Robust Inference Under Heteroskedasticity via the Hadamard Estimator

Edgar Dobriban* and Weijie J. Su†

July 22, 2018

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 develop 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, using the second-order Poincaré inequality. Our work provides improved statistical theory and methods for linear regression in high dimensions.

1 Introduction

Drawing statistical inferences from large datasets in a way that is robust to model assumptions is an important problem in statistics and data science. In this paper, we study a central question in this area, performing statistical inference for the unknown regression parameters in linear models.

1.1 Linear models and heteroskedastic noise

The linear regression model

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

is widely used and fundamental 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.

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

*Wharton Statistics Department, University of Pennsylvania. E-mail: dobriban@wharton.upenn.edu.
†Wharton Statistics Department, University of Pennsylvania. E-mail: suw@wharton.upenn.edu.

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:

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

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 26,000 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 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 two works do not appear to be aware of the earlier ones. The estimator is also not 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 also re-discovered the Hadamard estimator in 2017 while studying the bias of White's estimator, and were surprised to find out about the interesting earlier works. We emphasize that the three papers did not study many of the important properties of this estimator, and 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), signal strength, 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). Finally, we also establish the rate of convergence and approximate normality of the estimator, using the second-order Poincaré inequality (Section 4). We also 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`.

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 individual 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.

Estimating SNR. The signal-to-noise ratio (SNR)

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

of the linear model (1) is a fundamental measure that quantifies the fraction of an observational unit's variability explained by its covariates. Here $\|x\| = (\sum_i x_i^2)^{1/2}$ is the usual Euclidean norm of a vector x. 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

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

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.

Estimating signal squared magnitude. A further fundamental problem is estimating the magnitude of the regression coefficient $\|\beta\|^2$. From the identity

$$\mathbb{E}\,\|\hat{\beta}\|^2 = \|\beta\|^2 + \mathrm{tr}\left(\mathrm{Cov}(\hat{\beta})\right),$$

it follows that an unbiased estimator of $\mathrm{tr}\left(\mathrm{Cov}(\hat{\beta})\right)$ is $1_p^\top \widehat{V}$. Thus, an unbiased estimator of the squared signal magnitude is given as

$$\|\hat{\beta}\|^2 - 1_p^\top \widehat{V}.$$

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

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

That is, the vector $\mathrm{diag}(\Sigma)$ 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

$$1_p^\top (Q \odot Q)^{-1} (\widehat{\varepsilon} \odot \widehat{\varepsilon})$$

to estimate the total noise level $\mathrm{tr}(\Sigma) = \sum_{i=1}^n \mathrm{Var}(\varepsilon_i)$ in an unbiased way.

Estimating 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 "Estimating signal squared magnitude," 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 the product of two matrices

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

To clarify, note that $(Q \odot Q)^{-1}$ 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*

$$n \geq p + \frac{1}{2} + \sqrt{2p + \frac{1}{4}}. \tag{6}$$

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 (6).

In light of the above, it is tempting to ask whether (6) 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*

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

has Lebesgue measure zero in $\mathbb{R}^{np}$ if the inequality (6) *is satisfied.*

Therefore, the lower bound in Proposition 2.1 is sharp. Roughly speaking, $n \geq p + O(\sqrt{p})$ 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. As a comparison, our invertibility result is qualitatively stronger as it applies to almost every design matrix under more widely applicable distribution-free models.

To better appreciate this main theoretical contribution of the paper, we consider a random matrix X in the following corollary, which ensures that the Hadamard estimator is well-defined almost surely for many 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 iid 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 additonal conditions.

Specifically, we will consider a *high-dimensional asymptotic* setting, where the dimension n and the sample size p 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, 2009; 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 iid from a distribution with covariance matrix Γ. Let $\Gamma^{1/2}$ be the symmetric square root of Γ.

Theorem 2 (Eigenvalue bounds)**.** *Suppose the rows x_i of X are iid and have the form $x_i = \Gamma^{1/2} z_i$, where z_i have iid entries with mean zero and variance $1/p$, and uniformly bounded $(8 + \delta)$-th moment. Suppose that Γ is invertible. Then, as $n, p \to \infty$ such that $p/n \to \gamma < 1/2$, the matrix $T = Q \odot Q$ satisfies the following eigenvalue bounds almost surely:*

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

Practically speaking, the condition number of T is at most $1/(1 - 2\gamma)$ with high probability. See Section A.3 for a proof. We note that 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 much beyond those considered in their work. They consider only Gaussian designs without correlations.

