. \tag{17}$$

For a sequence of vectors $(w_p)_{p \geq 1}$, where $w_p \in \mathbb{R}^p$ for all $p \geq 1$, of unit norm, we have that

$$\frac{w_p^\top S(\operatorname{diag} \widehat{\vec{\Sigma}}) S^\top w_p - w_p^\top S \Sigma S^\top w_p}{\sqrt{\operatorname{Var}[w_p^\top S(\operatorname{diag} \widehat{\vec{\Sigma}}) S^\top w_p]}} \Rightarrow \mathcal{N}(0, 1). \tag{18}$$

We do not necessarily require Z_i, $i \in [p]$, to have unit variances, since we can normalize them and absorb the constants into Σ. In addition, as a special case of this theorem, taking w_p to be the i-th canonical basis vectors in $\mathbb{R}^p$, for all p, leads to

$$\frac{\hat{V}_i - V_i}{\sqrt{\operatorname{Var} \hat{V}_i}} \Rightarrow \mathcal{N}(0, 1)$$

if $\|G(e_i)\|/\|G(e_i)\|_F \to 0$, where $G(e_i) = \Sigma^{1/2} Q \operatorname{diag}[e_i^\top (S \odot S)(Q \odot Q)^{-1}] Q \Sigma^{1/2}$.

In principle, this result could justify using normal confidence intervals for inference on V_i as soon (17) holds. Moreover, the upper bound in Theorem 4 can be simplified as follows. Denote $A(w_p) = [(w_p^\top S) \odot (w_p^\top S)](Q \odot Q)^{-1}$. We have the upper bound $\|G(w_p)\| \leq \|\Sigma\| \|Q \operatorname{diag}[A(w_p)] Q\|$ and the lower bound $\|G(w_p)\|_{\text{Fr}} \geq \lambda_{\min}(\Sigma) \|Q \operatorname{diag}[A(w_p)] Q\|_{\text{Fr}}$. Therefore, (17) simplifies to

$$C_0 \kappa(\Sigma) \frac{\|Q \operatorname{diag}[A(w_p)] Q\|}{\|Q \operatorname{diag}[A(w_p)] Q\|_{\text{Fr}}} \to 0.$$

This bound decouples as the product of a term depending on the unknown covariance matrix Σ, and the known design matrix X. Therefore, in practice one can evaluate the second term. Thus, the deviation from normality only depends on the unknown Σ through its condition number.

The subsequent proposition further characterizes conditions on the design matrix X that lead to $\|G(w_p)\|/\|G(w_p)\|_{\text{Fr}} \to 0$, thereby resulting in the asymptotic normality of the variance estimator of $w_p^\top S \Sigma S^\top w_p$. We also verify that the random design matrix specified in Theorem 2 satisfies these conditions with probability tending to one. See Section A.12 for the proof.

Proposition 4.1 (Conditions for asymptotic normality)**.** *For $j \in [n]$, let $S_{\cdot j}$ be the j-th column of $S = (X^\top X)^{-1} X^\top$. Consider the following conditions, where $T = Q \odot Q$, $v^\top = \left[(w_p^\top S) \odot (w_p^\top S)\right] T^{-1}$, and c is some positive constant:*

1. $\lambda_{\min}(T) > c$.
2. $\kappa(\Sigma) \max_{i \in [n]} |v_i|/\|v\| \to 0$.

Then, for $G(w_p)$ from (16)*, we have $\|G(w_p)\|/\|G(w_p)\|_{Fr} \to 0$. Moreover, for a random design matrix satisfying the conditions in Theorem 2 that is independent of the noise vector ε, if $\kappa(\Gamma)$ is bounded and $\kappa(\Sigma) = o(n^{1/4})$, the above conditions hold with probability tending to one; and thus $w_p^\top S(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top w_p$ is asymptotically normal.*

Our results allow $\kappa(\Sigma)$ to grow to infinity, extending the conditions on the noise assumed in the literature, such as in Kline et al. (2020). The fact that the estimator can handle such unbounded heteroskedasticity has been conjectured in Cattaneo et al. (2018).

After this detailed analysis of $w_p^\top S\Sigma S^\top w_p$ and its estimator $w_p^\top S(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top w_p$, we discuss inference for $w_p^\top \beta$. This task crucially relies on the first conclusion in the following ratio-consistency lemma, which is a simple consequence of several bounds obtained in Proposition 4.1. The second conclusion in this lemma shows that although $\widehat{V}_i$—the unbiased estimator of V_i—can be negative with a small probability, all $\widehat{V}_i$, for $i \in [n]$, are simultaneously positive with probability tending to one. This finding is consistent with our numerical experiments. The proof of Lemma 4.2 is in Section A.14.

Lemma 4.2 (Ratio-consistency)**.** *We have the following ratio-consistency results under the conditions of Theorem 4 on the noise ε:*

1. *Under Condition 1 from Proposition 4.1, and the assumption that the bound in Condition 2 of Proposition 4.1 holds with $\kappa(\Sigma)$ replaced by $\kappa(\Sigma)^2$, we have $w_p^\top S(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top w_p / w_p^\top S\Sigma S^\top w_p \to_P 1$.*
2. *Assume that $\max_{i\in[n]} \mathbb{E}\,|Z_i|^8$ is bounded. For the random design considered in Theorem 2, if $\kappa(\Gamma)$ is bounded and $\kappa(\Sigma)$ is bounded, we have*

$$\max_{i\in[n]} |\widehat{V}_i/V_i - 1| = o_P(1).$$

If we further assume that $x_i = \Gamma^{1/2} z_i$ where z_i has sub-Gaussian entries, and relax the condition on $\kappa(\Sigma)$ from bounded to $\kappa(\Sigma) = o(n^{1/4-\varepsilon})$ for any small constant $\varepsilon > 0$, then the same conclusion holds.

The following result provides an inferential method for contrasts $w_p^\top \beta$. Its proof is built on Lemma 4.2, and is included in Section A.14.

Theorem 5 (Inference for contrasts)**.** *Under the conditions of claim 1 of Lemma 4.2, and further assuming that, with $S = (X^\top X)^{-1} X^\top$,*

$$\sqrt{\kappa(\Sigma)}\frac{\max_{j\in[n]} |w_p^\top S e_j|}{\|w_p^\top S\|} \to 0, \tag{19}$$

we have

$$(w_p^\top \hat{\beta} - w_p^\top \beta)/\sqrt{w_p^\top S \operatorname{diag}\left[(Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon})\right] S^\top w_p} \Rightarrow \mathcal{N}(0,1).$$

Both the condition given in (19) and Condition 2 in Proposition 4.1 are delocalization-type conditions that ensure no observation is too influential along the direction determined by w_p.

Condition 2 of Proposition 4.1 implies that there are no large outliers among all entries of v. Let $u = Tv$. If

$$\kappa(\Sigma) \max_i |u_i|/\|u\| \to 0, \tag{20}$$

then (19) is satisfied. This follows since for any vector $a \in \mathbb{R}^n$, $|a_j|/(\sum_{i=1}^n a_i^2)^{1/2} \leq |a_j|/(\sum_{i=1}^n a_i^4)^{1/4}$, and taking $a = S^\top w_p$ yields the result. In this sense (19) is weaker than Condition 2 in Proposition 4.1.

In this work, we focus on the fundamental setting where the noise is independent of X. It is possible to extend the application of the Hadamard estimator to settings where noise is correlated with the predictor, such as when omitted variables are correlated with the observed covariates (Hsiao and Zhou, 2024).

4.2 Estimation and inference on quadratic forms

Inference on quadratic forms of the type $\beta^\top A\beta$ for some non-random matrix A has broad applications, including global testing for hypotheses such as $\|\beta\|^2 = 0$, and analysis of variance with a growing number of groups (Kline et al., 2020; Guo and Cheng, 2022). These tasks can be effectively addressed using the Hadamard estimator for the noise covariance.

We know that $\hat{\beta}^\top A\hat{\beta} - \operatorname{tr} A \operatorname{Cov}(\hat{\beta})$ is an unbiased estimate of $\beta^\top A\beta$. Plugging-in the Hadamard estimator of $\operatorname{Cov}(\hat{\beta})$, given by $S(\operatorname{diag} \widehat{\vec{\Sigma}})S^\top$, we obtain $\hat{\beta}^\top A\hat{\beta} - \operatorname{tr} AS(\operatorname{diag} \widehat{\vec{\Sigma}})S^\top$, which is an unbiased estimator of $\beta^\top A\beta$. We have the following result, which characterizes the distribution of this estimator and facilitates the inference procedures discussed below.

Theorem 6. *Assume the noise vector ε satisfies the conditions stated in Theorem 4 with bounded $\kappa(\Sigma)$, and that condition 1 of Proposition 4.1 hold. Define $\zeta = \Sigma^{1/2}S^\top A\beta$ and $B = \Sigma^{1/2}S^\top AS\Sigma^{1/2}$. We further require the following conditions:*

1. $\max_{i\in[n]} |\zeta_i|/\|\zeta\| \to 0$.
2. $\sum_{i=1}^n B_{ii}^2/\|B\|_{\mathrm{Fr}}^2 \to 0$.
3. *either* $\frac{\|B\|}{\|B\|_{\mathrm{Fr}}} \to 0$, *or both* $\frac{\|B\|}{\|\zeta\|} \to 0$ *and* $\frac{\|B\|_{\mathrm{Fr}}}{\|B\|} \leq C$ *for some positive constant* C.

Then

$$\frac{\hat{\beta}^\top A\hat{\beta} - \operatorname{tr} AS(\operatorname{diag} \widehat{\vec{\Sigma}})S^\top - \beta^\top A\beta}{\sqrt{2\|B\|_{\mathrm{Fr}}^2 + 4\|\zeta\|^2}} \Rightarrow \mathcal{N}(0,1).$$

When $\zeta = 0$, the conclusion holds under condition 2, when $\|B\|/\|B\|_{\mathrm{Fr}} \to 0$.

The second condition is not required when the third moments of the noise components are zero, which holds for symmetric distributions.

For the last condition, the first part $\|B\|/\|B\|_{\mathrm{Fr}} \to 0$ is a reasonable assumption when A has growing rank. The growing rank scenario is particularly relevant for inference on the variation of fixed effecxts when the number of groups increases in ANOVA, as elaborated below. Moreover, it also enables inference on the global effect of dense but weak signals, such as those characterized by $\|\beta\|^2$.

When the rank of A is fixed, in other words, $\|B\|_{\mathrm{Fr}}/\|B\| \leq C$, then $\|B\|/\|\zeta\| \to 0$ is required. This means that when the inference target is on a quadratic form of the projection of

β onto a low-dimensional space spanned by columns of A, then we require the signal strength, reflected by norm of ζ as β is involved, to be strong enough to achieve valid inference. This is not a surprise since a quadratic form of type $Z^\top AZ$ where Z has i.i.d. entries and A is low-rank may have irregular distribution that depends on the entries of Z. When $\|B\|/\|\zeta\| \to 0$, the fluctuation of the statistic $\hat{\beta}^\top A\hat{\beta}$ is due to a linear form of the type $\zeta^\top Z$, thus the asymptotic normality holds.

Although these conditions appear to preclude inference on $\beta^\top A\beta$ when A is of low rank r and $\beta = 0$, we can instead test $U^\top\beta = 0$ where $U \in \mathbb{R}^{n\times p}$ is the matrix of the eigenvectors of A corresponding to the non-zero eigenvalues. Extending Theorem 5 to a multivariate case, we have $(U^\top S\Sigma S^\top U)^{-1/2}U^\top\hat{\beta} \Rightarrow \mathcal{N}(0, I_r)$ and consequently $\hat{\beta}^\top U(U^\top S\Sigma S^\top U)^{-1}U^\top\hat{\beta} \Rightarrow \chi^2_r$. The asymptotic covariance $U^\top S\Sigma S^\top U$ can be consistently estimated by $U^\top S\widehat{\vec{\Sigma}}S^\top U$ under the same condition as that in Lemma 4.2.

To perform inference on $\beta^\top A\beta$, we need to estimate the asymptotic variance of its estimator, given by $2\|B\|^2_{\mathrm{Fr}} + 4\|\zeta\|^2$ implied by Theorem 6. An application of Theorem 6, with A replaced by $AS\Sigma S^\top A$, yields that $\hat{\beta}^\top AS\Sigma S^\top A\hat{\beta} - \operatorname{tr} AS\Sigma S^\top AS(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top$ is an estimate of $\|\zeta\|^2 = \beta^\top AS\Sigma S^\top A\beta$, under conditions similar to those of Theorem 6. By substituting Σ with $\widehat{\vec{\Sigma}}$, we obtain an estimator of $\|\zeta\|^2$. Additionally, a plug-in estimator of $\|B\|^2_{\mathrm{Fr}}$ can be constructed. Combining these components, we arrive at an estimator of $2\|B\|^2_{\mathrm{Fr}}+4\|\zeta\|^2$, given by

$$4\hat{\beta}^\top AS(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top A\hat{\beta} - 2\operatorname{tr} AS(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top AS(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top. \tag{21}$$

In Section 5 we use simulations to demonstrate that this estimator enables valid inference on $\beta^\top A\beta$ across a broad range of settings. The simplicity of this estimator compared with the one from Kline et al. (2020) could be an advantage. The latter uses sample splitting, which leads to a relatively involved implementation in practice. The theoretical justification of our estimator is left to future work.

Analysis of variance with growing number of groups. The quadratic form $\beta^\top A\beta$ can represent the heterogeneity of fixed effects in ANOVA study by specifying A appropriately (Kline et al., 2020). To illustrate this, we revisit the one-way fixed effects model as in (Kline et al., 2020), where the number of groups can be as large as the the sample size. The one-way ANOVA model takes the form

$$y_i = \alpha_{c(i)} + z_i^\top\mu + \varepsilon_i, i \in [n]$$

where $c(i) \in [1, J]$ is the group label for the i-th observation. Here α_g represent the group effects and $z_i \in \mathbb{R}^d$ are common covariates. This model can be written as $y_i = x_i^\top\beta + \varepsilon_i$, where

$$\beta = (\alpha_1, \cdots, \alpha_J, \mu^\top)^\top, \quad x_i = (\ell_i^\top, z_i^\top)^\top \in \mathbb{R}^p$$

with $\ell_i = (I(i = 1), \cdots, I(i = J))^\top$.

Let n_j denote the number of datapoints that belong to group j. The statistic of interest is

$$\sigma^2_\alpha = \frac{1}{n}\sum_{j=1}^{J} n_j(\alpha_j - \bar{\alpha})^2, \tag{22}$$

where $\bar{\alpha} = \frac{1}{n}\sum_{j=1}^{J} n_j \alpha_j$. This measures the variability of the fixed effect α. It can be written as $\beta^\top A \beta$ where

$$A = \begin{pmatrix} A_\alpha A_\alpha^\top & 0_{J\times d} \\ 0_{d\times J} & 0_{d\times d} \end{pmatrix} \tag{23}$$

with $A_\alpha = \frac{1}{\sqrt{n}}(\ell_1 - \bar{\ell}, \cdots, \ell_n - \bar{\ell})$ and where $\bar{\ell} = n^{-1}\sum_{i=1}^{n} \ell_i$. Theorem 6 implies that $\hat{\beta}^\top A \hat{\beta} - \operatorname{tr} AS(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top$ is a consistent estimator of σ_α^2.

We verify the conditions in Theorem 6 under the one-way ANOVA design without common covariates. In this case, we have $J = p$. Without loss of generality, we consider the following design: $X_{j,1} = 1$ for $j \in [1, n_1]$, $X_{j,2} = 1$ for $j \in [n_1 + 1, n_1 + n_2]$, and so on, until $X_{j,p} = 1$ for $j \in [\sum_{i=1}^{p-1} n_i + 1, \sum_{i=1}^{p} n_i]$. All other entries of X are zero.

Let $C_i = I_{n_i} - \frac{1}{n_i} 1_{n_i} 1_{n_i}$ denote the centering matrix of size n_i. Then $A = X^\top C_n X$. We can find that $S^\top AS$ is a block-diagonal matrix given by $\operatorname{diag}(I_{n_1} - C_1, I_{n_2} - C_2, \cdots, I_{n_p} - C_p)$, and $\zeta = \Sigma^{1/2}\left[(\alpha_1 - \bar{\alpha})1_{n_1}^\top, \cdots, (\alpha_p - \bar{\alpha})1_{n_p}^\top\right]^\top$. As the eigenvalues of Σ are assumed to be bounded away from zero and infinity, Condition 1 in Thereom 6 is equivalent to

$$\max_{i\in[n]} (\alpha_i - \bar{\alpha})^2 / (n\sigma_\alpha^2) \to 0.$$

Condition 2 is equivalent to requiring that the group size grows for at least one group. Condition 3 translates to either $1/\sqrt{p} \to 0$, or $(n\sigma_\alpha^2)^{-1} \to 0$ when $\sqrt{p}$ is fixed. Notably, the second condition is not necessary if the noise has a symmetric distribution.

Relation to the KSS estimator from Kline et al. (2020). The KSS estimator proposed in Kline et al. (2020) uses a leave-one-out approach to estimate Σ_i by $y_i(y_i - x_i^\top \hat{\beta}_{-i})$ where $\hat{\beta}_{-i}$ is the leave-i-th-sample-out estimator of β. In vector form, the estimator of $\vec{\Sigma}$ can be written as $(\operatorname{diag}(Q))^{-1}(y \odot \widehat{\varepsilon})$. This is then used to construct the unbiased estimator of $\beta^\top A \beta$—given by $\hat{\beta}^\top A \hat{\beta} - \operatorname{tr} A \operatorname{Cov}(\hat{\beta})$—via plugging into $\operatorname{Cov}(\hat{\beta}) = S\Sigma S^\top$, recalling that $S = (X^\top X)^{-1} X^\top$.

The validity of the KSS estimator *requires the signal strength to not be too strong*, as reflected by Assumption 1 (iii) in Kline et al. (2020), while the Hadamard estimator does not require this constraint. If the signal is too strong, the KSS estimator tends to yield many negative variance estimates for the noise variance, leading to large type-I error in tests of coordinates, as demonstrated by simulations in Section 5.2. On the other hand, the KSS estimator allows leverage scores to be smaller than any constant smaller than one, while the Hadamard estimator essentially requires leverage scores to be bounded above by $1/2$. Therefore, when the design matrix X is mostly discrete-valued (for instance, in ANOVA settings) and groups are large—say, when each group has only two paired observations—the Hadamard estimator may not be well-defined, as the matrix $Q \odot Q$ becomes singular.

While Hadamard estimators for $\vec{\Sigma}$ differ from KSS estimators, they can yield equivalent estimators for certain special design matrices X of interest. In particular, we have the following result.

Proposition 4.3. *For any one-way ANOVA design without common continuous covariates, the Hadamard, KSS, and MW estimators for* $\operatorname{Cov}(\hat{\beta})$ *are equivalent.*

As a direct consequence, for any one-way ANOVA design without common continuous covariates, the Hadamard, KSS, and MW estimators for σ_α^2 defined in (22) are equivalent.

5 Numerical Results

In this section, we present several numerical simulations supporting our theoretical results. We consider the following cases:

> Case 1: Take X to have i.i.d. standard normal entries, and the noise to be $\varepsilon = \Sigma^{1/2} Z$, where Z has i.i.d. standard normal entries. The noise covariance matrix Σ is the diagonal matrix of eigenvalues of an AR-1 covariance matrix T, with $T_{ij} = \rho^{i-j}$.
>
> Case 2: Take X to have i.i.d. t_{10} entries, and the noise to be $\varepsilon = \Sigma^{1/2} Z$ where Z has i.i.d. standard normal entries. The noise covariance matrix Σ is the diagonal matrix consisting of the coordinates of $|Xc|$ where $c = (1, 0, \ldots, 0)^\top \in \mathbb{R}^p$.
>
> Case 3: Take X to contain both discrete variables representing one-hot encodings for group labels and continuous variables. Specifically, assume that in the first N columns of X, each column contains exactly n/N entries equal to one, and in the remaining $p-N$ columns, the entries are i.i.d. t_{10} random variables. The noise is generated in the same way as in Case 1. Fix $N = 100$, and $\rho = 0.9$.