2.3 Degree-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 $V_j = \mathrm{Var}\,\widehat{\beta}_j$, the expected value of $\widehat{V}_j$. Suppose the degrees of freedom of $\widehat{V}_j$ are d_j. Using the 4-th 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

$$d_j = \frac{2}{\frac{\mathbb{E}\,\widehat{V}_j^2}{V_j^2} - 1} = \frac{2V_j^2}{\mathbb{E}\,\widehat{V}_j^2 - V_j^2}. \tag{7}$$

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

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

Proposition 2.3 (Degrees of freedom)**.** *Under homoskedasticity, we have that the vector of degrees of freedom of $\widehat{V}$, defined in equation* (7)*, has the form*

$$d = \frac{2E}{\operatorname{diag}\left[(S \odot S)1_n 1_n^\top (S \odot S)^\top\right] + 2\operatorname{diag}\left[(S \odot S)(Q \odot Q)^{-1}(S \odot S)^\top\right] - E}, \tag{8}$$

where the division is understood to be entrywise.

See Section A.6 for a proof.

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:

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

In the last line, we have chosen $c = (\Sigma_{\max} + \Sigma_{\min})/2$, where $\Sigma_{\max}$ and $\Sigma_{\min}$ denote the maximal and minimial entries of Σ, respectively. Now, for designs X of aspect ratios $n \times p$ that are not close to 1, and with iid 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

$$\widehat{V}_W = \sum_{j=1}^{n} X_j^2 \widehat{\varepsilon}_j^2,$$

while MacKinnon-White's variance estimator (MacKinnon and White, 1985) can be seen to take the form

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

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 (8) simplifies to

$$d = 1 + \frac{1}{\sum_{j=1}^{n} \frac{X_j^4}{1 - 2X_j^2}}.$$

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:

$$\widehat{C}_{MW} = (X^\top X)^{-1} [X^\top \widehat{\Sigma}_{MW} X] (X^\top X)^{-1}, \tag{9}$$

where $\widehat{\Sigma}_{MW} = \operatorname{diag}(Q)^{-1} \operatorname{diag}(\widehat{\varepsilon})^2$. 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), eq (2.6).

Proposition 2.4 (Bias of classical estimators)**.** *Consider White's covariance estimator defined in* (3) *and MacKinnon-White's estimator defined in* (9)*. Their bias for estimating the coordinate-wise variances of the OLS estimator equals, respectively*

$$b_W = (S \odot S)[(Q \odot Q) - I_n]\Sigma_{vec}$$

for White's covariance estimator, and

$$b_{MW} = (S \odot S)[\operatorname{diag}(Q)^{-1}(Q \odot Q) - I_n]\Sigma_{vec}$$

for MacKinnon-White's estimator. Here Σ_{vec} 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, $\Sigma_{vec} = 1_n$, the vector of all ones. Therefore, $(Q \odot Q)\Sigma_{vec} = vec(\|q_j\|^2)$, 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

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

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

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.

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., 2017; 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

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

where $f_1(X)$ and $f_2(X)$ are polynomials in 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 $f_1(X)$ is 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 $u_1, \dots, u_r \in \mathbb{R}^n$, write $\operatorname{rank}^{\odot}(u_1, \dots, u_r)$ 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 $\{u_1, u_2, \dots, u_r\}$ and $\{u'_1, u'_2, \dots, u'_{r'}\}$ are linearly equivalent, meaning that one can be linearly represented by the other. Then,*

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

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

$$\operatorname{rank}(P \odot P) = \operatorname{rank}^{\odot}(u_1, \dots, u_r).$$

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

Lemma 3.6. *There exists (not necessarily normalized or orthogonalized) $u_1, \dots, u_r$ such that*

$$\operatorname{rank}^{\odot}(u_1, \dots, u_r) = n$$

if $r(r+1)/2 \ge n$.

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

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

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

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

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 $1 \le i, j \le n$. Apparently, $\operatorname{rank}(P \odot P) = \operatorname{rank}(R)$ since $P \odot P = RR^\top$. The (column) rank of R is just $\operatorname{rank}^{\odot}(u_1, \ldots, u_r)$ by Definition 3.3 (recognize that $u_i \odot u_j = u_j \odot u_i$). Hence, $\operatorname{rank}(P \odot P) = \operatorname{rank}^{\odot}(u_1, \ldots, u_r)$.