We set β to zero for most of the simulations in Sections 5.1-5.4 which are related to inference on coordinates of β, unless otherwise stated. The Hadamard estimator, the MW estimator and the White estimator of the noise variances do not depend on β, that is, for y generated from linear models with different values of β, the variance estimators depend only on X and ε. An exception is the leave one out (KSS) estimator proposed in Kline et al. (2020) as reviewed in Section 4.2, and some simulations in Section 5.2 show the difference. In Section 5.5, we conduct simulations to evaluate the method in Section 4.2 and provide further comparisons with the KSS estimator for estimating quadratic forms.

5.1 Mean type I error over all coordinates

We show the mean type I error of the normal confidence intervals based on the White, MacKinnon-White, and Hadamard methods over all coordinates of the OLS estimator.

In Figure 1, we show the results for Case 1. We take $n = 1000$, and three aspect ratios, $\gamma = 0.1, 0.5, 0.75$, varying p. We consider $\rho = 0$ (homoskedasticity), and $\rho = 0.9$ (heteroskedasticity). We draw one instance of X, and draw 1000 Monte Carlo repetitions of ε.

We observe that the CIs based on White's covariance matrix estimator are inaccurate for the aspect ratios considered. They have inflated type I error rates. All other estimators are more accurate. The MW confidence intervals are quite accurate for each configuration. The Hadamard estimator using the degrees of freedom correction is comparable, and noticably better if the dimension is high.

5.2 Coordinate-wise type I error

Figure 2 displays the mean type I error of each coordinate in Case 2. We also compare with the leave out (KSS) estimator developed in Kline et al. (2020). The results for Case 1 are in Section B of the appendix. To measure the overall accuracy, we also report the mean absolute deviation $\mathrm{MAD} = p^{-1} \sum_{j=1}^{p} |\mathrm{err}_j - 0.05|$ where err_j refers to the type I error for the j-th coordinate. The White estimator has inflated type I errors especially for larger p. We also observe that

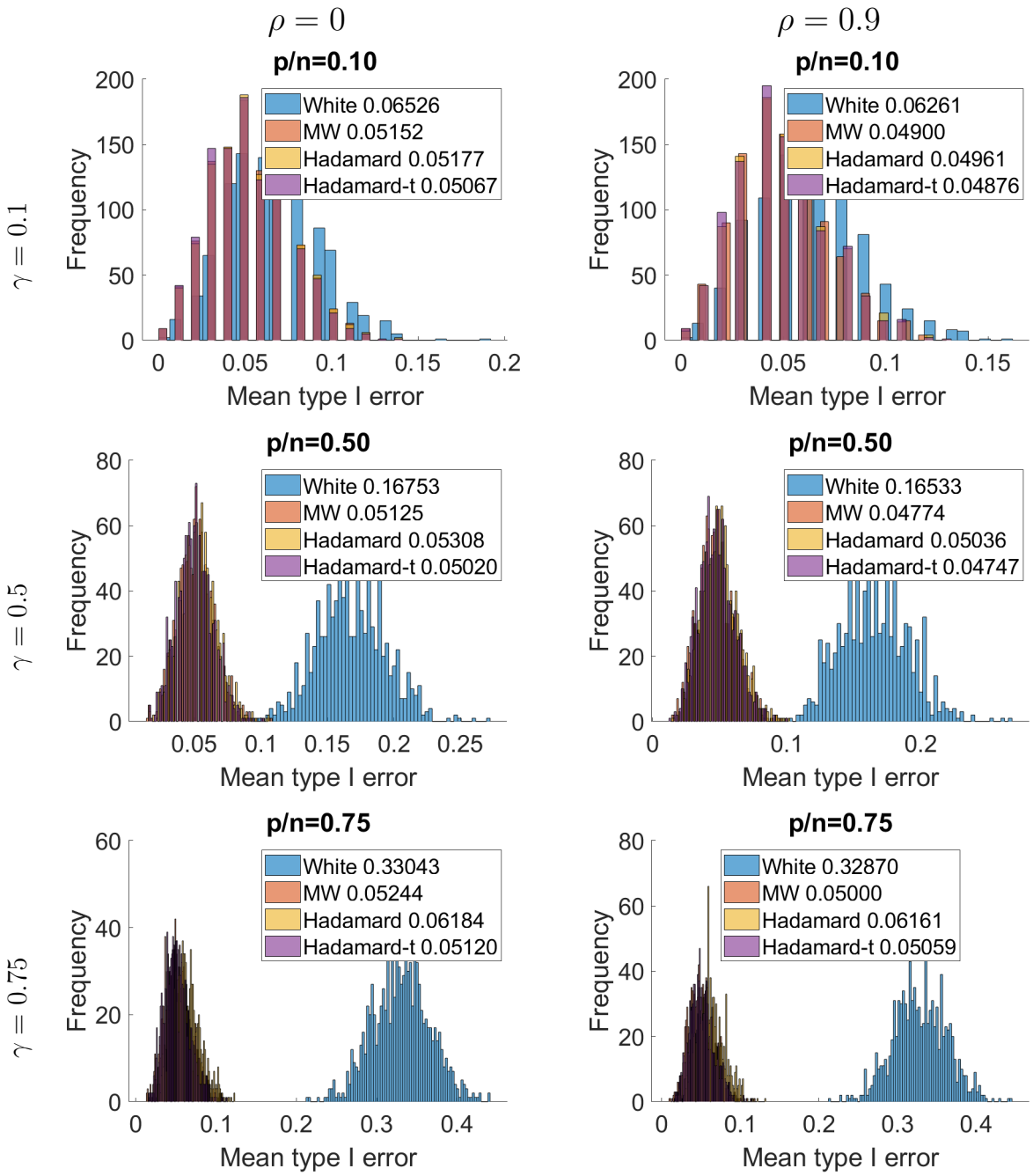


Figure 1: Mean type I error over all coordinates.

although the MAD of the Hadamard estimator is large when $p = 800$, the degrees-of-freedom adjustment significantly improves the performance and achieves performance comparable to the MW estimator. The performance of the KSS estimator resembles that of the Hadamard-t estimator for Case 2.

Figure 3 displays the mean type I error for each coordinate in Case 3. The coefficient vector

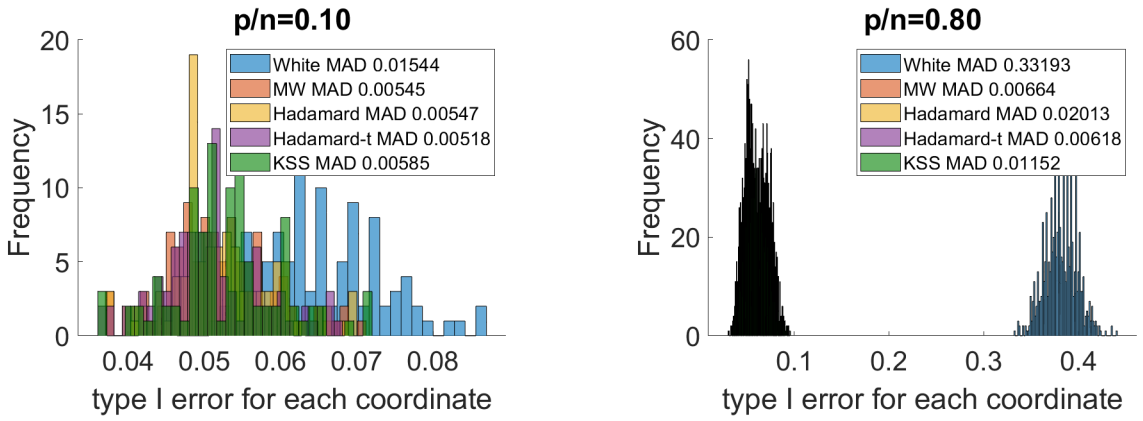


Figure 2: Mean type-I error for each coordinate over 1000 simulations for Case 2.

β is generated with coordinates drawn independently from $\mathcal{N}(0, 9)$. When $p = 100$, the design matrix corresponds to the ANOVA design discussed at the end of Section 2.2. In this setting, the MW, Hadamard, and KSS estimator coincide exactly, as shown in the left panel of Figure 3. We also observe that the Hadamard-t estimator significantly improves the performance.

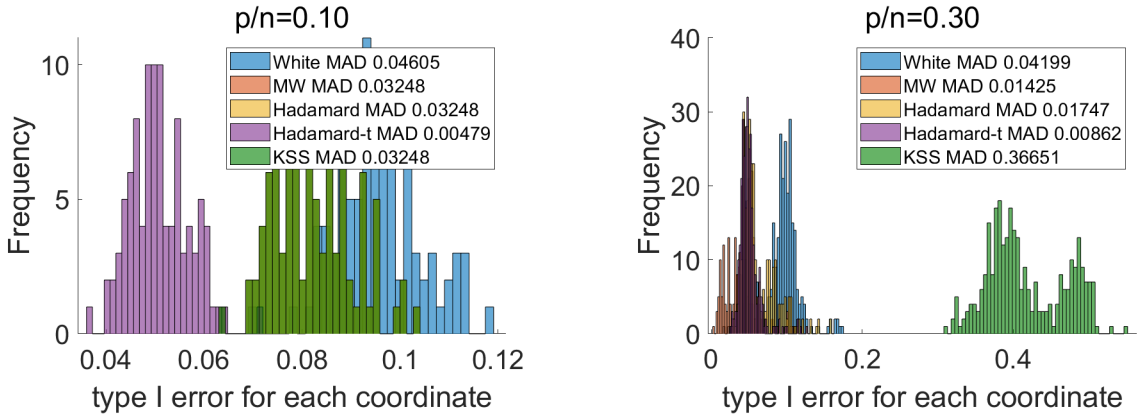


Figure 3: Mean type-I error for each coordinate over 1000 simulations for Case 3.

Figure 4 further shows that when $p/n = 0.3$, the type I error of the MW estimator for the first one hundred variables, which correspond to the discrete group label variables in the design matrix, fluctuates around the nominal level of 0.05. The Hadamard estimator exhibits some undercoverage for these variables. The Hadamard-t estimator substantially improves the performance, which can be explained by the necessity of a degrees-of-freedom adjustment for inference of a group fixed effect when each group contains only ten observations.

The undesirable performance of the KSS estimator can be attributed to the strong signal in this case. This results in roughly 38.9% of all KSS variance estimators $\text{Var}(\hat{\beta}_i)$ being negative-valued across 1000 replications and all coordinates i, with individual coordinates showing proportions ranging from 0.259 to 0.544. Thus, the KSS estimator has a non-negligible probability of being undefined. The other estimators do not have this issue. In Section B.4 of the

appendix, we include additional simulations where the noise variances tend to infinity, resulting in weak signal-to-noise ratios.

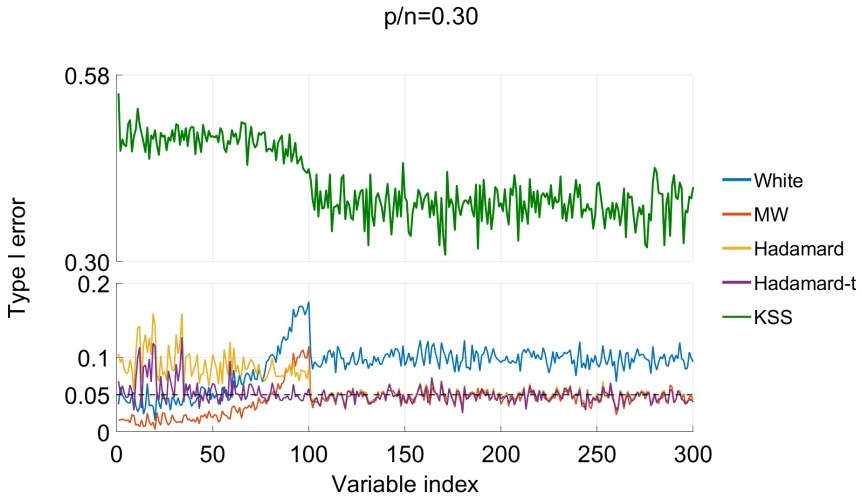


Figure 4: Type I error as a function of variable index in the right panel of Figure 3.

To further compare with other methods, we plot the mean type I error as a function of p by taking p equally spaced from 100 to 800 with gaps of 100, see Figure 5. The results including the KSS estimator and for Case 1 are reported in Section B of the appendix. We observe that the MW estimators are liberal for the first coordinate but accurate for the second coordinate. Section B.1 contains a discussion of when the MW estimator is problematic, and explains the performance of MW estimators in this experiment. The CI based on the Hadamard estimator has a slightly inflated type I error for both coordinates for larger p, but the Hadamard-t estimator is accurate.

For comparison, we also conduct experiments using two variants of the bootstrap. First, we use the pairs bootstrap (Freedman, 1981), where each observation of the bootstrap sample $[X^*, Y^*]$ is sampled randomly with replacement from the rows of $[X, Y]$, see Figure 5. We also include the residual bootstrap, which samples with replacement the residuals $Y - X\hat{\beta}$, and adds them to $X\hat{\beta}$ to get the new Y^*. Intuitively, this method is justified if the error terms are independent and identically distributed (see e.g., MacKinnon (2006) for a discussion). The corresponding results are in Figure 11.

Since the rows are sampled with replacement for the pairs bootstrap, when p is close to n, the resampled $X^{*\top}X^*$ may be ill-conditioned or non-invertible. Figure 5 shows the results where the matrix inversion is done using a pseudoinverse when $p > 500$. This leads to unstable confidence intervals, coverage and length. The lengths of the confidence intervals are in the right panel of Figure 11.

We also conduct experiments using the jackknife (or equivalently HC3 in MacKinnon and White (1985)), and see from Figures 10 and 12 that the jackknife is not accurate. This can be explained by the fact that the expression for the jackknife estimator therein was derived for a relatively small dimension p.

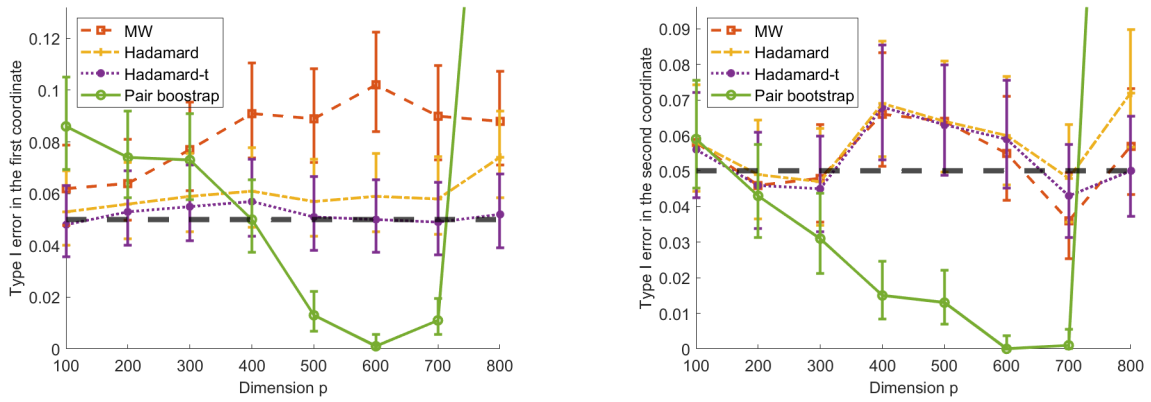
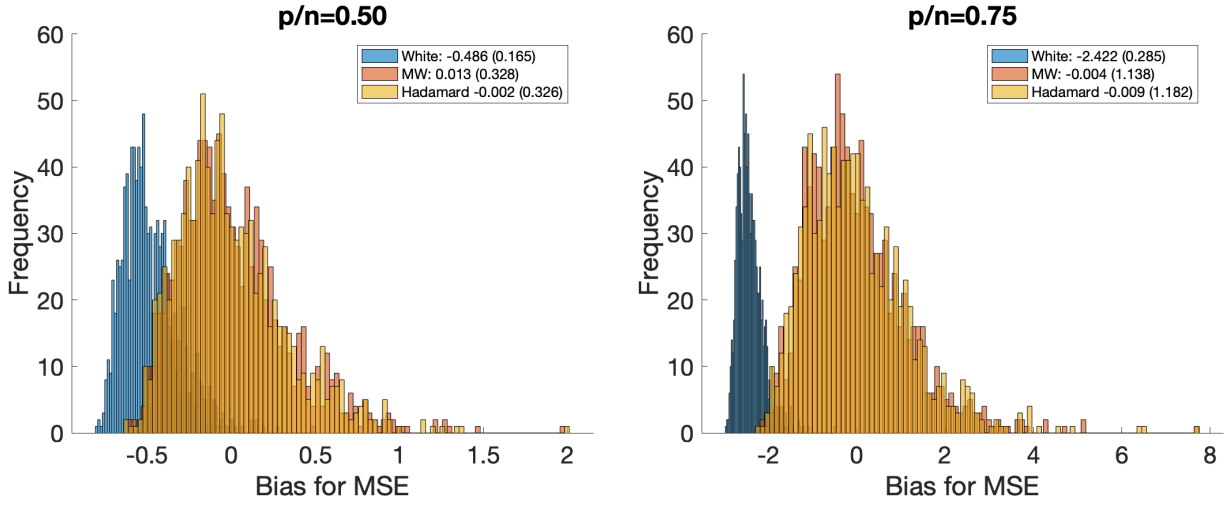


Figure 5: Mean type-I error in the first coordinate and second coordinate over 1000 simulations each. The error bars represent 95% Clopper-Pearson intervals for the coverage.

5.3 Estimating the MSE



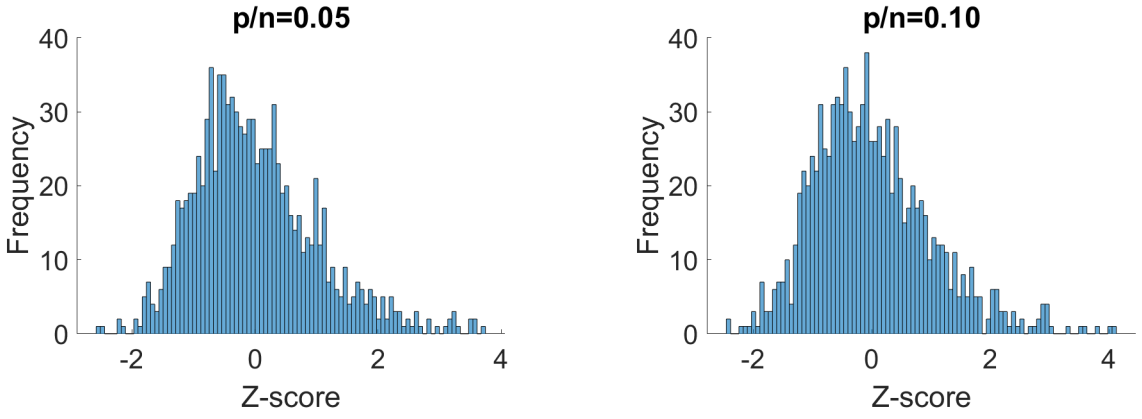
(a) $p/n = 0.5$ (b) $p/n = 0.75$

Figure 6: Bias in estimating MSE.

In Figure 6, we show the bias in estimating the MSE of the OLS estimator in Case 1 for the three methods where the numbers outside the bracket correspond to the mean biases of the 1000 Monte Carlo runs and the numbers inside the bracket stand for the standard deviation. For each method, we use the estimator which equals the sum of the variances of the individual component estimators.

The results are in line with those from the previous sections. Both MacKinnon-White's and the Hadamard estimator perform much better than the White estimator. In addition, the Hadamard estimator is usually comparable to MacKinnon-White's. More results for Case 2 are included in Section B of the appendix.

5.4 Approximate Normality



(a) $p/n = 0.05$ (b) $p/n = 0.1$

Figure 7: Distribution of z-scores of a fixed coordinate of the Hadamard estimator.

In Figure 7, we show the distribution of z-scores of a fixed coordinate of the Hadamard estimator in Case 1. We use a similar setup to the previous sections, but we choose a larger sample size $n = 1,000$, and also smaller aspect ratios $p/n = 0.05$ and $p/n = 0.1$. We observe a relatively good fit to the normal distribution, but it is also clear that a chi-squared approximation may lead to a better fit.

5.5 Estimating the quadratic form $\beta^\top A\beta$

We evaluate the performance of the Hadamard estimator and the KSS estimator for quadratic forms $\beta^\top A\beta$ under both continuous and ANOVA design matrices. We draw 1000 Monte Carlo repetitions of ε for each setting.

Table 1 reports the mean and standard deviation of the Hadamard estimator and KSS estimators for $\|\beta\|^2$ under design in Case 1, and β is generated as $s \cdot Z$, with Z drawn from a standard multivariate normal distribution, and $s = 1$. The results indicate that the Hadamard estimator exhibits high accuracy and performs comparably to the KSS estimator. Additionally, we report the coverage error of confidence intervals for $\|\beta\|^2$, using the Hadamard methods described in Equation (21), which closely aligns with the nominal significance level of 0.05. Further results for varying signal strengths (i.e., smaller s) are provided in Section B.5 of the appendix.

Table 2 presents the results for estimating (22) under an ANOVA design, as discussed in Section 4.2. Here we consider a model without continuous covariates, and p represents the number of groups. The group means α_i, for $i \in [p]$, are independently generated from a $\mathcal{N}(0, 100)$ distribution. For each group, the number of datapoints is chosen uniformly at random between three and $\lceil 2n/p \rceil$. The lower bound of three ensures that the Hadamard estimator is well defined, as explained at the end of Section 2.4. The noise is generated as in Case 2. Table 2 shows that the Hadamard estimator performs well and matches the KSS estimator exactly, consistent with the theoretical equivalence established in Proposition 4.3.

Further simulations under weaker signal strengths and with additional continuous covariates are reported in Section B.5 of the appendix.

Table 1: Comparison of Hadamard and KSS Estimators for Case 1

p	Hadamard			KSS		True Value
	Mean	SD	Coverage	Mean	SD	
50	57.2678	0.4660	0.042	57.2678	0.4665	57.2787
100	101.7548	0.7327	0.053	101.7546	0.7323	101.7473
150	156.4364	0.9363	0.055	156.4364	0.9352	156.4411
200	192.5218	1.0060	0.044	192.5218	1.0077	192.5599
250	283.9984	1.2391	0.050	283.9988	1.2399	283.9917
300	266.6129	1.1814	0.059	266.6129	1.1853	266.6609

Table 2: Comparison of Hadamard and KSS Estimators for σ_α^2

# groups	Hadamard			KSS		True Value
	Mean	SD	Coverage	Mean	SD	
50	104.8587	0.5962	0.0500	104.8587	0.5962	104.8356
100	92.7884	0.5685	0.0410	92.7884	0.5685	92.7796
150	108.8100	0.6129	0.0480	108.8100	0.6129	108.8000
200	91.5151	0.5893	0.0550	91.5151	0.5893	91.5247
250	102.8100	0.6198	0.0470	102.8100	0.6198	102.7500
300	97.1664	0.5807	0.0560	97.1664	0.5807	97.1904

5.6 Empirical data analysis example

The PISA (Programme for International Student Assessment) has become a gold standard for evaluating education systems worldwide. It records student performance across mathematics, science, and reading from over 80 countries and districts, tracking progress since 2000. We use the PISA dataset across several years to evaluate the variation of grades attributed to both country and parents' education level.

We consider a two-way fixed effects model $y_i = \alpha_{c(i)} + \eta_{e(i)} + z_i^\top \mu + \varepsilon_i$ where $c(i)$ and $e(i)$ represent the country group and education group for the i-th student, respectively. Denote the number of countries by J and parents' education levels by K; then $c(i) \in [1:J]$ and $e(i) \in [1:K]$. The covariates z_i include the socioeconomic status and gender. This model allows each individual to have more than one group membership. Let n_j denote the number of students in country j and m_k the number of students whose parents have education level k. We aim to conduct inference on

$$\sigma_\alpha^2 = \frac{1}{n}\sum_{j=1}^{J} n_j(\alpha_j - \bar{\alpha})^2, \quad \sigma_\eta^2 = \frac{1}{n}\sum_{k=1}^{K} m_k(\eta_k - \bar{\eta})^2,$$

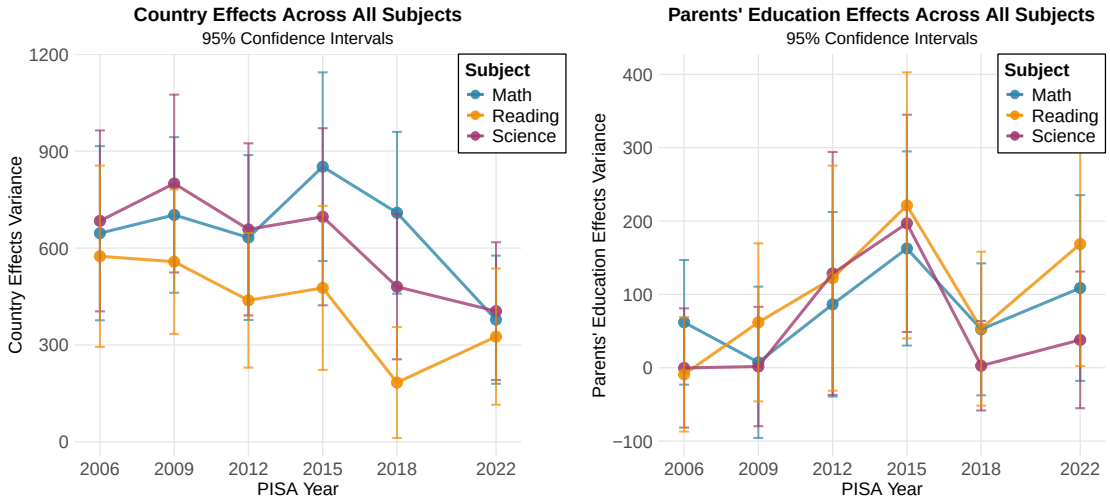


Figure 8: Variance in students' performance across three subjects attributed to country and parental education effects from 2006 to 2022.

where $\bar{\alpha} = n^{-1}\sum_{j=1}^{J} n_j\alpha_j$ and $\bar{\eta} = n^{-1}\sum_{k=1}^{K} m_k\eta_k$. These two quantities represent the variation of students' performance due to the country effect and the effect of the parents' education level, respectively. They can be expressed in the form $\beta^\top A\beta$ where the matrix A is defined analogously to the form in (23).

Since the PISA tests mainly focus on three subjects—mathematics, science, and reading—we will consider these measures for each subject and over several years. We use the subsampled PISA data from Wang et al. (2024). We preprocess their data by removing NA values, keeping groups with at least three observations to avoid singularity issues. The parents' education is a factor that combines the father's and mother's education levels. After these steps, we obtain six datasets from 2006 to 2022, which have sample sizes ranging from 1,135 to 1,256 for different years, and contain 36-38 country groups and 10-16 education groups.

Figure 8 illustrates the contributions of country and parents' education effect to student performance in three subjects from 2006 to 2022. A notable pattern emerges: the country effect is consistently weaker in reading compared to math and science, while the parental education effect tends to be stronger in reading than in math and science. These trends are stable over time, especially for the country effect.

6 Discussion and Future Work

In this paper, we have provided several fundamental theoretical results for the Hadamard estimator, which can be used to construct confidence intervals for the OLS estimator of coefficients in a linear model under heteroskedasticity. We showed that the Hadamard estimator is well-defined and well-conditioned for certain random design models. There are several important directions for future research. Can one develop similar results for nonlinear models? Is it possible to establish the non-negativity of the Hadamard estimator, possibly with some regularization? Is it possible to show approximate coverage results for our t-confidence intervals

based on the degrees of freedom correction as given in (11)? Such results have been obtained in the low-dimensional case by Kauermann and Carroll (2001), for instance. However, establishing such results in high dimensions seems to require different techniques.

Beyond our current investigations, an important direction is the development of tests for heteroskedasticity. White's original paper proposed such a test based on comparing his covariance estimator to the usual one under homoskedasticity. There are many other well-known proposals (Dette and Munk, 1998; Azzalini and Bowman, 1993; Cook and Weisberg, 1983; Breusch and Pagan, 1979; Wang et al., 2018). Perhaps most closely related to our work, Li and Yao (2019) have proposed tests for heteroskedasticity with good properties in low and high dimensional settings. Their tests rely on computing measures of variability of the estimated residuals, including the ratio of the arithmetic and geometric means, as well as the coefficient of variation. Their works and follow-ups such as Bai et al. (2016, 2018) show central limit theorems for these test statistics. They also show an improved empirical power compared to some classical tests for heteroskedasticity. It would be of interest to see if our covariance matrix estimator could be used to develop new tests for heteroskedasticity.

An important extension of the heteroskedastic model is the clustered observations model. Liang and Zeger (1986) proposed estimating equations for such longitudinal/clustered data. They allowed arbitrarily correlated observations for any fixed individual (i.e., within each cluster), and proposed a consistent covariance estimator in the low-dimensional setting. Can one extend our ideas to the clustered case?

Another important direction is to develop covariance estimators that have good performance in the presence of both heteroskedasticity and autocorrelation. The most well-known example is possibly the popular Newey-West estimator (West and Newey, 1987), which is a sum of symmetrized lagged autocovariance matrices with decaying weights. Is it possible to develop new methods inspired by our ideas suitable for this setting?

Our paper does not touch on the interesting but challenging regime where $n < p$. In that setting, Buhlmann, Dezeure, Zhang, (Dezeure et al., 2017) proposed bootstrap methods for inference with the lasso under heteroskedasticity, under the limited ultra-sparse regime, where the sparsity s of the regression parameter is $s \ll n^{1/2}$. These methods are limited as they apply only to the lasso, and because they only concern the ultra-sparse regime. It would be interesting to understand this regime better.

It is possible that one could sometimes get better confidence intervals by adding some regularization to the starting estimator, such as starting with an ℓ_2 regularized regression estimator. However, note that this would come at the cost of introducing some bias, so that each confidence interval would not be centered at the true parameter anymore. It is possible that by an appropriately small regularization, one could achieve a favorable "bias-variance" tradeoff for confidence intervals. However, the specific details are likely complex (e.g., how to tune the regularization) and deserve a separate investigation.

7 Acknowledgements

The authors thank Matias Cattaneo and Jason Klusowski for valuable discussions and feedback on an earlier version of the manuscript. This work is supported in part by the NSF grant DMS 2046874. Zhixiang Zhang is partially supported by the SRG2023-00053-FST and MYRG-GRG2024-00260-FST-UMDF from University of Macau, and the National Natural Science

Foundation of China (Grant No. 12401331).

References

A. Azzalini and A. Bowman. On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(2): 549–557, 1993.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer, 2010.

Z. Bai, G. Pan, and Y. Yin. Homoscedasticity tests for both low and high-dimensional fixed design regressions. *arXiv preprint arXiv:1603.03830*, 2016.

Z. Bai, G. Pan, and Y. Yin. A central limit theorem for sums of functions of residuals in a high-dimensional regression model with an application to variance homoscedasticity test. *Test*, 27:896–920, 2018.

R. M. Bell and D. F. McCaffrey. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–182, 2002.

A. K. Bera, T. Suprayitno, and G. Premaratne. On some heteroskedasticity-robust estimators of variance–covariance matrix of the least-squares estimators. *Journal of Statisticsistical Planning and Inference*, 108(1-2):121–136, 2002.

P. Billingsley. Probability and measure. 3rd wiley. *New York*, 1995.

T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 47(5):1287–1294, 1979.

M. Capitaine, C. Donati-Martin, and D. Féral. The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *The Annals of Probability*, 37(1):1–47, 2009.

M. D. Cattaneo, M. Jansson, and W. K. Newey. Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361, 2018.

M. Chen, C. Gao, and Z. Ren. A general decision theory for huber's epsilon-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

V. Chew. Covariance matrix estimation in linear models. *Journal of the American Statistical Association*, 65(329):173–181, 1970.

R. D. Cook and S. Weisberg. Diagnostics for heteroscedasticity in regression. *Biometrika*, 70 (1):1–10, 1983.

H. Dette and A. Munk. Testing heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):693–708, 1998.

R. Dezeure, P. Bühlmann, and C.-H. Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26:685–719, 2017.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2017.

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

D. Donoho and A. Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

F. Eicker. Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 59–82, 1967.

N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

D. A. Freedman. Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228, 1981.

W. H. Greene. *Econometric analysis*. Pearson, 2003.

X. Guo and G. Cheng. Moderate-dimensional inferences on quadratic functionals in ordinary least squares. *Journal of the American Statistical Association*, 117(540):1931–1950, 2022.

H. Hartley, J. Rao, and G. Kiefer. Variance estimation with one unit per stratum. *Journal of the American Statistical Association*, 64(327):841–851, 1969.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.

R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, 1994.

C. Hsiao and Q. Zhou. Statistical inference for the low dimensional parameters of linear regression models in the presence of high-dimensional data: An orthogonal projection approach. *Journal of Econometrics*, page 105851, 2024.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA, 1967.

P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley, 2011.

G. W. Imbens and M. Kolesar. Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712, 2016.

L. Janson, R. F. Barber, and E. Candès. Eigenprism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79 (4):1037–1065, 2017.

G. Kauermann and R. J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001.

H. H. Kelejian and I. R. Prucha. On the asymptotic distribution of the moran i test statistic with applications. *Journal of econometrics*, 104(2):219–257, 2001.

P. Kline, R. Saggio, and M. Sølvsten. Leave-out estimation of variance components. *Econometrica*, 88(5):1859–1898, 2020.

E. Lehmann and G. Casella. Theory of point estimation. *Springer Texts in Statistics*, 1998.

Z. Li and J. Yao. Testing for heteroscedasticity in high-dimensional regressions. *Econometrics and Statistics*, 9:122–139, 2019.

K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

J. G. MacKinnon. Bootstrap methods in econometrics. *Economic Record*, 82:S2–S18, 2006.

J. G. MacKinnon and H. White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325, 1985.

D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statisticsistical Planning and Inference*, 150:1–29, 2014.

P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era—concepts and misconceptions. *Nature reviews genetics*, 9(4):255–266, 2008.

H. Wang, P.-S. Zhong, and Y. Cui. Empirical likelihood ratio tests for coefficients in high-dimensional heteroscedastic linear models. *Statistica Sinica*, 28(4):2409–2433, 2018.

K. Wang, P. Yacobellis, E. Siregar, S. Romanes, K. Fitter, G. V. Dalla Riva, D. Cook, N. Tierney, P. Dingorkar, S. Sai Subramanian, and G. R. Chen. *learningtower: OECD PISA Datasets from 2000-2022 in an Easy-to-Use Format*, 2024. URL `https://kevinwang09.github.io/learningtower/`. R package version 1.1.0, https://github.com/kevinwang09/learningtower.

K. D. West and W. K. Newey. A simple, positive semi-definite, heteroskedasticity and auto-correlation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

C.-F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.

F. Yang. Linear spectral statistics of eigenvectors of anisotropic sample covariance matrices. *arXiv preprint arXiv:2005.00999*, 2020.

J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.

L. Zhou and H. Zou. Cross-fitted residual regression for high-dimensional heteroscedasticity pursuit. *Journal of the American Statistical Association*, 118(542):1056–1065, 2023.

W.-X. Zhou, K. Bose, J. Fan, and H. Liu. A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Annals of Statistics*, 46(5):1904, 2018.

Appendix

Notation. For two positive sequences $(a_n)_{n\geq 1}$, $(b_n)_{n\geq 1}$, we write $a_n \asymp b_n$ if $C^{-1}b_n \leq a_n \leq Cb_n$ for some positive constant C.

A Proofs

A.1 Proof of unbiasedness of the Hadamard estimator

We consider estimators of the vector of variances of $\hat{\beta}$ of the form $\widehat{V} = A \cdot (\widehat{\varepsilon} \odot \widehat{\varepsilon})$ where A is a $p \times n$ matrix, and $M \odot M$ is the element-wise (or Hadamard) product of the vector or matrix M with itself. Our goal is to find A such that $\mathbb{E}\,\widehat{V} = V$, where $V = \operatorname{diag} \operatorname{Cov}(\hat{\beta})$. Here the diag operator returns the vector of diagonal entries of the matrix M, that is $\operatorname{diag} M = (M_{11}, M_{22}, \ldots, M_{nn})^\top$.

Recall that $S = (X^\top X)^{-1} X^\top$ is a $p \times n$ matrix. We have that $\hat{\beta} = Sy = S\varepsilon + \beta$. Since $\operatorname{Cov}(\varepsilon) = \Sigma$, we have that $\operatorname{Cov}(\hat{\beta}) = S\Sigma S^\top$. Thus, our goal is to find unbiased estimates of

the diagonal of this matrix. The following key lemma re-expresses that diagonal in terms of Hadamard products:

Lemma A.1. *Let v be a zero-mean random vector, and M be a fixed matrix. Then,*

$$\mathbb{E}(M \odot M)(v \odot v) = \operatorname{diag}[M \operatorname{diag} \operatorname{Cov}(v) M^\top].$$

In particular, let Σ be a diagonal matrix. and let $\vec{\Sigma}$ be the vector of diagonal entries of Σ. Then

$$(M \odot M)\vec{\Sigma} = \operatorname{diag}[M \Sigma M^\top].$$

Alternatively, let u be a vector. Then $(M \odot M)u = \operatorname{diag}[M \operatorname{diag}(u) M^\top]$.

Proof. Suppose M has k rows, and denote them by m_i, $i \in [k]$. Let also $\Sigma = \operatorname{diag} \operatorname{Cov}(v)$. Then, for any $i \in [k]$, the i-th entry of the left hand side equals, with l denoting the number of columns of M,

$$\mathbb{E}(m_i \odot m_i)^\top (v \odot v) = \mathbb{E} \sum_{j \in [l]} m_{ij}^2 v_{j \in [l]}^2 = \sum_{j \in [l]} m_{ij}^2 \Sigma_j.$$

The i-th entry of the right hand side equals $m_i^\top \Sigma m_i = \sum_{j \in [l]} m_{ij}^2 \Sigma_{j \in [l]}$. Thus, the two sides are equal, which proves the first claim of the lemma. The second claim follows directly from the first claim, from the special case when the covariance of v is diagonal. The third claim is simply a restatement of the second one.

□

We now apply the lemma as follows.

1. Let us use the lemma for $v = \varepsilon$ and $M = S$. Notice that we have $\operatorname{Cov}(v) = \Sigma$ is diagonal, so the right hand side of the lemma is $\operatorname{diag} S\Sigma S^\top = \operatorname{diag} \operatorname{Cov}(\hat{\beta})$, where the equality follows from our calculation before the lemma. Moreover, the left hand side is $\mathbb{E}(S \odot S)(\varepsilon \odot \varepsilon) = (S \odot S)\vec{\Sigma}$, where we vectorize Σ, writing $\vec{\Sigma} = (\Sigma_{11}, \dots, \Sigma_{nn})^\top$. The equality follows because $\operatorname{Cov}(\varepsilon) = \Sigma$ is diagonal. Thus, by the lemma, we have $V = \operatorname{diag} \operatorname{Cov}(\hat{\beta}) = (S \odot S)\vec{\Sigma}$.
2. Let us now use the lemma for a second time, with $M = I$ and $v = \widehat{\varepsilon}$. This shows that $\mathbb{E}(\widehat{\varepsilon} \odot \widehat{\varepsilon}) = \operatorname{diag} \operatorname{Cov}(\widehat{\varepsilon})$. By linearity of expectation, we obtain $\mathbb{E}\, \widehat{V} = A \cdot \mathbb{E}(\widehat{\varepsilon} \odot \widehat{\varepsilon}) = A \cdot \operatorname{diag} \operatorname{Cov}(\widehat{\varepsilon})$.
3. Finally, let us use the lemma for the third time, with $M = Q$ and $v = \varepsilon$. As in the first case, the left hand side equals $\mathbb{E}(\widehat{\varepsilon} \odot \widehat{\varepsilon}) = (Q \odot Q)\vec{\Sigma}$. The right hand side equals $\operatorname{diag}[M \operatorname{diag} \operatorname{Cov}(v) M^\top] = \operatorname{diag} Q\Sigma Q$, where we used that Q is a symmetric matrix. Now, $\operatorname{Cov}(\widehat{\varepsilon}) = \operatorname{Cov}(Q\varepsilon) = Q\Sigma Q$. Thus, we conclude $\operatorname{diag} \operatorname{Cov}(\widehat{\varepsilon}) = \operatorname{diag} Q\Sigma Q = (Q \odot Q)\vec{\Sigma}$.