□

4 Rate of Convergence

We next give two fundamental results characterizing the sampling properties of the Hadamard estimator. The first result bounds the relative error for estimating the vector of variances of all the entries of the OLS estimator. The result is completely explicit. It shows that the estimation error is small when the aspect ratio γ is small. The relative error converges to zero when γ goes to zero. This shows another desirable property of the Hadamard estimator.

Theorem 3 (Rate of convergence)**.** *Under the conditions of Theorem 2, assume in addition that the kurtosis of the entries ε_i of the noise is zero. Let also $V = \operatorname{Var} \hat{\beta}$ the vector of variances of the entries of the OLS estimator. Then, under high-dimensional asymptotics as $n, p \to \infty$ such that $p/n \to \gamma < 1/2$, we have*

$$P\left(\frac{\|\widehat{V} - V\|}{\|\Sigma_{vec}\|} \geq \frac{t}{n}\right) \leq \frac{2c}{t^2} \cdot \frac{1}{(1-\gamma^{1/2})^2 \cdot (1-2\gamma)}$$

a.s., for any constant $c > 1$.

See Section A.9 for a proof. The theorem assumes that the kurtosis of the entries of the noise is zero, but this can be relaxed. Assuming that the fourth moment of the entries is less than a constant $C \geq 3$ times the variance squared of the entries, the result still holds, with the constant 2 in the bound changed to a larger constant $C - 1 \geq 2$.

4.1 Approximate normality

We already know that the estimators $\widehat{V}_i$ are unbiased for the variances of the coordinates of the OLS estimator $V_i = \operatorname{Var} \hat{\beta}_i$, and in the previous section we have seen an inequality bounding the error $\|\widehat{V} - V\|$. In this section, we give a deeper result on the distribution of each V_i.

To study this problem, for simplicity we will assume Gaussian noise, though much of it generalizes to distributions where the noise is approximately Gaussian. Under normality, we can express the residuals as $\widehat{\varepsilon} = Q\Sigma^{1/2}Z$, where Z is a vector of standard normal random variables, $Z \sim \mathcal{N}(0, I_n)$. 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. Therefore, its exact distribution can be obtained as a weighted linear combination of chi-squared random variables. The mean of that distribution is $V_i = \operatorname{Var} \hat{\beta}_i$. We will bound the deviation from normality of the coordinates $\widehat{V}_i$. Since they are linear combinations of chi-squared random variables, this should be true if none of the weights is too large. This is true in fact, and is formalized by a so-called second order Poincaré inequality (Chatterjee, 2009). We will use this result to get our approximate normality result.

Theorem 4 (Approximate normality)**.** *Consider the linear model $y = X\beta + \varepsilon$, where the noise is normally distributed, so that $\varepsilon \sim \mathcal{N}(0, \Sigma)$. Let B_i be a normal random variable with the same mean and variance as the $\widehat{V}_i$ entry of the Hadamard estimator. Then we have the total variation error bound*

$$d_{TV}(\widehat{V}_i, B_i) \leq C \frac{\lambda_{\max}}{(\sum_j \lambda_j^2)^{1/2}},$$

where $C = 4 \cdot 5^{1/2} \cdot 3^{1/4}$ is a numerical constant, and λ_j are the eigenvalues of

$$W_i = \Sigma^{1/2} Q \operatorname{diag}(A_i) Q \Sigma^{1/2}.$$

Here $A_i^\top$ is the i-th row $A = (S \odot S)(Q \odot Q)^{-1}$. Moreover $\lambda_{\max}$ is the largest eigenvalue of W_i.

See Section A.10 for a proof. In principle, this result could justify using normal confidence intervals for inference on V_i as soon the upper bound provided is small. Moreover, the upper bound in result can

Table 1: Type I error for the first coordinate.

$\gamma = p/n$	White	MW	Hadamard	Hadamard-t
0.5	0.172	0.045	0.042	0.039
0.75	0.347	0.059	0.053	0.047

be simplified as follows. First, we can upper bound $\lambda_{\max}(W_i) \leq \Sigma_{\max}\lambda_{\max}(Q\,\mathrm{diag}(A_i)Q)$. Second, we can lower bound

$$\sum_j \lambda_j^2 = \|W_i\|_{Fr}^2 \geq \Sigma_{\min}\|Q\,\mathrm{diag}(A_i)Q\|_{Fr}^2 = \Sigma_{\min}A_i^\top QA_i.$$

Therefore, defining $\kappa := \kappa(\Sigma)$ as the condition number of Σ, we obtain the simplified upper bound

$$C \cdot \kappa(\Sigma)\frac{\lambda_{\max}(Q\,\mathrm{diag}(A_i)Q)}{A_i^\top QA_i}.$$

The improvement from the upper bound stated in the theorem is that this bound decouples simply 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., Then, for any guess on the condition number of Σ, one gets an upper bound on the deviation from normality.

5 Numerical Results

In this section, we present several numerical simulations supporting our theoretical results.

5.1 Mean type I error over all coordinates

In Figure 1, 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. We take X to have iid standard normal entries, and the noise to be $\varepsilon = \Sigma^{1/2}Z$, where Z has iid standard normal entries. The noise covariance matrix Σ is the diagonal matrix of eigenvalues of an AR-1 covariance matrix T, with $T_{ij} = \rho^{i-j}$. We take $n = 100$, 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 samples 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 Type I error over one coordinate

The situation is more nuanced, however, when we look at individual coordinates. In Table 1, we report the empirical type I error of the methods for the first coordinate, where the average is over the Monte Carlo trials. In this case, the MW estimator can be both liberal and conservative, while the Hadamard estimator is closer to having the right coverage.