Putting these conclusions together, we obtain that $\widehat{V}$ is unbiased, namely $\mathbb{E}\, \widehat{V} = \operatorname{diag} \operatorname{Cov}(\hat{\beta})$, if $A(Q \odot Q)\vec{\Sigma} = (S \odot S)\vec{\Sigma}$. This system of linear equations holds for any Σ if and only if $A(Q \odot Q) = (S \odot S)$. If $Q \odot Q$ is invertible, then we have (7). This shows that the original estimator $\widehat{V}$ has the required form, finishing the proof.

For a sequence $(w_p)_{p \geq 1}$, we can also estimate $w_p^\top S\Sigma S^\top w_p$, the variance of $w_p^\top \hat{\beta}$. Since we can use $(Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon})$ to estimate $\operatorname{diag}(\Sigma)$ in an unbiased way, and considering that Σ is diagonal, an unbiased estimator of $w_p^\top S\Sigma S^\top w_p$ is $w_p^\top S \operatorname{diag}[(Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon})] S^\top w_p$.

A.2 Proof of Proposition 2.1

To prove the lower bound, we first claim that for any symmetric matrix A,

$$\operatorname{rank} A \odot A \le \binom{\operatorname{rank} A + 1}{2}.$$

Therefore, in order for $Q \odot Q$ to be invertible, we need $n \le \binom{n-p+1}{2}$. By solving the quadratic inequality, this is equivalent to $p \le [2n + 1 - (8n + 1)^{1/2}]/2$.

To prove the claim about ranks, let r be the rank of A, and let $A = \sum_{i=1}^{r} v_i v_i^\top$ be its eigendecomposition. Here $(v_i)_{i\in[r]}$ are orthogonal, but not necessarily of unit norm. Then,

$$A \odot A = \left(\sum_{i=1}^{r} v_i v_i^\top\right) \odot \left(\sum_{i=1}^{r} v_i v_i^\top\right) = \sum_{i=1}^{r} (v_i \odot v_i)(v_i \odot v_i)^\top + 2 \sum_{1\le i<j\le r} (v_i \odot v_j)(v_i \odot v_j)^\top.$$

This shows that the rank of $A \odot A$ is at most $r + r(r-1)/2 = r(r+1)/2$, as desired.

A.3 Proof of Theorem 2

Our first step is to reduce to the case $\Gamma = I_p$. Indeed, we notice that we can write $X = Z\Gamma^{1/2}$, where Z is the matrix with rows z_i. Hence, $Q = I_n - X(X^\top X)^{-1}X^\top = I_n - Z(Z^\top Z)^{-1}Z^\top$. Therefore, we can take $\Gamma = I_p$.

The next step is to reduce the bounds on eigenvalues to bounds on certain quadratic forms. For all $i \in [p]$, let us define the $n \times n$ matrices $R_i = X^\top X - x_i x_i^\top = \sum_{j\ne i} x_j x_j^\top$. See Section A.4 for a proof of the following result.

Lemma A.2 (Reduction to quadratic forms)**.** *We have the following two bounds on the eigenvalues of $T = Q \odot Q$ with $Q = I_n - X(X^\top X)^{-1}X^\top$:*

$$\lambda_{\max}(T) \le \max_{i\in[p]} \frac{1}{1 + x_i^\top R_i^{-1} x_i},$$

and

$$\lambda_{\min}(T) \ge \min_{i\in[p]} \frac{1 - x_i^\top R_i^{-1} x_i}{(1 + x_i^\top R_i^{-1} x_i)^2}.$$

To bound these expressions, we will use the following well-known statement about concentration of quadratic forms; its short proof is provided in Section A.5.

Lemma A.3 (Concentration of quadratic forms, consequence of Lemma B.26 in Bai and Silverstein (2010))**.** *Let $x \in \mathbb{R}^p$ be a random vector with independent entries and $\mathbb{E}[x] = 0$, for which $\mathbb{E}\left[(\sqrt{n}x_i)^2\right] = \sigma^2$ and $\sup_i \mathbb{E}\left[(\sqrt{n}x_i)^{4+\eta}\right] < C$ for some $\eta > 0$ and $C < \infty$. Moreover, let A_p be a random $p \times p$ symmetric matrix independent of x, with uniformly bounded eigenvalues. Then*

$$\mathbb{P}(|x^\top A_p x - n^{-1}\sigma^2 \operatorname{tr} A_p|^{2+\eta/2} > t) \le Ct^{-1}n^{-(1+\eta/4)}.$$

Proceeding with the proof of Theorem 2, we can scale X so that the variances of the entries of X are $1/n$. Define the following events:

$$\Xi_1 = \bigcap_{i=1}^{n} \left\{ \left| \frac{1}{n} \operatorname{tr} R_i^{-1} - \frac{\gamma_{p,n}}{1-\gamma_{p,n}} \right| \le n^{-0.9999} \right\},$$
$$\Xi_2 = \bigcap_{i=1}^{n} \left\{ \left| x_i^\top R_i^{-1} x_i - n^{-1} \operatorname{tr} R_i^{-1} \right| < \xi \right\}.$$

We obtain $P(\Xi_2^c) \le \xi^{-4-\delta/2} n^{-1-\delta/4}$ from Lemma A.3 and by taking a union bound over $[n]$. Next, we will verify that $\mathbb{P}(\Xi_1^c) \le 5n^{-1-\delta/4}$. Then by Lemma A.2, and applying the inequality $\mathbb{P}(A \cap B) \ge 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c)$ with $A = \Xi_1$, $B = \Xi_2$, the proof will be concluded.

Now to bound $\mathbb{P}(\Xi_1^c)$, according to the rank inequality, see Theorem A.43 of Bai and Silverstein (2010), we can equivalently show $\mathbb{P}\left(\left| n^{-1} \operatorname{tr} R^{-1} - \frac{\gamma_{p,n}}{1-\gamma_{p,n}} \right| \le n^{-0.9999} \right) > 1 - 5n^{-1-\delta/4}$ for sufficiently large n. We further define

$$\Omega_1 := \left\{ \max_{i \in [n],\, j \in [p]} |x_{ij}| \le n^{-0.001} \right\},$$
$$\Omega_2 := \bigcap_{i=1}^{n} \{ (1 - \sqrt{\gamma_{p,n}})^2/2 \le \lambda_{\min}(R) \le \lambda_{\max}(R) \le 2(1 + \sqrt{\gamma_{p,n}})^2 \}.$$

Since $\mathbb{E}\,|X_{ij}|^{8+\delta} < \infty$ for $i \in [n]$, $j \in [p]$, it can be checked that $\mathbb{P}(\Omega_1^c) \le n^{-1-\delta/4}$ by Chebyshev's inequality and taking a union bound. To bound $\mathbb{P}(\Omega_2^c)$, by the argument from Section 9.12.5 of Bai and Silverstein (2010), it can be readily checked that $\mathbb{P}(\Omega_2^c | \Omega_1) < n^{-\ell}$ for any fixed $\ell > 0$. Then by taking ℓ large enough, $\mathbb{P}(\Omega_2^c) \le \mathbb{P}(\Omega_2^c | \Omega_1) + \mathbb{P}(\Omega_1^c) \le 2\,\mathbb{P}(\Omega_1^c)$.

Let v_0 be a small positive constant, and take $x_\ell = (1 - \sqrt{\gamma_{p,n}})^2/3$, $x_r = 3(1 + \sqrt{\gamma_{p,n}})^2$. Denoting the imaginary unit by i, define the rectangular region

$$\Upsilon = \{ z = x_\ell + \mathrm{i}v : v \in [-v_0, v_0] \} \cup \{ z = x_r + \mathrm{i}v : v \in [-v_0, v_0] \} \cup \{ z = x \pm \mathrm{i}v_0 : x \in [x_\ell, x_r] \}$$

in the complex plane.

For any z in Υ, let $m_n(z) := p^{-1} \operatorname{tr}(R - zI)^{-1}$, and $m_c(z)$ be the solution to the self-consistent equation $[m_c(z)]^{-1} = -z + (p/n)^{-1}[1 + m_c(z)]^{-1}$ whose imaginary part has the same sign as that of z. This equation has a unique solution (Bai and Silverstein, 2010).

By Cauchy's integral formula, we have

$$\left(n^{-1} \operatorname{tr} R^{-1} - \frac{\gamma_{p,n}}{1-\gamma_{p,n}} \right) I(\Omega_1 \cap \Omega_2) = \frac{\gamma_{p,n}}{2\pi \mathrm{i}} \oint_\Upsilon \frac{m_n(z) - m_c(z)}{z} I(\Omega_1 \cap \Omega_2) dz. \quad (24)$$

Defining $\Omega_3 = \{ |m_n(z) - m_c(z)| \le n^{-0.9999} \}$, we have $\mathbb{P}(\Omega_3^c | \Omega_1) < n^{-\ell}$ for any fixed $\ell > 0$. This is a consequence of the averaged local law from random matrix theory, see Theorem 3.5 of Yang (2020) for instance. Specifically, we apply their result for $d_N = \gamma_{p,n}$ and $\Sigma = I$. Note that for ρ_{2c} defined in their equation (2.10), $\operatorname{supp}(\rho_{2c}) = [(1 - \sqrt{\gamma_{p,n}})^2, (1 + \sqrt{\gamma_{p,n}})^2]$ thus with D_{out} from their equation (3.10) and some small $\omega > 0$, $\Upsilon \subset D_{out}$. Additionally, the bounded support condition in their equation (3.15) is satisfied on Ω_1. Then (3.21) in their Theorem 3.5 implies our desired bound.

By taking ℓ large enough, $\mathbb{P}(\Omega_3^c) \leq \mathbb{P}(\Omega_3^c|\Omega_1) + \mathbb{P}(\Omega_1^c) \leq 2\,\mathbb{P}(\Omega_1^c)$. Therefore, combining the above bounds for the probabilities of $\Omega_1, \Omega_2, \Omega_3$ and by (24), we have

$$\mathbb{P}\left(\left|n^{-1}\operatorname{tr} R^{-1} - \frac{\gamma_{p,n}}{1-\gamma_{p,n}}\right| \leq n^{-0.9999}\right) \geq \mathbb{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3)$$
$$\geq 1 - \mathbb{P}(\Omega_1^c) - \mathbb{P}(\Omega_2^c) - \mathbb{P}(\Omega_3^c) \geq 1 - 5\,\mathbb{P}(\Omega_1^c).$$

This finishes the argument.

A.4 Proof of Lemma A.2

We need to bound the smallest and largest eigenvalues of $T = Q \odot Q$. Now, for all $i, j \in [n]$, $T_{ij} = Q_{ij}^2 = (\delta_{ij} - x_i^\top R^{-1} x_j)^2$, where $R = X^\top X$ and δ_{ij} is the Kronecker delta which equals unity if $i = j$, and zero otherwise. We will use the following well-known rank-one perturbation formula for an invertible matrix T and a vector u of conformable size:

$$(uu^\top + T)^{-1} = T^{-1} - \frac{T^{-1}uu^\top T^{-1}}{1 + u^\top T^{-1} u}.$$

We will also use a "leave-one-out" argument which has roots in random matrix theory (see e.g., Bai and Silverstein, 2010; Paul and Aue, 2014; Yao et al., 2015). For any $i \in [n]$, letting $R_i = X^\top X - x_i x_i^\top = \sum_{j \neq i} x_j x_j^\top$, we have $R^{-1} = R_i^{-1} - \frac{R_i^{-1} x_i x_i^\top R_i^{-1}}{1 + x_i^\top R_i^{-1} x_i}$. For all $i, j \in [n]$, the quantity that is squared in the i, j-th entry of T is thus

$$x_i^\top R^{-1} x_j = x_i^\top R_i^{-1} x_j - \frac{x_i^\top R_i^{-1} x_i \cdot x_i^\top R_i^{-1} x_j}{1 + x_i^\top R_i^{-1} x_i} = \frac{x_i^\top R_i^{-1} x_j}{1 + x_i^\top R_i^{-1} x_i}.$$

Also, for all $i \in [n]$, we have $x_i^\top R^{-1} x_i = \frac{x_i^\top R_i^{-1} x_i}{1 + x_i^\top R_i^{-1} x_i}$, so that the diagonal terms are

$$T_{ii} = (1 - x_i^\top R^{-1} x_i)^2 = \frac{1}{(1 + x_i^\top R_i^{-1} x_i)^2}.$$

By the Gershgorin disk theorem (Horn and Johnson, 1990, Thm 6.1.1), we have $\lambda_{\max}(T) \leq \max_{i \in [n]} \left(T_{ii} + \sum_{j \neq i} |T_{ij}|\right)$. Thus,

$$\lambda_{\max}(T) \leq \max_{i \in [n]} \frac{1 + \sum_{j \neq i} (x_i^\top R_i^{-1} x_j)^2}{(1 + x_i^\top R_i^{-1} x_i)^2}.$$

Now, the sum in the numerator can be written as $x_i^\top R_i^{-1} (\sum_{j \neq i} x_j x_j^\top) R_i^{-1} x_i = x_i^\top R_i^{-1} x_i$. Thus, the upper bound simplifies to $\max_{i \in [n]} 1/(1 + x_i^\top R_i^{-1} x_i)$. Similarly, for the smallest eigenvalue, by the Gershgorin disk theorem (Horn and Johnson, 1990, Thm 6.1.1), we have $\lambda_{\min}(T) \geq \min_{i \in [n]} \left(T_{ii} - \sum_{j \neq i} |T_{ij}|\right)$. We can express for all $i \in [n]$

$$T_{ii} - \sum_{j \neq i} |T_{ij}| = \frac{1 - x_i^\top R_i^{-1} x_i}{(1 + x_i^\top R_i^{-1} x_i)^2}.$$

Hence $\lambda_{\min}(T) \geq \min_{i \in [n]} (1 - x_i^\top R_i^{-1} x_i)(1 + x_i^\top R_i^{-1} x_i)^2$, finishing the proof.

A.5 Proof of Lemma A.3

We will use the following Lemma quoted from Bai and Silverstein (2010).

Lemma A.4 (Trace Lemma, Lemma B.26 of Bai and Silverstein (2010))**.** *Let y be a p-dimensional random vector of independent elements with mean zero. Suppose that $\mathbb{E}\left[y_i^2\right] = 1$ and $\mathbb{E}\,|y_i|^\ell \le m_\ell$ for all $i \in [p]$, and let A_p be a fixed $p \times p$ matrix. Then for any $q \ge 2$,*

$$\mathbb{E}\left[|y^\top A_p y - \operatorname{tr} A_p|^q\right] \le C_q \left\{ \left(m_4 \operatorname{tr}[A_p A_p^\top] \right)^{q/2} + m_{2q} \operatorname{tr}[(A_p A_p^\top)^{q/2}] \right\},$$

for some constant C_q that only depends on q.

Proof. Under the conditions of Lemma A.3, the operator norms $\|A_p\|_2$ are bounded by a constant C, thus $\operatorname{tr}[(A_p A_p^\top)^{q/2}] \le pC^q$ and $\operatorname{tr}[A_p A_p^\top] \le pC^2$. Consider now a random vector x with the properties assumed in the present lemma. For $y = \sqrt{n}x/\sigma$ and $q = 2 + \eta/2$ with $\eta > 0$, using $\mathbb{E}\left[y_i^{2q}\right] \le C$ and the other conditions in Lemma A.3, Lemma A.4 yields

$$\frac{n^q}{\sigma^{2q}} \mathbb{E}\left[|x^\top A_p x - \frac{\sigma^2}{n} \operatorname{tr} A_p|^q\right] \le C \left\{ \left(pC^2\right)^{q/2} + pC^q \right\},$$

or equivalently $\mathbb{E}\left[|x^\top A_p x - \frac{\sigma^2}{n} \operatorname{tr} A_p|^{2+\eta/2}\right] \le C n^{-(1+\eta/4)}$. By Markov's inequality applied to the $2+\eta$-th moment of $\varepsilon_p = x^\top A_p x - \frac{\sigma^2}{n} \operatorname{tr} A_p$, we obtain as required that for $t > 0$, $\mathbb{P}(|\varepsilon_p|^{2+\eta/2} > t) \le C t^{-1} n^{-(1+\eta/4)}$. □

A.6 Proof of Proposition 2.3

We need to evaluate $\mathbb{E}\,\widehat{V}^{\odot 2} = \mathbb{E}\,\widehat{V} \odot \widehat{V} \in \mathbb{R}^p$. Note that this vector is the diagonal of $\mathbb{E}\,\widehat{V}\widehat{V}^\top$, which is equal to

$$\mathbb{E}\,\widehat{V}\widehat{V}^\top = \mathbb{E}\,A(\widehat{\varepsilon} \odot \widehat{\varepsilon})(\widehat{\varepsilon} \odot \widehat{\varepsilon})^\top A^\top = \mathbb{E}\,A\left[(\widehat{\varepsilon}\widehat{\varepsilon}^\top) \odot (\widehat{\varepsilon}\widehat{\varepsilon}^\top)\right] A^\top = A\,\mathbb{E}\left[(\widehat{\varepsilon}\widehat{\varepsilon}^\top) \odot (\widehat{\varepsilon}\widehat{\varepsilon}^\top)\right] A^\top.$$

Note that $\widehat{\varepsilon}\widehat{\varepsilon}^\top = Q\varepsilon\varepsilon^\top Q$, since the residuals $\widehat{\varepsilon} = Q\varepsilon$. Using this expression and recognizing that ε has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, for any $i, j \in [n]$, the (i,j)-th element of $\mathbb{E}(\widehat{\varepsilon}\widehat{\varepsilon}^\top)^{\odot 2}$ is

$$\begin{aligned}
&\mathbb{E}\left(\sum_{1\le l,k\le n} Q_{il}\varepsilon_l\varepsilon_k Q_{kj}\right)^2 \\
&= \sum_{l\neq k} \mathbb{E}\left(Q_{il}^2 Q_{jk}^2 \varepsilon_l^2\varepsilon_k^2 + Q_{il}Q_{jk}Q_{ik}Q_{jl}\varepsilon_k^2\varepsilon_l^2 + Q_{il}Q_{jl}Q_{ik}Q_{jk}\varepsilon_l^2\varepsilon_k^2\right) + \sum_{l=1}^n \mathbb{E}\,Q_{il}^2 Q_{jl}^2 \varepsilon_l^4 \\
&= \sum_{l\neq k} \left(Q_{il}^2 Q_{jk}^2 \sigma^4 + Q_{il}Q_{jk}Q_{ik}Q_{jl}\sigma^4 + Q_{il}Q_{jl}Q_{ik}Q_{jk}\sigma^4\right) + \sum_{l=1}^n Q_{il}^2 Q_{jl}^2 3\sigma^4 \\
&= \sigma^4 \sum_{l\neq k} \left(Q_{il}^2 Q_{jk}^2 + 2Q_{il}Q_{jk}Q_{ik}Q_{jl}\right) + 3\sigma^4 \sum_{l=1}^n Q_{il}^2 Q_{jl}^2 \\
&= \sigma^4 \sum_{1\le l,k\le n} \left(Q_{il}^2 Q_{jk}^2 + 2Q_{il}Q_{jk}Q_{ik}Q_{jl}\right) = \sigma^4 \sum_{1\le l,k\le n} Q_{il}^2 Q_{jk}^2 + 2\sigma^4 \left(\sum_{l=1}^n Q_{il}Q_{jl}\right)^2.
\end{aligned}$$

To proceed, we recognize that $\sum_{1\le l,k\le n} Q_{il}^2 Q_{jk}^2$ is the (i,j)-th element of

$$[(Q\odot Q)1_n]\,[(Q\odot Q)1_n]^\top = (Q\odot Q)1_n 1_n^\top (Q\odot Q),$$

and $\left(\sum_{l=1}^n Q_{il}Q_{jl}\right)^2$ is the (i,j)-th element of $Q^2\odot Q^2 = Q\odot Q$.

Summarizing the calculation above, we obtain

$$\mathbb{E}(\widehat{\varepsilon}\widehat{\varepsilon}^\top)^{\odot 2} = \sigma^4 (Q\odot Q)1_n 1_n^\top (Q\odot Q) + 2\sigma^4 Q\odot Q,$$

from which it follows that

$$\begin{aligned}
\mathbb{E}\,\widehat{V}\odot\widehat{V} &= \operatorname{diag}\left[A\left(\sigma^4(Q\odot Q)1_n1_n^\top(Q\odot Q) + 2\sigma^4 Q\odot Q\right)A^\top\right] \\
&= \sigma^4 \operatorname{diag}\left[A(Q\odot Q)1_n1_n^\top(Q\odot Q)A^\top\right] + 2\sigma^4\operatorname{diag}\left[A(Q\odot Q)A^\top\right] \\
&= \sigma^4 \operatorname{diag}\left[(S\odot S)1_n1_n^\top(S\odot S)^\top\right] + 2\sigma^4\operatorname{diag}\left[(S\odot S)(Q\odot Q)^{-1}(S\odot S)^\top\right].
\end{aligned}$$

Note that $V = \sigma^2 \operatorname{diag}\left[(X^\top X)^{-1}\right]$ due to the assumption of homoskedasticity. Recalling (10), we find (11), finishing the proof.

A.7 Calculation for the case when $p = 1$

We compute each part of the unbiased estimator in turn. We start by noticing that $S = (X^\top X)^{-1}X^\top = X^\top$ is a $1\times n$ vector. We continue by calculating $Q\odot Q$, where $Q = I - X(X^\top X)^{-1}X^\top = I - XX^\top$. Thus,

$$Q_{ij}^2 = \begin{cases} X_i^2 X_j^2, & i\neq j \\ (1-X_i^2)^2, & \text{else.}\end{cases}$$

Denoting $u = X\odot X$, and $D = I - 2\operatorname{diag}(X\odot X)$, we can write $Q\odot Q = D + uu^\top$. Now, the estimator takes the form $\widehat{V} = (S\odot S)(Q\odot Q)^{-1}(\widehat{\varepsilon}\odot\widehat{\varepsilon})$. Hence, we need to calculate $(S\odot S)(Q\odot Q)^{-1} = (X\odot X)(D+uu^\top)^{-1}$. We use the rank-one perturbation formula $u^\top(D+uu^\top)^{-1} = \frac{u^\top D^{-1}}{u^\top D^{-1}u+1}$. In our case,

$$u^\top D^{-1}u = \sum_{j=1}^n \frac{u_j^2}{D_j} = \sum_{j=1}^n \frac{X_j^4}{1-2X_j^2},$$

and $u^\top D^{-1}$ has entries $X_j^2/(1-2X_j^2)$ for $j\in[n]$. This leads to the desired final answer:

$$\widehat{V} = u^\top(D+uu^\top)^{-1}\widehat{\varepsilon}\odot\widehat{\varepsilon} = \frac{\sum_{j=1}^n \frac{X_j^2}{1-2X_j^2}\widehat{\varepsilon}_j^2}{1+\sum_{j=1}^n \frac{X_j^4}{1-2X_j^2}}.$$

Next, since $X^\top X = 1$, we have for E from (10) that $E = 1$. Finally, for d from (11), since $S = X^\top$, $u = X\odot X$, and $Q\odot Q = D + uu^\top$, so that $u^\top 1_n = 1$, we find

$$d = \frac{1}{u^\top(D+uu^\top)^{-1}u} = 1 + \frac{1}{u^\top D^{-1}u} = 1 + \frac{1}{\sum_{j=1}^n \frac{X_j^4}{1-2X_j^2}},$$

as desired.

A.8 Proof of Proposition 2.4

To compute the bias of White's estimator defined in (3), we proceed as follows. First we need to compute its expectation,

$$\mathbb{E}\,\widehat{C}_{\mathrm{W}} = (X^\top X)^{-1}[X^\top\,\mathbb{E}\,\mathrm{diag}(\widehat{\varepsilon}\odot\widehat{\varepsilon})X](X^\top X)^{-1}.$$

As we saw, $\mathbb{E}(\widehat{\varepsilon}\odot\widehat{\varepsilon}) = \mathrm{diag}\,\mathrm{Cov}(\widehat{\varepsilon}) = \mathrm{diag}\,Q\Sigma Q = (Q\odot Q)\vec{\Sigma}$. Thus,

$$\mathrm{diag}\,\mathbb{E}\,\widehat{C}_{\mathrm{W}} = \mathrm{diag}[S\,\mathrm{diag}[(Q\odot Q)\vec{\Sigma}]S^\top] = (S\odot S)(Q\odot Q)\vec{\Sigma}.$$

Again, as we saw, $V = \mathrm{diag}\,\mathrm{Cov}(\hat{\beta}) = (S\odot S)\vec{\Sigma}$. Therefore, the bias of White's estimator is as in (14).

To compute the bias of MacKinnon-White's estimator, we proceed similarly, starting with its expectation:

$$\mathbb{E}\,\widehat{C}_{\mathrm{MW}} = (X^\top X)^{-1}[X^\top\,\mathbb{E}\,\mathrm{diag}(Q)^{-1}\,\mathrm{diag}(\widehat{\varepsilon}\odot\widehat{\varepsilon})X](X^\top X)^{-1}.$$

In this equation, the expression $\mathrm{diag}(Q)$ is interpreted as the diagonal matrix whose entries are those on the diagonal of Q. Thus,

$$\mathrm{diag}\,\mathbb{E}\,\widehat{C}_{\mathrm{MW}} = \mathrm{diag}[S\,\mathrm{diag}(Q)^{-1}\,\mathrm{diag}[(Q\odot Q)\vec{\Sigma}]S^\top] = (S\odot S)(Q\odot Q)\,\mathrm{diag}(Q)^{-1}\vec{\Sigma}.$$

Thus the bias is as in (15), finishing the proof.

A.9 Proof of Theorem 3

We aim to bound $\|\widehat{V} - V\|$, where $\|\cdot\|$ denotes usual Euclidean vector norm. Recalling that $V = (S\odot S)\vec{\Sigma}$ and $\widehat{V} = (S\odot S)(Q\odot Q)^{-1}(\widehat{\varepsilon}\odot\widehat{\varepsilon})$, where $S = (X^\top X)^{-1}X^\top$, we have

$$\|\widehat{V} - V\| \le \|S\odot S\|_{\mathrm{op}}\|(Q\odot Q)^{-1}\|_{\mathrm{op}}\|(\widehat{\varepsilon}\odot\widehat{\varepsilon}) - (Q\odot Q)\vec{\Sigma}\|. \tag{25}$$

We will find upper bounds for each term in the above product.

1. To bound $\|S\odot S\|_{\mathrm{op}}$, Schur's inequality (e.g., Horn and Johnson, 1994, Thm. 5.5.1), states that $\|S\odot S\|_{\mathrm{op}} \le \|S\|_{\mathrm{op}}^2$. Moreover, as $\|S\|_{\mathrm{op}} = 1/\sigma_{\min}(X)$, it follows by (9.7.9) in Bai and Silverstein (2010) that for any constant $c > 0$ and $\ell > 0$, $\sigma_{\min}(X) \ge \sigma_{\min}(\Gamma^{1/2})(n^{1/2} - p^{1/2} - c)$ holds with probability $1 - o(n^{-\ell})$. Thus, denoting

$$\mathcal{E}_{1,n} := \left\{ n\|S\odot S\| \le c\frac{1}{\sigma_{\min}(\Gamma)(1-\gamma_{p,n}^{1/2})^2} \right\},$$

   we have $\mathbb{P}(\mathcal{E}_{1,n}) \ge 1 - o(n^{-\ell})$ for any constant $c > 1$ and $\ell > 0$.

2. To bound $\|(Q\odot Q)^{-1}\|_{\mathrm{op}}$, denoting

$$\mathcal{E}_{2,n} := \left\{ \|(Q\odot Q)^{-1}\|_{\mathrm{op}} \le c\frac{1}{(1-\gamma_{p,n})(1-2\gamma_{p,n})} \right\},$$

   by Theorem 2, we have $\mathbb{P}(\mathcal{E}_{2,n}) \ge 1 - O(n^{-1-\delta/4})$ for any $c > 1$.

3. To bound $\alpha := \|(\widehat{\varepsilon} \odot \widehat{\varepsilon}) - (Q \odot Q)\vec{\Sigma}\|$, we can express $\alpha^2 = \sum_{i=1}^n \alpha_i^2$, where for all $i \in [n]$, $\alpha_i^2 = (\widehat{\varepsilon}_i^2 - (q_i \odot q_i)^\top \vec{\Sigma})^2$. From the earlier unbiasedness argument, $\mathbb{E}\,\widehat{\varepsilon}_i^2 = (q_i \odot q_i)^\top \vec{\Sigma}$, and thus $\mathbb{E}\,\alpha_i^2 = \operatorname{Var}\widehat{\varepsilon}_i^2$. A simple calculation shows that, with $\Gamma_k = \mathbb{E}\,\varepsilon_k^4$ for all $k \in [n]$, we have

$$\operatorname{Var}\widehat{\varepsilon}_i^2 = \sum_{k=1}^n q_{ik}^4(\Gamma_k - 3\Sigma_k^2) + 2[(q_i \odot q_i)^\top \vec{\Sigma}]^2.$$

Now for all $k \in [n]$, the excess kurtosis can be bounded as $\Gamma_k - 3\Sigma_k^2 \le (C-3)\Sigma_k^2$. Therefore, we can bound by Markov's inequality:

$$\mathbb{P}(\alpha \ge t) \le \frac{\sum_{i=1}^n \mathbb{E}\,\alpha_i^2}{t^2} = \frac{(C-1)\sum_{i=1}^n [(q_i \odot q_i)^\top \vec{\Sigma}]^2}{t^2} \le \frac{(C-1)\cdot \left\|(Q \odot Q)\vec{\Sigma}\right\|^2}{t^2}.$$

According to Theorem 2, $\|Q \odot Q\|_{\mathrm{op}} \le c(1-\gamma_{p,n})$ holds with probability $1 - C' n^{-1-\delta/4}$ for some positive constant C' and any constant $c > 1$. Hence

$$\mathbb{P}(\alpha \ge t) \le \frac{2c(1-\gamma_{p,n})^2 \|\vec{\Sigma}\|^2}{t^2} + C' n^{-1-\delta/4}.$$

In conclusion, we have for sufficiently large n and p with $\gamma_{p,n} < 1/2$ that

$$\begin{aligned}
&\mathbb{P}\left(\frac{\|\widehat{V} - V\|}{\|\vec{\Sigma}\|} > \frac{t}{n}\right) \le \mathbb{P}\left(\alpha n \|S \odot S\|_{\mathrm{op}} \|(Q \odot Q)^{-1}\|_{\mathrm{op}} > t\|\vec{\Sigma}\|\right) \\
&\le \mathbb{P}\left(\alpha n \|S \odot S\|_{\mathrm{op}} \|(Q \odot Q)^{-1}\|_{\mathrm{op}} > t\|\vec{\Sigma}\| \,\middle|\, \mathcal{E}_{1,n}, \mathcal{E}_{2,n}\right) + \mathbb{P}(\mathcal{E}_{1,n}^c) + \mathbb{P}(\mathcal{E}_{2,n}^c) \\
&\le \frac{2c}{t^2} \frac{1}{\left[\sigma_{\min}(\Gamma)(1-\gamma_{p,n}^{1/2})^2(1-2\gamma_{p,n})\right]^2} + C' n^{-1-\delta/4}.
\end{aligned}$$

This proves the required result.

A.10 A lemma on the joint distribution of quadratic and linear forms

We state a lemma that establishes the joint distribution of a random quadratic form $\mathcal{Z}_n^\top A_n \mathcal{Z}_n$ and a linear form $B_n^\top \mathcal{Z}_n$ where $\mathcal{Z}_n$ has independent entries that have zero mean and unit variance. The marginal distribution of the quadratic form is used in the proof of Theorem 4 and the joint distribution plays a key role in the proof of Theorem 6.

Lemma A.5. *Let $\mathcal{Z}_n = (Z_1, \cdots, Z_n)^\top \in \mathbb{R}^n$ be a sequence of random vectors with independent entries, satisfying $\mathbb{E}\,Z_i = 0$ and $\mathbb{E}\,Z_i^2 = 1$. Let $A_n = \{(a_{ij}(n))\}_n$ be a sequence of $n \times n$ symmetric matrices, and $B_n = \{(b_i(n))\}_n$ be a sequence of $n \times 1$ vectors, both independent of $\mathcal{Z}_n$. Define*

$$\Sigma_n = \begin{pmatrix} \sum_{i=1}^n (\mathbb{E}\,Z_i^4 - 3)a_{ii}^2 + 2\,\mathrm{tr}\,A_n^2, & \sum_{i=1}^n \mathbb{E}\,Z_i^3 a_{ii} b_i \\ \sum_{i=1}^n \mathbb{E}\,Z_i^3 a_{ii} b_i & \|B_n\|^2 \end{pmatrix}. \tag{26}$$

Assume the following conditions:

1. $\|A_n\|^2 / \operatorname{tr} A_n^2 = o_P(1)$ *and* $\max_{k\in[n]} |b_k(n)|/\|B_n\| = o_P(1)$.
2. *(a)* $\max_{i\in[n]} \mathbb{E}\, Z_i^4 < \infty$*; or*
   (b) $a_{ii} = 0$ *for* $i \in [n]$, $\max_{i\in[n]} \mathbb{E}\, Z_i^{2+\delta} < \infty$ *for some small constant* $\delta > 0$*, and* $\max_{i\in[n]} \mathbb{E}\, Z_i^4 \cdot \|A_n\|^2 / \operatorname{tr} A_n^2 = o_P(1)$.
3. $((\Sigma_n)_{12})^2/[(\Sigma_n)_{11}(\Sigma_n)_{22}] < 1 - c'$ *and* $(\Sigma_n)_{11} > c' \operatorname{tr} A_n^2$ *for some small constant* $c' > 0$.

Then we have for any fixed nonzero $c = (c_1, c_2)^\top \in \mathbb{R}^2$*,*

$$(c^\top \Sigma_n c)^{-1/2} c^\top \begin{pmatrix} \mathcal{Z}_n^\top A_n \mathcal{Z}_n - \operatorname{tr} A_n \\ B_n^\top \mathcal{Z}_n \end{pmatrix} \Rightarrow \mathcal{N}(0,1).$$

The condition $((\Sigma_n)_{12})^2/[(\Sigma_n)_{11}(\Sigma_n)_{22}] < 1 - c'$ *is not required for the marginal distribution of* $\mathcal{Z}_n^\top A_n \mathcal{Z}_n$*.*

This lemma generalizes several results in the literature. It extends Lemma B.1 in Kline et al. (2020) to allow for general A rather than restricting to the diagonal-free case that suffices for their approach. Moreover, we do not require the existence of fourth moments of Z_i when A_n is diagonal-free. Compared with Theorem 1 in Kelejian and Prucha (2001), we do not require $\|A_n\|_1 < \infty$. The marginal distribution of the quadratic form extends Theorem 5.2 in Capitaine et al. (2009), though our proof strategy is inspired by their approach.

Another feature of our result is its scale invariance property; the conclusion remains valid when A_n and B_n are replaced with $s_n A_n, t_n B_n$ for any sequences $(s_n)_{n\geq 1}$ and $(t_n)_{n\geq 1}$. This eliminates the need to consider proper scaling, distinguishing this lemma from other related results in the literature.

The conditions imposed are sufficient and nearly necessary to guarantee the asymptotic normality. When $\operatorname{tr} A_n^2$ and $\|B_n\|^2$ are of the same order, Condition 1 is necessary for the asymptotic normality. Condition 2(b) implies that when the diagonal entries of A are zero, the fourth-moment assumption on Z_i can be relaxed to some extent. Condition 3 rules out singular cases that may arise in special settings. For example, if $\mathbb{P}(Z_i = \pm 1) = 1/2$, then $(\Sigma_n)_{11} = 2\sum_{i\neq j} a_{ij}^2$, which is zero when all off-diagonal entries of A_n vanish.

Proof. We consider the case when the sequences A_n and B_n are nonrandom. The random case follows essentially the same argument by replacing the deterministic bounds for A_n- and B_n-related quantities with bounds in probability. We drop the subscript n from A_n, B_n and $\mathcal{Z}_n$ for simplicity.

Let both c_1 and c_2 be nonzero. The marginal cases are easier to handle separately.

We begin by assuming Conditions 1, 3 and 2(a), and we specify the modifications required under Condition 2(b) at the end. Let

$$\mathcal{M}_k := c_1(Z_k^2 - 1)a_{kk} + 2c_1 Z_k \sum_{j<k} Z_j a_{jk} + c_2 b_k Z_k.$$

Then we can write $c_1(\mathcal{Z}^\top A \mathcal{Z} - \operatorname{tr} A) + c_2 B^\top \mathcal{Z} = \sum_{k=1}^n \mathcal{M}_k$. Let $\mathcal{F}_k$ be the σ-field generated by $\{Z_1, \cdots, Z_k\}$, and $\mathbb{E}_k$ the conditional expectation with respect to $\mathcal{F}_k$. It is readily verified that $\{\mathcal{M}_k\}$ is martingale difference sequence, as $\mathbb{E}\, \mathcal{M}_k = 0$ and $\mathbb{E}_{k-1}\, \mathcal{M}_k = 0$. We apply the

martingale central limit theorem (e.g., see Theorem 35.12 of Billingsley (1995)). It suffices to show that

$$\sum_{k=1}^{n} \mathbb{E}\left(\left(\frac{\mathcal{M}_k}{\sqrt{c^\top \Sigma_n c}}\right)^2 I_{\left|\frac{\mathcal{M}_k}{\sqrt{c^\top \Sigma_n c}}\right|>\varepsilon}\right) \to 0 \tag{27}$$

and

$$\sum_{k=1}^{n} \mathbb{E}_{k-1}\left(\frac{\mathcal{M}_k}{\sqrt{c^\top \Sigma_n c}}\right)^2 \to_P 1. \tag{28}$$

Let $\mathcal{M}_k/\sqrt{c^\top \Sigma_n c} = c_1\mathcal{M}_{1k} + 2c_1\mathcal{M}_{2k} + c_2\mathcal{M}_{3k}$ where $\mathcal{M}_{1k} = (Z_k^2-1)a_{kk}/\sqrt{c^\top \Sigma_n c}$, $\mathcal{M}_{2k} = Z_k \sum_{j<k} Z_j a_{jk}/\sqrt{c^\top \Sigma_n c}$, and $\mathcal{M}_{3k} = b_k Z_k/\sqrt{c^\top \Sigma_n c}$. To show (27), it suffices to verify the bound for $\mathcal{M}_{1k}$, $\mathcal{M}_{2k}$ and $\mathcal{M}_{3k}$ individually.