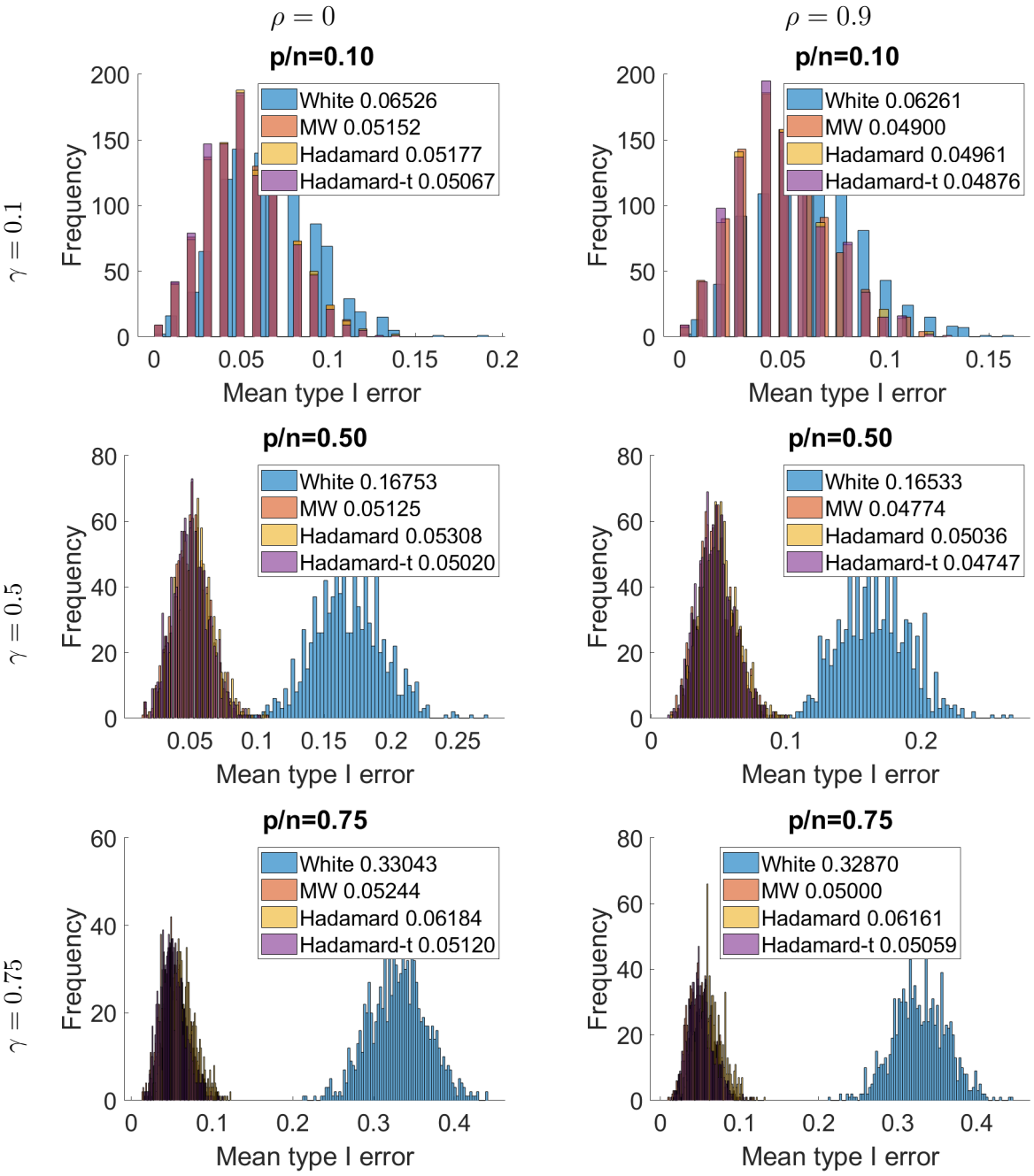
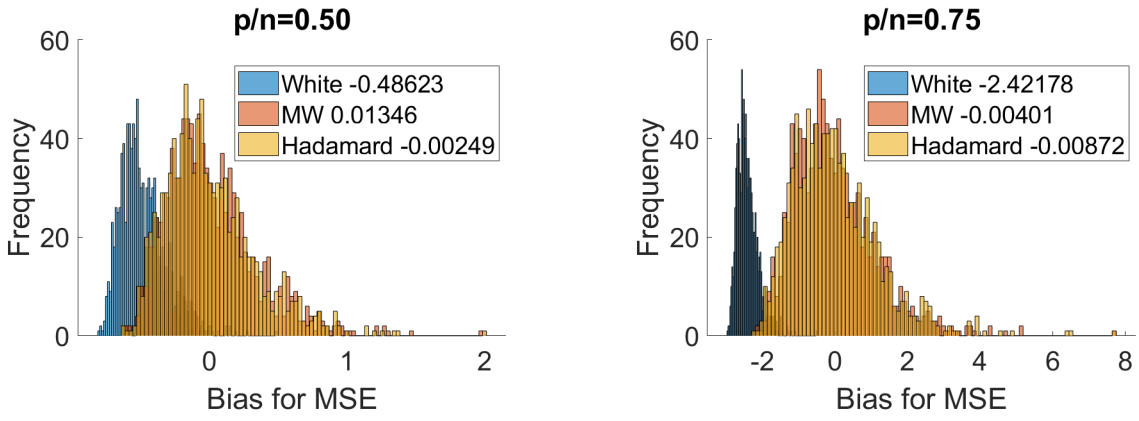
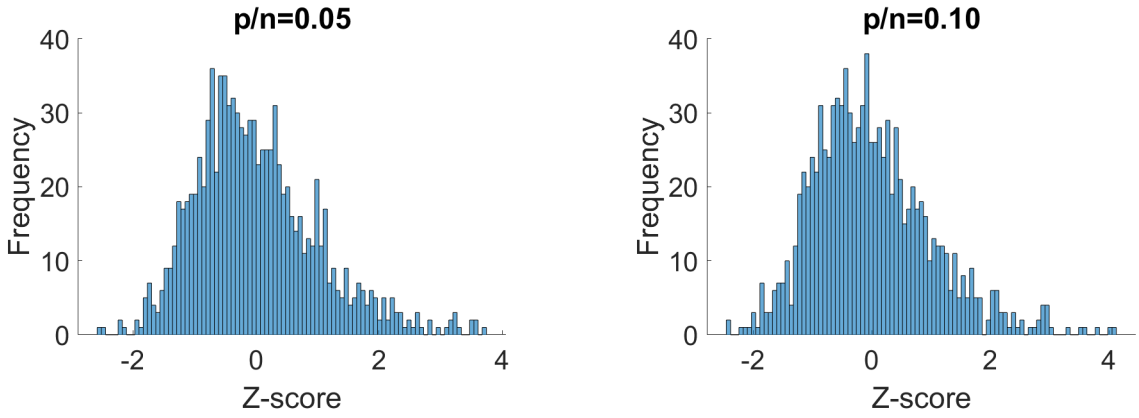


Figure 1: Mean type I error over all coordinates.