The conditions on Σ_n imply that for any fixed nonzero vector c,

$$\frac{\operatorname{tr} A^2}{c^\top \Sigma_n c} \asymp \frac{\operatorname{tr} A^2}{(\Sigma_n)_{11}} = O(1). \tag{29}$$

For

$$\sum_{k=1}^{n} \mathbb{E}\left(\mathcal{M}_{1k}^2 I_{|\mathcal{M}_{1k}|>\varepsilon}\right) = \sum_{k=1}^{n} \mathbb{E}\left(\frac{(Z_k^2-1)^2}{c^\top \Sigma_n c} a_{kk}^2 I_{\{|Z_k|^2-1>\varepsilon\sqrt{c^\top \Sigma_n c}/a_{kk}\}}\right),$$

since $\max_k |a_{kk}|/\sqrt{c^\top \Sigma_n c} \le \|A\|/c^\top \Sigma_n c = o(1)$, we have

$$f_{n,k} := \mathbb{E}(Z_k^2-1)^2 I_{\{|Z_k|^2-1>\varepsilon\sqrt{c^\top \Sigma_n c}/a_{kk}\}} \to 0$$

as $n \to \infty$ for any k, and the sequence $\{f_{n,k}\}_k$ is uniformly bounded. Using $\sum_{k=1}^n a_{kk}^2/c^\top \Sigma_n c = O(1)$ and the dominated convergence theorem, we conclude

$$\sum_{k=1}^{n} \mathbb{E}\left(\mathcal{M}_{1k}^2 I_{|\mathcal{M}_{1k}|>\varepsilon}\right) \to 0. \tag{30}$$

By the Rosenthal inequality, we have

$$\mathbb{E}\,|\sum_{j<k} Z_j a_{jk}|^{2+\delta} \le C\left(\sum_{j<k} \mathbb{E}\, Z_j^{2+\delta} a_{jk}^{2+\delta} + (\sum_{j<k} a_{jk}^2)^{1+\delta/2}\right).$$

Using

$$\sum_{k=1}^{n}\sum_{j<k} \mathbb{E}\, Z_j^{2+\delta} a_{jk}^{2+\delta} \le \max_{j\in[n]} \mathbb{E}\, Z_j^{2+\delta} \cdot \|A\|^\delta \|A\|_{Fr}^2,$$

and

$$\sum_{k=1}^{n} (\sum_{j<k} a_{jk}^2)^{1+\delta/2} \le (\max_k \sum_{j=1}^{n} a_{jk}^2)^{\delta/2} \sum_{j<k} a_{jk}^2 \le \|A\|^\delta \|A\|_{Fr}^2,$$

we obtain

$$\mathbb{E} \sum_{k=1}^{n} \frac{1}{(c^\top \Sigma_n c)^{1+\delta/2}} |\sum_{j<k} Z_j a_{jk}|^{2+\delta} = o(1).$$

Thus we find

$$\sum_{k=1}^{n} \mathbb{E}\left(\mathcal{M}_{2k}^2 I_{|\mathcal{M}_{2k}|>\varepsilon}\right) \le \frac{1}{\varepsilon^{1+\delta/2}(c^\top \Sigma_n c)^{1+\delta/2}} \mathbb{E} \sum_{k=1}^{n} |\sum_{j<k} Z_j a_{jk}|^{2+\delta} = o(1).$$

We also find

$$\sum_{k=1}^{n} \mathbb{E}\left(\mathcal{M}_{3k}^2 I_{|\mathcal{M}_{3k}|>\varepsilon}\right) = \sum_{k=1}^{n} \frac{b_k^2}{c^\top \Sigma_n c} \mathbb{E}\, Z_k^2 I_{\{Z_k > \varepsilon \sqrt{c^\top \Sigma_n c}/b_k\}} = o(1),$$

where the second step uses $\max |b_k|/\sqrt{c^\top \Sigma_n c} = o(1)$ and arguments analogous to the proof of (30). Therefore we conclude (27).

Next, we verify (28). Below we calculate $\sum_{k=1}^n \mathbb{E}_{k-1}(c_1\mathcal{M}_{1k}+c_2\mathcal{M}_{3k})^2$, $\sum_{k=1}^n \mathbb{E}_{k-1}(\mathcal{M}_{2k})^2$, and $\sum_{k=1}^n \mathbb{E}_{k-1}(c_1\mathcal{M}_{1k} + c_2\mathcal{M}_{3k})\mathcal{M}_{2k}$.

We find

$$\sum_{k=1}^{n} \mathbb{E}_{k-1}(c_1\mathcal{M}_{1k} + c_2\mathcal{M}_{3k})^2 = \frac{1}{c^\top \Sigma_n c} \sum_{k=1}^{n} \left(c_1^2(\mathbb{E}\, Z_k^4 - 1)a_{kk}^2 + c_2^2 b_k^2 + 2c_1c_2\, \mathbb{E}\, Z_k^3 a_{kk} b_k\right). \tag{31}$$

Let A_L be the strictly lower triangular part of A. We have

$$\sum_{k=1}^{n} \mathbb{E}_{k-1}\, \mathcal{M}_{2k}^2 = \frac{1}{c^\top \Sigma_n c} \sum_{k=1}^{n} \left(\sum_{j<k} Z_j a_{jk}\right)^2 = \frac{\mathcal{Z}^\top A_L^\top A_L \mathcal{Z}}{c^\top \Sigma_n c}.$$

For any symmetric matrix A, we have the bound

$$\|A_L\| \le \sqrt{6}\|A\|, \tag{32}$$

which will be proved at the end. This and $\|A\|^2 = o(c^\top \Sigma_n c)$ imply that $\|A_L\|^2 = o(c^\top \Sigma_n c)$. Hence we have

$$\begin{aligned} \mathbb{E}(\mathcal{Z}^\top A_L^\top A_L \mathcal{Z} - \operatorname{tr} A_L^\top A_L)^2 &\le \sum_{i=1}^{n} (\mathbb{E}\, Z_i^4 - 3)((A_L^\top A_L)_{ii})^2 + 2\operatorname{tr}(A_L^\top A_L)^2 \\ &\le C\|A_L\|^2 \operatorname{tr}(A_L^\top A_L) = \|A_L\|^2 \sum_{j<k} a_{jk}^2 = o((c^\top \Sigma_n c)^2). \end{aligned} \tag{33}$$

Therefore, we find

$$\sum_{k=1}^{n} \mathbb{E}_{k-1}\, \mathcal{M}_{2k}^2 - \frac{1}{2c^\top \Sigma_n c}\left(\operatorname{tr} A^2 - \sum_{k=1}^{n} a_{kk}^2\right) = o_P(1). \tag{34}$$

We next analyze the cross-terms. We find

$$\mathbb{E}_{k-1}(\mathcal{M}_{1k}\mathcal{M}_{2k}) = \frac{1}{c^\top \Sigma_n c} \sum_{k=1}^{n} (\mathbb{E}\, Z_k^3) a_{kk} \sum_{j<k} a_{jk} Z_j.$$

Since $\mathbb{E}(\mathcal{M}_{1k}\mathcal{M}_{2k}) = 0$, and

$$\begin{aligned}\mathbb{E}\left(\sum_{k=1}^{n} a_{kk}\sum_{j<k} a_{jk}Z_j\right)^2 &= \sum_{k_1,k_2=1}^{n} a_{k_1k_1}a_{k_2k_2}\sum_{j<k_1,j<k_2} a_{jk_1}a_{jk_2}\\ &= 1_n^\top \operatorname{diag}(A)A_L A_L^\top \operatorname{diag}(A)1_n\\ &= \left(\sum_{i=1}^{n} a_{ii}^2\right)\|A_L A_L\| = o((c^\top\Sigma_n c)^2),\end{aligned}$$

where the last step uses $\|A_L\|^2 = o(c^\top\Sigma_n c)$, we have

$$\mathbb{E}_{k-1}(\mathcal{M}_{1k}\mathcal{M}_{2k}) = o_P(1). \tag{35}$$

Similarly, using $\mathbb{E}(\mathcal{M}_{3k}\mathcal{M}_{2k}) = 0$ and

$$\mathbb{E}\left(\sum_{k=1}^{n} b_k\sum_{j<k} a_{jk}Z_j\right)^2 = B^\top A_L A_L^\top B = \|B\|^2\|A_L A_L\| = o((c^\top\Sigma_n c)^2),$$

we find

$$\mathbb{E}_{k-1}(\mathcal{M}_{3k}\mathcal{M}_{2k}) = o_P(1). \tag{36}$$

Combining (31), (34), (35) and (36), we establish (28). The proof is complete under Conditions 1, 3 and 2(a).

Under Condition 2(b), the term $\mathcal{M}_{1k}$ vanishes. The only modification required is for finding $\sum_{k=1}^{n}\mathbb{E}_{k-1}\mathcal{M}_{2k}^2$ by modifying (33). The condition $\max_{i\in[n]}\mathbb{E}\,Z_i^4\cdot\|A_n\|^2/\operatorname{tr}A_n^2 = o(1)$ implies that

$$\sum_{i=1}^{n}(\mathbb{E}\,Z_i^4 - 3)((A_L^\top A_L)_{ii})^2 = \sum_{i=1}^{n}(\mathbb{E}\,Z_i^4 - 3)(\sum_{j>i} a_{ji}^2)^2 = o((\operatorname{tr}A^2)^2),$$

hence the bound in (33) still holds. This completes the proof.

Verification of (32). Denote the diagonal matrix of A by D. For any vector $x\in\mathbb{R}^n$, we have

$$\begin{aligned}x^\top A^2 x &= x^\top(A_L + A_L^\top + D)(A_L + A_L^\top + D)x\\ &= x^\top A_L A_L^\top x + x^\top A_L^\top A_L x + x^\top(A_L^2 + (A_L^\top)^2)x\\ &\quad + 2x^\top A_L D x + 2x^\top A_L^\top D x + x^\top D^2 x.\end{aligned}$$

We have

$$x^\top(A_L^2 + (A_L^\top)^2)x = \frac{1}{2}x^\top\left((A_L + A_L^\top)(A_L + A_L^\top) + (A_L - A_L^\top)(A_L - A_L^\top)\right)x \geq 0.$$

Moreover, applications of the Cauchy-Schwarz inequality and the AM-GM inequality yield

$$|x^\top A_L D x| \leq \sqrt{x^\top A_L A_L^\top x}\sqrt{x^\top D^2 x} \leq \frac{x^\top A_L A_L^\top x}{4} + x^\top D^2 x,$$

and

$$|x^\top A_L^\top Dx| \le \sqrt{x^\top A_L^\top A_L x}\sqrt{x^\top D^2 x} \le \frac{x^\top A_L^\top A_L x}{2} + \frac{x^\top D^2 x}{2}.$$

Plugging these bounds into (A.10), we find

$$\begin{aligned} x^\top A^2 x &\ge x^\top A_L A_L^\top x + x^\top A_L^\top A_L x - 2|x^\top A_L^\top Dx| - 2|x^\top A_L Dx| + x^\top D^2 x \\ &\ge \frac{1}{2} x^\top A_L A_L^\top x + 2x^\top D^2 x. \end{aligned}$$

Therefore $\|A_L A_L^\top\| \le 2\|A\|^2 + 4\|D\|^2 \le 6\|A\|^2$. This implies (32).

□

A.11 Proof of Theorem 4

For the given Z, any matrices $M_1, M_2 \in \mathbb{R}^{n\times n}$ and any vector $v \in \mathbb{R}^n$, we find

$$\begin{aligned} v^\top \operatorname{diag}\{M_1[(M_2 Z)\odot(M_2 Z)]\} v &= \sum_{i=1}^n v_i^2 \sum_{j=1}^n M_{1,ij} \left(\sum_{k=1}^n M_{2,jk} Z_k\right)^2 \\ = \sum_{k_1,k_2=1}^n Z_{k_1} Z_{k_2} M_{2,jk_1} \left(\sum_{i,j=1}^n v_i^2 M_{1,ij}\right) M_{2,jk_2} &= Z^\top M_2^\top \operatorname{diag}[(v\odot v)^\top M_1] M_2 Z. \end{aligned}$$

Taking $v = S^\top w_p$, $M_1 = (Q\odot Q)^{-1}$, $M_2 = Q\Sigma^{1/2}$ above, we find that

$$w_p^\top S(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top w_p = Z^\top G(w_p) Z, \tag{37}$$

where $G(w_p)$ is defined in (16). An application of Lemma A.5 concludes the proof.

A.12 Proof of Proposition 4.1

Proof. We consider $\|G(w_p)\|_{\mathrm{Fr}}$ first. For any vector $v \in \mathbb{R}^n$, we find that

$$\begin{aligned} \|Q\operatorname{diag}(v)Q\|_{\mathrm{Fr}}^2 &= \sum_{j,k=1}^n \left(\sum_{\ell=1}^n Q_{j\ell} v_\ell Q_{\ell k}\right)^2 \\ = \sum_{j,k=1}^n \sum_{\ell_1,\ell_2=1}^n Q_{j\ell_1} v_{\ell_1} Q_{\ell_1 k} Q_{j\ell_2} v_{\ell_2} Q_{\ell_2 k} &= \sum_{\ell_1,\ell_2=1}^n v_{\ell_1}[(Q^2)_{\ell_1\ell_2}]^2 v_{\ell_2} = v^\top (Q\odot Q) v, \end{aligned} \tag{38}$$

where we use $Q^2 = Q$ in the last step. Substituting $v^\top = [(w_p^\top S)\odot(w_p^\top S)](Q\odot Q)^{-1}$ into the above equation, and by condition 1, we conclude that

$$\|G(w_p)\|_{\mathrm{Fr}} = \Omega(\lambda_{\min}(\Sigma)\|(w_p^\top S)\odot(w_p^\top S)\|).$$

The upper bound for $\|G(w_p)\|$ is obtained by

$$\begin{aligned} \|G(w_p)\| &\le \|Q\Sigma Q\|\|\operatorname{diag}(\left[(w_p^\top S)\odot(w_p^\top S)\right](Q\odot Q)^{-1})\| \\ &\le \lambda_{\max}(\Sigma)\max_{j\in[n]}\left|\left[(w_p^\top S)\odot(w_p^\top S)\right](Q\odot Q)^{-1} e_j\right|, \end{aligned}$$

where the second step uses $\|Q\| \le 1$. Therefore we conclude $\|G(w_p)\|/\|G(w_p)\|_{Fr} \to 0$ by condition 2.

In the remainder, we show that the first two conditions hold with probability tending to one under the random design X satisfying the conditions of Theorem 2. In matrix form, write $X = Z\Gamma^{1/2}$ with the i-th row being $x_i = \Gamma^{1/2} z_i$ for all $i \in [n]$, where z_i has independent entries with zero mean, unit variance and finite $(8+\delta)$-th moment. The bound for $T = Q \odot Q$ in condition 1 holds with probability tending to one, as a consequence of Theorem 2.

Next we verify condition 2. Let $u^\top = (w_p^\top S) \odot (w_p^\top S)$. The verification consists of two steps. In the first step, we show that $\max_{k\in[n]} \kappa(\Sigma)|u_k|/\|u\| = o_P(1)$ if $\kappa(\Sigma) = o(n^{1/4})$. Recall that for $j \in [n]$, $S_{\cdot j}$ is the j-th column of S. Then

$$\begin{aligned} \|u\| = \left(\sum_{j=1}^n (w_p^\top S_{\cdot j})^4 \right)^{1/2} &\ge n^{-1/2} \sum_{j=1}^n (w_p^\top S_{\cdot j})^2 \\ &= n^{-1/2} w_p^\top S S^\top w_p = n^{-1/2} w_p^\top (X^\top X)^{-1} w_p. \end{aligned} \tag{39}$$

For any sequence of deterministic vectors $w_p \in \mathbb{R}^p$ of bounded norm, we have

$$\mathbb{E}\,|w_p^\top (Z^\top Z)^{-1} z_k|^{8+\delta} \le \mathbb{E}\left| w_p^\top \big(\sum_{i\neq k} z_i z_i^\top \big)^{-1} z_k \right|^{8+\delta} \le C\,\mathbb{E}\left\| w_p^\top \left(\sum_{i\neq k} z_i z_i^\top \right)^{-1} \right\|^{8+\delta},$$

which further equals $O(n^{-8-\delta})$, where in the first step we use the Sherman–Morrison formula

$$w_p^\top (Z^\top Z)^{-1} z_k = w_p^\top \left(\sum_{i\neq k} z_i z_i^\top \right)^{-1} z_k \Big/ \left[1 + z_k^\top \left(\sum_{i\neq k} z_i z_i^\top \right)^{-1} z_k \right] \le w_p^\top \left(\sum_{i\neq k} z_i z_i^\top \right)^{-1} z_k.$$

and in the second step we use the moment bound $\mathbb{E}\,|w_p^\top z_k|^q \le C\|w_p\|^q$ for any $0 \le q \le 8+\delta$. This moment bound can be checked by applying Lemma A.4 with $A_p = w_p w_p^\top$ and using $\mathrm{tr}(w_p w_p^\top) = \|w_p\|^2$.

Therefore,

$$\begin{aligned} &\mathbb{P}\left(\max_{k\in[n]} \left| w_p^\top S_{\cdot k} \right| > n^{-7/8} \lambda_{\min}^{-1/2}(\Gamma) \right) = \mathbb{P}\left(\max_{k\in[n]} \left| w_p^\top \Gamma^{-1/2} (Z^\top Z)^{-1} z_k \right| > n^{-7/8} \lambda_{\min}^{-1/2}(\Gamma) \right) \\ &\le n\mathbb{P}\left(\left| w_p^\top \Gamma^{-1/2} (Z^\top Z)^{-1} z_k \right| > n^{-7/8} \lambda_{\min}^{-1/2}(\Gamma) \right) \\ &\le n^{1+\frac{7}{8}(8+\delta)} \lambda_{\min}^{\frac{1}{2}(8+\delta)}(\Gamma)\, \mathbb{E}\,|w_p^\top \Gamma^{-1/2} (Z^\top Z)^{-1} z_k|^{8+\delta} = O(n^{-\delta/8}). \end{aligned}$$

We further find $\lambda_{\min}^{-1}(\Gamma) = O_P(n[\lambda_{\min}(X^\top X)]^{-1})$, where we use $\lambda_{\min}(X^\top X) \le \lambda_{\max}(\Gamma) \cdot \lambda_{\min}(Z^\top Z) = O_P(n\lambda_{\max}(\Gamma))$ and that $\kappa(\Gamma)$ is bounded. It follows that

$$\max_{k\in[n]} |w_p^\top S_{\cdot k}| = O_P(n^{-3/8}[\lambda_{\min}(X^\top X)]^{-1/2}]). \tag{40}$$

By this bound, (39), and $\kappa(\Sigma) = o(n^{1/4})$, we conclude that $\max_{k\in[n]} \kappa(\Sigma)|u_k|/\|u\| = o_P(1)$.

In the second step, we show that $\max_{k\in[n]} \kappa(\Sigma)|u^\top T^{-1} e_k|/\|u\| = o_P(1)$. We prove this by contradiction. If this does not hold, then it will imply a bound regarding the maximum entry of u that contradicts the conclusion established in the first step.

Let $v = T^{-1}u$ and suppose that there exist positive constants ε_0, δ_0, such that for any n, there exists $n_0 \geq n$, and $i_0(n_0)$ such that with probability greater than δ_0, $\kappa(\Sigma)v_{i_0}/\|v\| > \varepsilon_0$. Define $\mathcal{E}_1(c) = \{T_{i_0 i_0} > c\}$, and $\mathcal{E}_2 = \{\max_{j \neq i_0} |T_{i_0 j}| \leq n^{-3/4}\}$. We will show later that there exists a positive constant c_0 such that $\mathcal{E}_1(c_0)$ and $\mathcal{E}_2$ hold with probability tending to one. Thus on the intersection of the events $\{v_{i_0}/\|v\| > \varepsilon_0\}$, $\mathcal{E}_1(c_0)$ and $\mathcal{E}_2$, which holds with probability greater than $\delta_0/2$ for sufficiently large n_0, we have

$$\kappa(\Sigma)\frac{e_{i_0}^\top T v}{\|v\|} \geq \frac{T_{i_0 i_0}\kappa(\Sigma)v_{i_0} - \kappa(\Sigma)\sum_{j \neq i_0}|T_{i_0 j}v_j|}{\|v\|} \geq c_0\varepsilon_0 - \kappa(\Sigma)(n_0-1)^{-1/4},$$

where we use $\sum_{j \neq i_0}|T_{i_0 j}v_j| \leq (\sum_{j \neq i_0} T_{i_0 j}^2)^{1/2}\|v\|$ in the second step. For n_0 sufficiently large, the above lower bound is greater than $c_0\varepsilon_0/2$. Combining this with $\|u\| = \|Tv\| \asymp \|v\|$, we find $u_{i_0}/\|u\|$ is bounded away from zero with probability greater than $\delta_0/2$, contradicting the conclusion of step one that $\max_k |u_k|/\|u\| = o_P(1)$.

In what follows, we show that the above defined events $\mathcal{E}_i$ hold with probability tending to one. For the event $\mathcal{E}_1$, note the relation $T_{ii} = (1 + x_i^\top R_i^{-1}x_i)^{-1}$, which has been verified in the proof of Lemma A.2. Therefore, according to the proof of Theorem 2, specifically, recall the event Ω_1 and Ω_2 defined therein, there exists some positive c_0 such that $\mathcal{E}_1(c_0)$ holds with probability tending to one. For $\mathcal{E}_2$, applying the Sherman-Morrison formula we find that for $j \neq i$, $T_{ij} = (x_i^\top(X^\top X)^{-1}x_j)^2 = (z_i^\top(Z^\top Z)^{-1}z_j)^2 \leq (z_i^\top R_{ij}^{-1}z_j)^2$ where $R_{ij} = \sum_{k \notin \{i,j\}} z_k z_k^\top$. Using the moment bound $\mathbb{E}\,|w_p^\top z_i|^q \leq C\|w_p\|^q$ for any $0 \leq q \leq 8+\delta$, we find $\mathbb{E}\, n^{-8-\delta}|z_i^\top A_n z_j|^{8+\delta} = O(n^{-4-\delta/2})$ for any A_n with bounded spectral norm that is independent of z_i, z_j. We apply this with $A_n = n^{-1}R_{ij}^{-1}$ and by the fact that $n^{-1}R_{ij}^{-1}$ has a bounded spectral norm with probability greater than $1 - n^{-\ell}$ for any $\ell > 0$, it can be readily checked that $\mathbb{P}(|T_{i_0 j}| > n^{-3/4}) \leq n^{3+3\delta/8}\,\mathbb{E}\,T_{i_0 j}^{4+\delta/2} = O(n^{-1-\delta/8})$. By taking a union bound over $j \neq i_0$, we conclude that $\mathcal{E}_2$ holds with probability tending to one.

□

A.13 Consistency of the SNR estimator

The following result shows that the SNR estimator given in (5) is ratio-consistent.

Proposition A.6 (Ratio-consistency of SNR estimator)**.** *Assume the conditions of Theorem 4 on the noise ε, and condition 1 of Proposition 4.1 on the data matrix X. In addition, suppose $\|\beta\| = \Omega(1)$, $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$ for positive constants $c \leq C$, and $\lambda_{\min}(n^{-1}X^\top X) \asymp 1$. For $\widehat{\mathrm{SNR}}$ from* (5)*, and the signal-to-noise ratio $\mathrm{SNR} = n\|\beta\|^2/\operatorname{tr}\Sigma$, we have $\widehat{\mathrm{SNR}}/\mathrm{SNR} \to_P 1$.*

Proof. We already know that the numerator and the denominator of $\widehat{\mathrm{SNR}}$ are unbiased estimators for $\|\beta\|^2$ and $n^{-1}\operatorname{tr}\Sigma$, respectively. The conclusion follows by applying Slutsky's theorem if we can show

$$(\|\hat{\beta}\|^2 - 1_p^\top \widehat{V})/\|\beta\|^2 \to_P 1, \tag{41}$$

and

$$\frac{1_p^\top (Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon})}{\operatorname{tr}\Sigma} \to_P 1. \tag{42}$$

We show (41) first by verifying that $\mathrm{Var}\left[(\|\hat{\beta}\|^2 - 1_p^\top \widehat{V})/\|\beta\|^2\right] \to 0$. We have a decomposition for $\|\hat{\beta}\|^2$, given as $\|\hat{\beta}\|^2 = \|\beta + S\varepsilon\|^2 = \|\beta\|^2 + 2\beta^\top S\varepsilon + \varepsilon^\top S^\top S\varepsilon$. By the assumptions $\|\beta\| = \Omega(1)$, $\lambda_{\min}(X^\top X) \asymp n$, and $\|\Sigma\| \asymp 1$, we find

$$\begin{aligned}\mathrm{Var}(\|\hat{\beta}\|^2) &\le 8\mathrm{Var}[\beta^\top S\varepsilon] + 2\mathrm{Var}[\varepsilon^\top S^\top S\varepsilon] \\ &\le C\beta^\top (X^\top X)^{-1} X^\top \Sigma X (X^\top X)^{-1}\beta + C\|\Sigma S^\top S\|_{\mathrm{Fr}}^2 = O(\|\beta\|^2 n^{-1}).\end{aligned} \tag{43}$$

For $1_p^\top \widehat{V}$, using (37) with $w_p = e_i$ and summing over $i \in [n]$, we find $1_p^\top \widehat{V} = Z^\top \Sigma^{1/2} Q \,\mathrm{diag}[1_p^\top (S \odot S)(Q \odot Q)^{-1}] Q \Sigma^{1/2} Z$. Then

$$\mathrm{Var}(1_p^\top \widehat{V}) \asymp \|\Sigma^{1/2} Q \,\mathrm{diag}[1_p^\top (S \odot S)(Q \odot Q)^{-1}] Q \Sigma^{1/2}\|_{\mathrm{Fr}}^2 \asymp 1_p^\top (S \odot S)(Q \odot Q)^{-1}(S \odot S) 1_p,$$

where in the second step we use (38) with $v^\top = 1_p^\top (S \odot S)(Q \odot Q)^{-1}$. By $\|S \odot S\| \le \|S\|_{\mathrm{op}}^2 \asymp n^{-1}$ and $(Q \odot Q)^{-1} \asymp 1$, we have $\mathrm{Var}(1_p^\top \widehat{V}) = O(n^{-1})$. Therefore, combining this with (43) and $\|\beta\| \asymp 1$, we conclude (41).

Then we verify (42) by checking that the variance of the left term tends to zero. Writing $1_p^\top (Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon}) = Z^\top \Sigma^{1/2} Q \,\mathrm{diag}[1_p^\top (Q \odot Q)^{-1}] Q \Sigma^{1/2} Z$, we obtain

$$\mathrm{Var}\left(1_p^\top (Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon})\right) \asymp \|\Sigma^{1/2} Q \,\mathrm{diag}[1_p^\top (Q \odot Q)^{-1}] Q \Sigma^{1/2}\|_{\mathrm{Fr}}^2 \asymp 1_p^\top (Q \odot Q)^{-1} 1_p \asymp n,$$

where in the second step we use (38). Since $\mathrm{tr}\,\Sigma \asymp n$, we conclude (42).

□

A.14 Proof of Lemma 4.2 and Theorem 5

A.14.1 Proof of Lemma 4.2

We start with the first conclusion. Denote the ℓ_1-norm of a vector $a \in \mathbb{R}^n$ by $\|a\|_1 = \sum_{i=1}^n |a_i|$, and ℓ_∞-norm by $\|a\|_\infty = \max_{i\in[n]} |a_i|$. Let $u^\top = (w_p^\top S) \odot (w_p^\top S)$ and $v^\top = u^\top (Q \odot Q)^{-1}$. By (37), we have

$$\begin{aligned}\mathrm{Var}(w_p^\top S(\mathrm{diag}\,\widehat{\vec{\Sigma}})S^\top w_p) &= \mathrm{Var}(Z^\top G(w_p) Z) \\ &= \sum_{i=1}^n \mathbb{E}\, Z_i^4 (G(w_p)_{ii})^2 + 2\,\mathrm{tr}(G(w_p)^2) \\ &\le C\,\mathrm{tr}(G(w_p)^2) \le C\lambda_{\max}^2(\Sigma) u^\top v \le C\lambda_{\max}^2(\Sigma) \max \|v\|_\infty \|u\|_1,\end{aligned} \tag{44}$$

where C is some positive constant, the second step uses the variance formula of random quadratic forms, see Lemma A.5, and the second last step uses (38). We also have

$$w_p^\top S\Sigma S^\top w_p = \sum_{i=1}^n \Sigma_i (w_p^\top S e_i)^2 \ge \lambda_{\min}(\Sigma)\|u\|_1 \ge \lambda_{\min}(\Sigma)\|u\|^{1/2}\|u\|_1^{1/2},$$

where the last step uses $\|u\|_1 \ge \|u\|$, which holds since $u_i \ge 0$ for $i \in [n]$. By Chebyshev's inequality and Condition 2 of Proposition 4.1, we conclude

$$\frac{w_p^\top S(\mathrm{diag}\,\widehat{\vec{\Sigma}})S^\top w_p}{w_p^\top S\Sigma S^\top w_p} \to_P 1.$$

Then consider the second conclusion. Define the events

$$\tilde{\Omega}_1 = \left\{ \max_{i\in[p],j\in[n]} |e_i^\top S_{\cdot j}| \le n^{-1/4-\delta/4}\lambda_{\min}(X^\top X)^{-1/2} \right\}, \quad \tilde{\Omega}_2(c) = \{\lambda_{\min}(Q\odot Q) > c\},$$

and $\tilde{\Omega} = \tilde{\Omega}_1 \cap \tilde{\Omega}_2(c)$. The event $\tilde{\Omega}$ holds with probability $1-o(1)$ for some $c>0$. For $\tilde{\Omega}_2$ it is according to Theorem 2. For $\tilde{\Omega}_1$, it can be verified by slightly modifying the arguments leading to (40). It suffices to prove that for $i\in[n]$,

$$\mathbb{E}(|\widehat{V}_i - V_i|^4 I_{\tilde{\Omega}}) = o(n^{-1}[\lambda_{\max}(\Sigma)]^4[\lambda_{\min}(X^\top X)]^{-4}). \tag{45}$$

This together with $V_i = e_i^\top S\Sigma S^\top e_i \asymp \lambda_{\min}(\Sigma)[\lambda_{\min}(X^\top X)]^{-1}$ implies that for any $\varepsilon_0>0$,

$$\begin{aligned}\mathbb{P}\left(\max_{i\in[n]}|\widehat{V}_i/V_i - 1| > \varepsilon_0 \Big| \tilde{\Omega}\right) &\le \sum_{i=1}^n \mathbb{P}\left(|\widehat{V}_i/V_i-1|>\varepsilon_0\Big|\tilde{\Omega}\right)\\ &\le \sum_{i=1}^n \mathbb{E}(|\widehat{V}_i - V_i|^4|\tilde{\Omega})(\varepsilon_0 V_i)^{-4} = o(1),\end{aligned} \tag{46}$$

which concludes the proof.

By (37), Lemma A.4, and $\max_{j\in[p]}\mathbb{E}\,|Z_j|^4<\infty$, we have $\mathbb{E}(|\widehat{V}_i - V_i|^4|X) \le C[\|G(e_i)\|_{\text{Fr}}^4 + (\max_{j\in[n]}\mathbb{E}\,Z_j^8)\,\mathrm{tr}\,G(e_i)^4]$ for some positive constant C. Recalling the lower bound in (39), we have that

$$\begin{aligned}\|(e_i^\top S)\odot(e_i^\top S)\| &\le \max_{j\in[n]}|w_p^\top S_{\cdot j}|\left(\sum_{j=1}^n (e_i^\top S_{\cdot j})^2\right)^{1/2}\\ &\le \max_{j\in[n]}|e_i^\top S_{\cdot j}|[\lambda_{\min}(X^\top X)]^{-1/2} = o(n^{-1/4}[\lambda_{\min}(X^\top X)]^{-1}).\end{aligned}$$

Then using (38) we have $\|G(e_i)\|_{\text{Fr}} = o(n^{-1/4}\lambda_{\max}(\Sigma)[\lambda_{\min}(X^\top X)]^{-1})$. We also have $\mathrm{tr}\,G(e_i)^4 \le \|G(e_i)\|^2\|G(e_i)\|_{Fr}^2 = o(n^{-1}[\lambda_{\max}(\Sigma)]^4[\lambda_{\min}(X^\top X)]^{-4})$. Therefore we conclude (45). This completes the proof for the design as in Theorem 2.

If we further assume $x_i = \Gamma^{1/2}z_i$ where each z_i has sub-Gaussian entries, then the event $\tilde{\Omega}_1' := \{\max_{i\in[p],j\in[n]}|e_i^\top S_{\cdot j}| \le n^{-1/2+\varepsilon}\lambda_{\min}(X^\top X)^{-1/2}\}$ holds with probability tending to one for any small $\varepsilon>0$. On the event $\tilde{\Omega}' := \tilde{\Omega}_1'\cap\tilde{\Omega}_2(c)$, we have $\|(e_i^\top S)\odot(e_i^\top S)\|\| = o(n^{-1/2+\varepsilon}[\lambda_{\min}(X^\top X)]^{-1})$ and thus $\|G(e_i)\|_{Fr} = o(n^{-1/2+\varepsilon}[\lambda_{\min}(X^\top X)]^{-1})$. A calculation same as before yields

$$\mathbb{E}(|\widehat{V}_i - V_i|^4 I_{\tilde{\Omega}'}) = O(n^{-2+2\varepsilon}[\lambda_{\max}(\Sigma)]^4[\lambda_{\min}(X^\top X)]^{-4}).$$

Since $\kappa(\Sigma)^4 = o(n^{1-2\varepsilon})$, and $\tilde{\Omega}'$ holds with probability $1-o(1)$, we can use the same steps as in (46) to conclude the proof. □

A.14.2 Proof of Theorem 5

By the first conclusion in Lemma 4.2 and Slutsky's theorem, it suffices to show that

$$\frac{w_p^\top\hat{\beta} - w_p^\top\beta}{\sqrt{w_p^\top S\Sigma S^\top w_p}} \Rightarrow \mathcal{N}(0,1). \tag{47}$$

Write $w_p^\top(\hat\beta-\beta)=w_p^\top S\Sigma^{1/2}Z$. Condition (19) implies that

$$\frac{\max_{j\in[n]}|w_p^\top S\Sigma^{1/2}e_j|^2}{w_p^\top S\Sigma S^\top w_p}\le\kappa(\Sigma)\frac{\max_{j\in[n]}|w_p^\top Se_j|^2}{\|w_p^\top S\|^2}\to 0.$$

We conclude (47) from Lyapunov's Central Limit Theorem.

A.15 Proof of Theorem

Proof. Let

$$D_1=\hat\beta^\top A\hat\beta-\operatorname{tr}AS(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top-\beta^\top A\beta$$

and

$$D_2=\operatorname{tr}AS(\operatorname{diag}\widehat{\vec{\Sigma}})S^\top-\operatorname{tr}A\operatorname{Cov}(\hat\beta).$$

Using $\hat\beta=S(X\beta+\Sigma^{1/2}Z)$ and $\operatorname{tr}A\operatorname{Cov}(\hat\beta)=\operatorname{tr}AS\Sigma S^\top$, we have

$$D_1+D_2=Z^\top BZ-\mathbb{E}\,Z^\top BZ+2\zeta^\top Z, \tag{48}$$

where

$$B=\Sigma^{1/2}S^\top AS\Sigma^{1/2},\quad \zeta=\Sigma^{1/2}S^\top A\beta.$$

If $\|B\|/\|B\|_{Fr}\to 0$, then due to Condition 1, we can apply Lemma A.5 to $A_n=B$, $B_n=\zeta$ and $c=(1,2)^\top$. By Condition 2, we obtain

$$\begin{aligned}\operatorname{Var}(D_1+D_2)&=\sum_{i=1}^n(\mathbb{E}\,Z_i^4-3)B_{ii}^2+2\operatorname{tr}B^2+4\|\zeta\|^2+4\sum\zeta_iB_{ii}\,\mathbb{E}\,Z_i^3\\&=(2\|B\|_{\mathrm{Fr}}^2+4\|\zeta\|^2)(1+o(1)).\end{aligned} \tag{49}$$

Therefore

$$\frac{D_1+D_2}{\sqrt{2\|B\|_{\mathrm{Fr}}^2+4\|\zeta\|^2}}\Rightarrow\mathcal{N}(0,1). \tag{50}$$

If $\|B\|/\|\zeta\|\to 0$ and $\|B\|_{Fr}/\|B\|=O(1)$, we have $\operatorname{Var}(Z^\top BZ/\|\zeta\|)=o(1)$, thus $(Z^\top BZ-\operatorname{tr}B)/\|\zeta\|=o_P(1)$. Condition 1 implies that $\zeta^\top Z/(\|\zeta\|)\Rightarrow\mathcal{N}(0,1)$. Therefore we have $(D_1+D_2)/(2\|\zeta\|)\Rightarrow\mathcal{N}(0,1)$. This combined with $2\|\zeta\|/\sqrt{2\|B\|_{Fr}^2+4\|\zeta\|^2}\to 1$ still implies (50).

Denote the diagonal part of $S^\top AS$ as $D_{S^\top AS}$. Note that D_2 can be expressed as a quadratic form:

$$\begin{aligned}D_2&=\operatorname{tr}[S^\top AS(\operatorname{diag}\widehat{\vec{\Sigma}})]-\operatorname{tr}S^\top AS\Sigma\\&=1_p^\top D_{S^\top AS}(Q\odot Q)^{-1}(\widehat{\varepsilon}\odot\widehat{\varepsilon})-\operatorname{tr}S^\top AS\Sigma\\&=Z^\top\Sigma^{1/2}Q\operatorname{diag}[1_p^\top D_{S^\top AS}(Q\odot Q)^{-1}]Q\Sigma^{1/2}Z-\operatorname{tr}S^\top AS\Sigma.\end{aligned}$$

We have

$$\begin{aligned}\operatorname{Var}(D_2)&\le C\|\Sigma^{1/2}Q\operatorname{diag}[1_p^\top D_{S^\top AS}(Q\odot Q)^{-1}]Q\Sigma^{1/2}\|_{\mathrm{Fr}}^2\\&\le C\lambda_{\max}(\Sigma)^2\|1_p^\top D_{S^\top AS}\|^2=C\lambda_{\max}(\Sigma)^2\sum_{i=1}^n(S_{\cdot i}^\top AS_{\cdot i})^2\end{aligned}$$

for some constant C, where the second step uses (38) and Condition 1 in Proposition 4.1. Since $\kappa(\Sigma) = O(1)$, we have $\mathrm{Var}(D_2)/\|B\|_{\mathrm{Fr}}^2 \to 0$. This implies that

$$\frac{D_2}{\sqrt{2\|B\|_{\mathrm{Fr}}^2 + 4\|\zeta\|^2}} \to_P 0. \tag{51}$$

By (50) and (51), we conclude the proof for the case when $\zeta \neq 0$.

When $\zeta = 0$, all the above arguments remain valid, as the terms involving ζ can be directly ignored; for example, in (49). Moreover, Condition 3 can be simplified to $\|B\|/\|B\|_{\mathrm{Fr}} \to 0$. □

Proof. Let $C_i = I_{n_i} - \frac{1}{n_i} 1_{n_i} 1_{n_i}$, which is the centering matrix of size n_i. We can readily verify that Q is a block-diagonal matrix given by $\mathrm{diag}(C_1, C_2, \cdots, C_p)$. For the Hadamard estimator, $\widehat{\vec{\Sigma}}_{\mathrm{Had}} = (Q \odot Q)^{-1}(\widehat{\varepsilon} \odot \widehat{\varepsilon}) = (Q \odot Q)^{-1}[(Qy) \odot (Qy)]$. Since Q is block-diagonal, we partition y accordingly as $y = (y_{[1]}, y_{[2]}, \ldots, y_{[p]})$. Thus $\widehat{\vec{\Sigma}}_{\mathrm{Had}}$ decomposes as $(C_i \odot C_i)^{-1}[C_i y_{[i]} \odot C_i y_{[i]}]$. At the same time, we can decompose the KSS estimator $\widehat{\vec{\Sigma}}_{\mathrm{KSS}}$ as $(\mathrm{diag}(C_i))^{-1}[y_{[i]} \odot (C_i y_{[i]})]$, and decompose the MW estimator as $(\mathrm{diag}(C_i))^{-1}[(C_i y_{[i]}) \odot (C_i y_{[i]})]$.

Consider $M := X^\top(\mathrm{diag}\,\widehat{\vec{\Sigma}})X$ for the three methods of obtaining $\widehat{\vec{\Sigma}}$. To verify that $\mathrm{Cov}(\hat{\beta}) = S(\mathrm{diag}\,\widehat{\vec{\Sigma}})S^\top$ are the same for the three methods, it suffices to show that M is the same. It can be readily verified that $M_{ij} = 0$ if $1 \leq i \neq j \leq p$. In what follows, we check the diagonal entries.