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

Figure 2: Bias in estimating MSE.



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

Figure 3: Distribution of z-scores of a fixed coordinate of the Hadamard estimator.

5.3 Estimating MSE

In Figure 2, we show the bias in estimating the MSE of the OLS estimator for the three methods. For each method, we use the estimator which equals the sum of the variances of the individual components. We use the same setup as above.

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. Moreover, when $\gamma = 1/2$, the Hadamard estimator is significantly more accurate than MacKinnon-White's.

5.4 Approximate Normality

In Figure 3, we show the distribution of z-scores of a fixed coordinate of the Hadamard estimator. 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.

6 Discussion and Future Work

In this paper we have developed a new method for constructing confidence intervals for the OLS estimator under heteroskedasticity. We have also provided several fundamental theoretical results. In particular, we have shown that the estimator is well-defined and well-conditioned for certain random design models.

There are several important directions for future research. A few came up during our investigations. 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 (8)? 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., 2017). Perhaps most closely related to our work, Li and Yao (2015) 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, 2017) 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., 2016) 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.

7 Acknowledgements

The authors thank Jason Klusowski for valuable discussions and feedback on an earlier version of the manuscript.

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 Y. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, 2009.

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*, pages 1–25, 2017.

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 Statistical Planning and Inference*, 108 (1-2):121–136, 2002.

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. D. Cattaneo, M. Jansson, and W. K. Newey. Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 0(0):1–12, 2018.

S. Chatterjee. Fluctuations of eigenvalues and second order poincaré inequalities. *Probability Theory and Related Fields*, 143(1-2):1–40, 2009.

M. Chen, C. Gao, Z. Ren, et al. A general decision theory for hubers 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. *arXiv preprint arXiv:1606.03940*, 2016.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. *arXiv preprint arXiv:1703.00893*, 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. *Proc. Natl. Acad. Sci. USA*, 110(36):14557–14562, 2013.

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

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. 1994.

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. 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.

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. *arXiv preprint arXiv:1510.00097*, 2015.

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

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 Statistical 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, 2008.

H. Wang, P.-S. Zhong, and Y. Cui. Empirical likelihood ratio tests for coefficients in high dimensional heteroscedastic linear models. *Statistica Sinica*, 2017.

K. D. West and W. K. Newey. A simple, positive semi-definite, heteroskedasticity and autocorrelation 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.

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

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. *arXiv preprint arXiv:1711.05381*, 2017.

A Proof details

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 Σ_{vec} be the vector of diagonal entries of Σ. Then

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

Alternatively, let u be a vector. Then

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

Proof. Let m_i be the rows of M. Let also $\Sigma = \operatorname{diag} \operatorname{Cov}(v)$. Then, the i-th entry of the left hand side equals

$$\mathbb{E}(m_i \odot m_i)^\top (v \odot v) = \mathbb{E} \sum_j m_{ij}^2 v_j^2 = \sum_j m_{ij}^2 \Sigma_j.$$

The i-th entry of the right hand side equals

$$m_i^\top \Sigma m_i = \sum_j m_{ij}^2 \Sigma_j.$$

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.

□

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 (RHS) 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 (LHS) is $\mathbb{E}(S \odot S)(\varepsilon \odot \varepsilon) = (S \odot S)\Sigma_{vec}$, where we vectorize Σ, writing $\Sigma_{vec} = (\Sigma_{11}, \dots, \Sigma_{nn})^\top$. The equality follows because $\operatorname{Cov}(\varepsilon) = \Sigma$ is diagonal. Thus, by the lemma, we have

$$V = \operatorname{diag} \operatorname{Cov}(\hat{\beta}) = (S \odot S)\Sigma_{vec}.$$

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 LHS equals $\mathbb{E}(\widehat{\varepsilon} \odot \widehat{\varepsilon}) = (Q \odot Q)\Sigma_{vec}$. The RHS 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, the conclusion of using the lemma for the third time is

$$\operatorname{diag} \operatorname{Cov}(\widehat{\varepsilon}) = \operatorname{diag} Q\Sigma Q = (Q \odot Q)\Sigma_{vec}$$

Putting together the above three equations, we obtain that $\widehat{V}$ is unbiased, namely $\mathbb{E}\, \widehat{V} = \operatorname{diag} \operatorname{Cov}(\hat{\beta})$, if

$$A(Q \odot Q)\Sigma_{vec} = (S \odot S)\Sigma_{vec}.$$

This is a system of linear equations. The equation holds for any Σ if and only if

$$A(Q \odot Q) = (S \odot S).$$

If $Q \odot Q$ is invertible, then we can write

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

This shows that the original estimator $\widehat{V}$ has the required form, finishing the proof.