For the Hadamard estimator, $M_{ii} = 1_{n_i}^\top (I - C_i)(C_i \odot C_i)^{-1}[C_i y_{[i]} \odot C_i y_{[i]}]$. We consider $i = 1$ without loss of generality. After direct calculations, we obtain $(C_1 \odot C_1)^{-1} = (\alpha_1 I_{n_1} - \theta_1 n_1^{-1} 1_{n_1} 1_{n_1}^\top)$ where $\alpha_1 = (1 - 2/n_1)^{-1}$ and $\theta_1 = n_1^{-1}(1 - 2/n_1)^{-1}(1 - 1/n_1)^{-1}$. Then

$$1_{n_1}^\top (I - C_1)(C_1 \odot C_1)^{-1} = 1_{n_1}^\top (C_1 \odot C_1)^{-1} = (\alpha_1 - \theta_1) 1_{n_1}^\top = (1 - 1/n_1)^{-1} 1_{n_1}^\top,$$

and thus

$$\begin{aligned} M_{11} &= (1 - 1/n_1)^{-1} \sum_{i=1}^{n_1} \sum_{j_1, j_2}^{n_1} (C_1)_{ij_1} y_{j_1} (C_1)_{ij_2} y_{j_2} \\ &= (1 - 1/n_1)^{-1} \sum_{j_1, j_2}^{n_1} y_{j_1} (C_1)_{j_1 j_2} y_{j_2}. \end{aligned}$$

For the KSS estimator,

$$\begin{aligned} M_{11} &= 1_{n_1}^\top (I - C_1)(\mathrm{diag}(C_1))^{-1}[y_{[1]} \odot (C_1 y_{[1]})] \\ &= (1 - 1/n_1)^{-1} \sum_{j_1, j_2}^{n_1} y_{j_1} (C_1)_{j_1 j_2} y_{j_2}. \end{aligned}$$

For the MW estimator,

$$\begin{aligned} M_{11} &= 1_{n_1}^\top (\mathrm{diag}(C_1))^{-1}[(C_1 y_{[1]}) \odot (C_1 y_{[1]})] \\ &= (1 - 1/n_1)^{-1} \sum_{j_1, j_2}^{n_1} y_{j_1} (C_1)_{j_1 j_2} y_{j_2}. \end{aligned}$$

This completes the proof. □

B Additional Simulation Results

This section includes a discussion on the bias of the MW estimator and additional simulation results mentioned in Section 5.

B.1 Bias of MW estimator

Due to the seemingly good performance in simulations of MW estimator, it is of interest to investigate further when MW can be problematic. Consider the inference on $w_p^\top \beta$. The bias of the MW estimator of the asymptotic variance can be written as

$$\sum_{i=1}^{n} (w^\top S_{\cdot i})^2 \sum_{j=1}^{n} \frac{Q_{ij}^2}{Q_{ii}} (\Sigma_j - \Sigma_i)$$

The sign of the bias is generally inconclusive. To gain more insight, consider the setting where $\Sigma_j = g(|w_p^\top S_{\cdot j}|)$ for some increasing function g. Such a structure relates to the case where $\Sigma_j = g(\gamma^\top x_j)$ where x_j is the j-th row of X and γ is a loading vector measuring how noise heteroskedasticity depends on observations. Taking g to be the square function corresponds to a noise model where $\varepsilon_i = \gamma^\top x_i Z_i$ with Z_i i.i.d. Taking g to be exponential function yields a similar model as studied in (Cook and Weisberg, 1983; Zhou and Zou, 2023).

Under this setting, let m be the median of $\{|w^\top S_{\cdot i}|\}_{i=1}^n$. We divide $[n]$ into two sets: $\mathcal{S}_1 = \{i : |w^\top S_{\cdot i}| \le m\}$ and $\mathcal{S}_2 = \{i : |w^\top S_{\cdot i}| > m\}$. Then we expect that $b_i := \sum_{j=1}^n Q_{ij}^2 (\Sigma_j - \Sigma_i)/Q_{ii}$ are positive for those $i \in \mathcal{S}_1$, and negative for those $i \in \mathcal{S}_2$. If Q_{ij} are very close for those $i \neq j$ and Q_{ii} are very close for $i \in [n]$, the magnitude of $|b_i|$ for $i \in \mathcal{S}_1$ and $\mathcal{S}_2$ are close. Then $\sum_{i\in\mathcal{S}_1} (w^\top S_{\cdot i})^2 b_i + \sum_{i\in\mathcal{S}_2} (w^\top S_{\cdot i})^2 b_i$ will be negative. This implies that the MW estimator tends to underestimate the variance of $w_p^\top \hat{\beta}$ in this situation. Case 2 in our experiment corresponds to such a scenario, and the undercoverage for the first coordinate can be attributed to this.

B.2 Case 1

Figure 9 displays the mean type-I error for each coordinate over 1000 simulations in Case 1. It exhibits a similar pattern to that shown in Figure 2. We also plot the mean type-I error in the first and second coordinates over 1000 simulations in Figure 10. Compared with case 2 shown in Figure 5, where MW has an inflated type-I error for the first coordinate, here the MW estimator is more accurate, though still performing slightly worse than the Hadamard-t estimator. We also observe that the Hadamard-t estimator is more accurate compared with the MW estimator for larger p, as also reflected by MAD reported in Figure 9.

B.3 Case 2

It is observed from Table 3 and Figure 12 that the jackknife estimator performs poorly. The performance of KSS is similar to that of the Hadamard-t estimator.

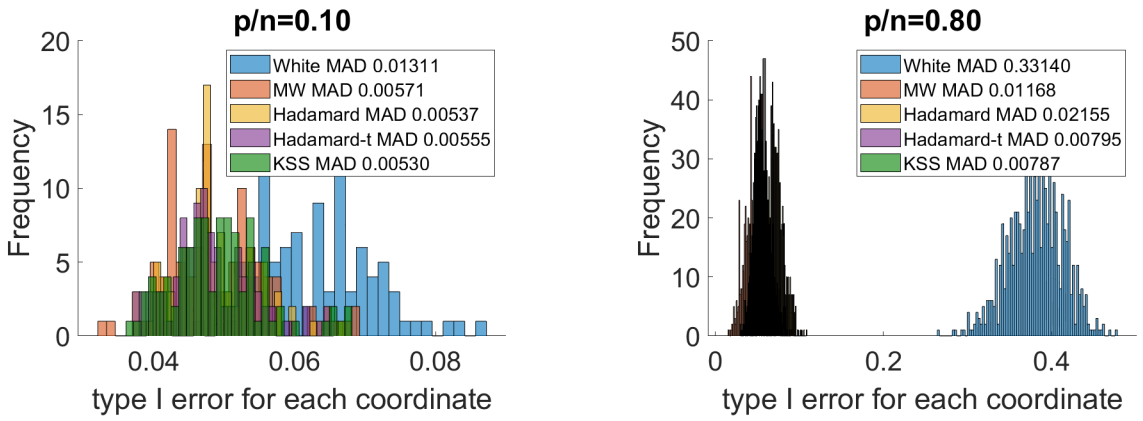


Figure 9: Mean type-I error for each coordinate over 1000 simulations for Case 1.

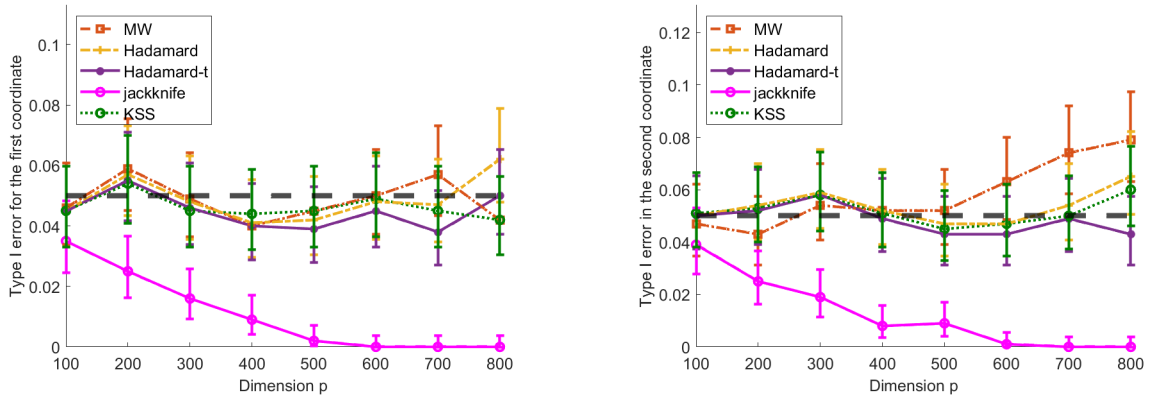


Figure 10: Mean type-I error in the first and second coordinate over 1000 simulations each for Case 1. The error bars represent 95% Clopper-Pearson intervals for the coverage.

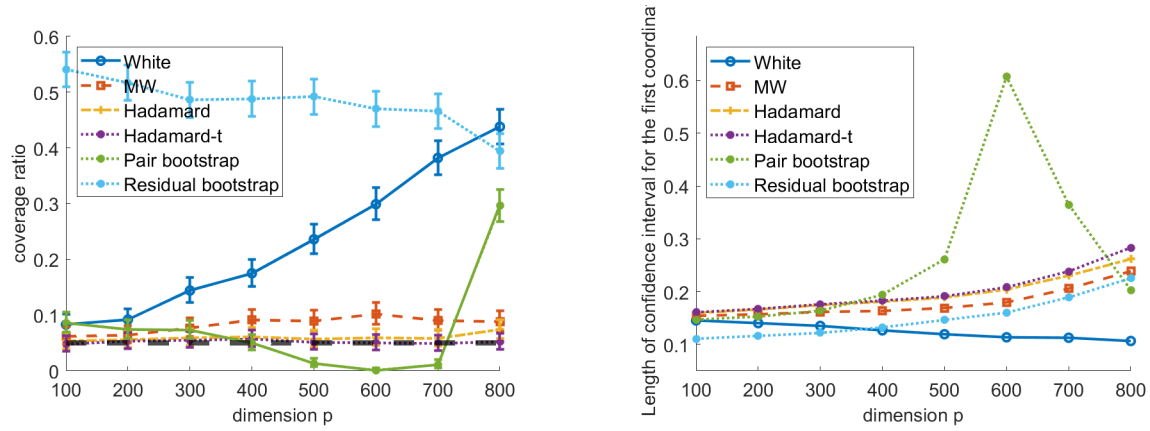


Figure 11: Results for data generated from model in Case 2. Left: Mean type-I error in the first coordinate over 1000 simulations; Right: Mean length of the confidence intervals. The error bars represent 95% Clopper-Pearson intervals for the coverage.

Table 3: Type-1 error in the first coordinate for various methods.

Method \ Dimension	100	200	300	400	500	600	700	800
MW	0.072	0.080	0.088	0.100	0.081	0.100	0.082	0.071
Hadamard	0.067	0.070	0.064	0.060	0.051	0.059	0.0560	0.056
Hadamard-t	0.063	0.067	0.061	0.058	0.045	0.056	0.048	0.040
jackknife	0.061	0.053	0.031	0.030	0.011	0.009	0.003	0.000

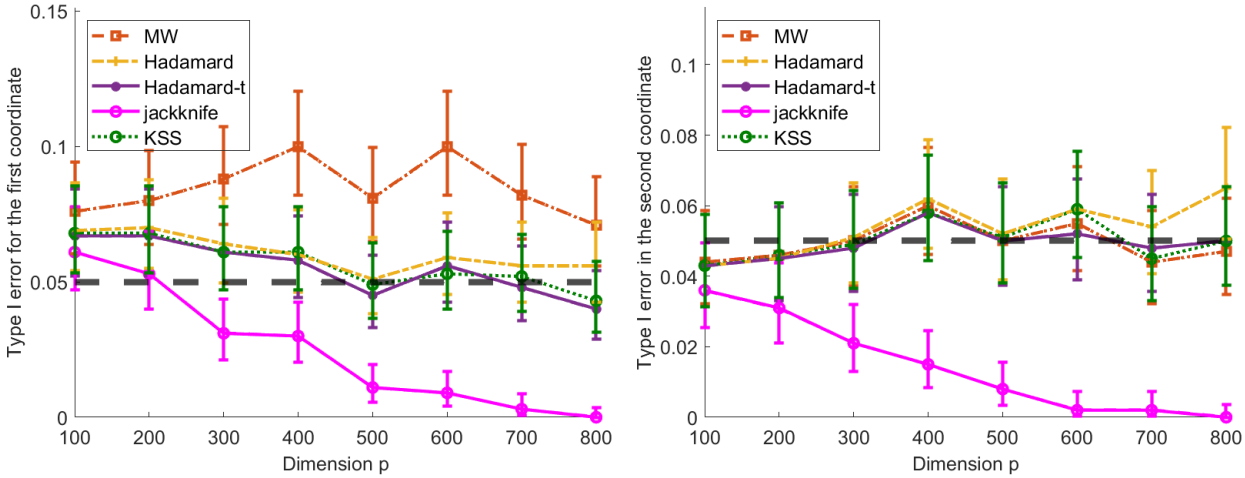


Figure 12: Mean type-I error in the first and second coordinate over 1000 simulations each for Case 2. The error bars represent 95% Clopper-Pearson intervals for the coverage.

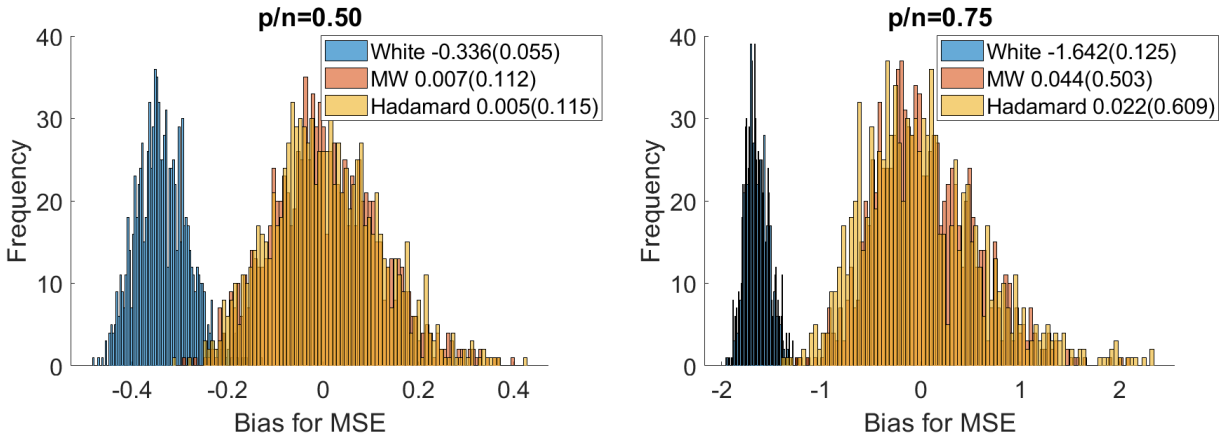
Figure 13 illustrates the bias in estimating the MSE of the OLS estimators in Case 2. We observe the same pattern as in Case 1, where the MW and Hadamard estimators have comparable performance and both are much better than the White estimator.

B.4 Additional simulations related to Case 3

We consider another setting similar to Case 3 but with increased noise variances. Specifically, use the same design as in Case 3 but let Σ be a diagonal matrix where the first $n/2$ diagonal entries equal p and the remaining diagonal entries equal one. This setting corresponds to strong heteroskedasticity. In our simulations with $p = 300$ and $n = 1000$, the R-squared is approximately 0.35, and the signal-to-noise ratio defined as $\|X\beta\|^2/(\|X\beta\|^2 + \mathrm{tr}(\Sigma))$ is 0.078. Figure 14 shows that the Hadamard-t estimator and the KSS estimator outperform the other methods. When $p/n = 0.3$, the Hadamard-t estimator performs better than the KSS estimator.

B.5 Estimating a quadratic form

Table 4 presents results using a similar protocol as in Table 1, except that β is generated from the distribution $s \cdot \mathcal{N}(0, I_p)$ with $s = 0.1$ and $s = 0$, representing weak signals and no regres-



(a) $p/n = 0.5$ (b) $p/n = 0.75$

Figure 13: Bias in estimating MSE for data from Case 2.

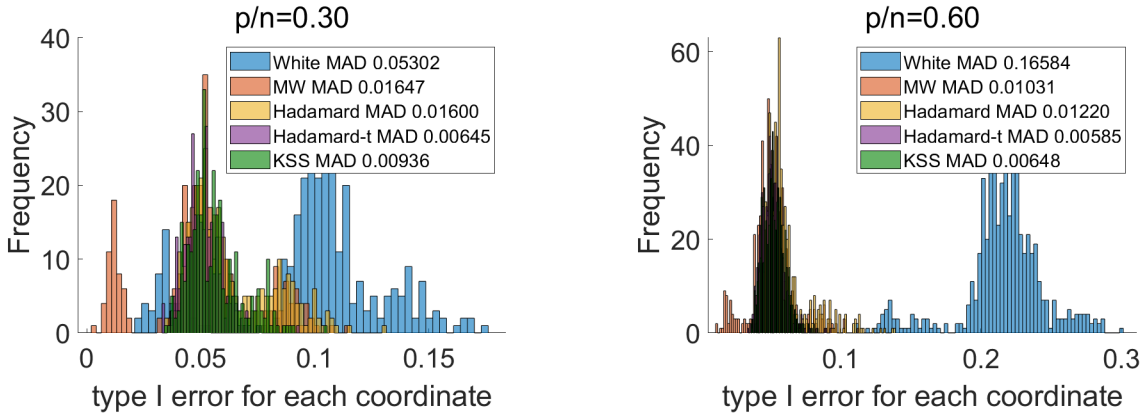


Figure 14: Mean type-I error for each coordinate over 1000 simulations for the design matrix in Case 3, $\Sigma_i = p$ for $i \leq n/2$ and $\Sigma_i = 1$ for $i > n/2$.

sors, respectively. The other settings are the same as those in Table 1. The coverage is accurate when $s = 0.1$. However, when $\beta = 0$, a slight undercoverage is observed when p is small. Table 5 shows the results for the ANOVA design, similar to Table 2, under a relatively weak signal strength. In this setting, the group means α_i for $i \in [J]$ are independently generated from $\mathcal{N}(0, 1)$. We also consider the same signal scenario with the addition of ten covariates, each independently drawn from a t_{10} distribution. The coverage remains accurate. Although the Hadamard estimator does not exactly match the KSS estimator, their performance is very similar across all simulations.

Table 4: Comparison of Hadamard and KSS Estimators for Case 1 using similar protocol as in Table 1 but with weak signal ($s = 0.1$) and zero signal ($s = 0$)

p	$s = 0.1$				$s = 0$			
	Hadamard Mean (SD)	Type-I error	KSS Mean (SD)	True Value	Hadamard Mean (SD)	Type-I error	KSS Mean (SD)	True Value
50	0.5719 (0.0477)	0.044	0.5719 (0.0477)	0.5728	2.4866e-04 (0.0110)	0.109	2.4790e-04 (0.0111)	0
100	1.0183 (0.0756)	0.059	1.0182 (0.0756)	1.0175	4.5982e-05 (0.0176)	0.076	4.3580e-05 (0.0177)	0
150	1.5640 (0.0957)	0.053	1.5640 (0.0956)	1.5644	7.1094e-05 (0.0229)	0.043	6.2418e-05 (0.0229)	0
200	1.9225 (0.1058)	0.057	1.9225 (0.1060)	1.9251	7.8716e-04 (0.0294)	0.043	8.1537e-04 (0.0295)	0
250	2.8398 (0.1298)	0.054	2.9812 (0.1299)	2.8399	-8.4038e-04 (0.0365)	0.032	-8.6540e-04 (0.0365)	0
300	2.6590 (0.1274)	0.058	2.6590 (0.1278)	2.6666	-0.0031 (0.0485)	0.049	-0.0031 (0.0486)	0

Table 5: Comparison of Hadamard and KSS Estimators for σ_α^2 under weak signal, without and with covariates

# groups	Without Covariates				With Covariates			
	Hadamard Mean (SD)	Type-I error	KSS Mean (SD)	True Value	Hadamard Mean (SD)	Type-I error	KSS Mean (SD)	True Value
50	1.0506 (0.0601)	0.0460	1.0506 (0.0601)	1.0484	0.8311 (0.0523)	0.0490	0.8311 (0.0525)	0.8327
100	0.9283 (0.0579)	0.0370	0.9283 (0.0579)	0.9278	1.0593 (0.0580)	0.0330	1.0592 (0.0582)	1.0609
150	0.0896 (0.0641)	0.0470	0.0896 (0.0641)	0.0880	0.9099 (0.0611)	0.0560	0.9101 (0.0622)	0.9080
200	0.9134 (0.0621)	0.0600	0.9134 (0.0621)	0.9152	0.9286 (0.0626)	0.0590	0.9284 (0.0644)	0.9297
250	1.0322 (0.0670)	0.0620	1.0322 (0.0670)	1.0276	0.9254 (0.0647)	0.0610	0.9257 (0.0665)	0.9243
300	0.9688 (0.0625)	0.0650	0.9688 (0.0625)	0.9719	0.8679 (0.0608)	0.0740	0.8679 (0.0620)	0.8712