A.2 Proof of Proposition 2.1

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

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

Therefore, in order for $Q \odot Q$ to be invertible, we need

$$n \leq \binom{n-p+1}{2}.$$

By solving the quadratic inequality, this is equivalent to $p \leq [2n + 1 - (8n+1)^{1/2}]/2$.

To prove the claim about ranks, let $A = \sum_{i=1}^{r} v_i v_i^\top$ be the eigendecomposition of A. Here v_i are orthogonal, but not necessarily of unit norm. Then,

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

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,

$$Q = X(X^\top X)^{-1} X^\top = Z(Z^\top Z)^{-1} Z^\top.$$

Therefore, we can work with $\Gamma = I_p$.

The next step is to reduce the bounds on eigenvalues to bounds on certain quadratic forms. Let us define the matrices $R_i = X^\top X - x_i x_i^\top = \sum_{j \neq i} x_j x_j^\top$. See Section A.4 for a proof.

Lemma A.2 (Reduction to quadratic forms)**.** *We have the following two bounds on the eigenvalues of T:*

$$\lambda_{\max}(T) \leq \max_i \frac{1}{1 + x_i^\top R_i^{-1} x_i},$$

and

$$\lambda_{\min}(T) \geq \min_i \frac{1 - x_i^\top R_i^{-1} x_i}{(1 + x_i^\top R_i^{-1} x_i)^2}.$$

To bound these expression, we will use the following well-known statement about concentration of quadratic forms.

Lemma A.3 (Concentration of quadratic forms, consequence of Lemma B.26 in Bai and Silverstein (2009))**.** *Let $x \in \mathbb{R}^p$ be a random vector with i.i.d. entries and $\mathbb{E}[x] = 0$, for which $\mathbb{E}\left[(\sqrt{p} x_i)^2\right] = \sigma^2$ and $\sup_i \mathbb{E}\left[(\sqrt{p} x_i)^{4+\eta}\right] < C$ for some $\eta > 0$ and $C < \infty$. Moreover, let A_p be a sequence of random $p \times p$ symmetric matrices independent of x, with uniformly bounded eigenvalues. Then the quadratic forms $x^\top A_p x$ concentrate around their means at the following rate*

$$P(|x^\top A_p x - p^{-1}\sigma^2 \operatorname{tr} A_p|^{2+\eta} > C) \leq C p^{-(1+\eta/4)}.$$

Lemma A.3 requires a small proof, see Section A.5.

By assumption, the rows of our matrix X satisfy the above assumptions, for $\sigma^2 = 1$, and some $\eta > 0$. In particular, x_{ij} are iid random variables of zero mean and variance $1/p$. By taking $\eta = 4 + \delta$ for some $\delta > 0$, we obtain by the Borel-Cantelli lemma that uniformly over all i

$$x_i^\top R_i^{-1} x_i - p^{-1} \operatorname{tr} R_i^{-1} \to_{a.s.} 0.$$

Therefore, from our earlier result we obtain

$$\limsup \lambda_{\max}(T) \leq \limsup \max_i \frac{1}{1 + p^{-1} \operatorname{tr} R_i^{-1}},$$

and

$$\liminf \lambda_{\min}(T) \geq \liminf \min_i \frac{1 - p^{-1} \operatorname{tr} R_i^{-1}}{(1 + p^{-1} \operatorname{tr} R_i^{-1})^2}.$$

Now, by the Marchenko-Pastur (MP) theorem (Bai and Silverstein, 2009, Theorem 3.6), the empirical spectral distribution of each γR_i converges to the standard MP law with parameter $\gamma < 1$. The reason for normalization by γ is that $\mathbb{E}\, x_{ij}^2 = 1/p$, whereas the MP law refers to matrices of the form $n^{-1} \sum_{i=1}^n z_i z_i^\top$, for z_i with unit variance entries.

Thus, $p^{-1} \operatorname{tr} R_i^{-1} \to \mathbb{E}\, T^{-1}$ a.s., where T is distributed as a MP random variable with parameter γ. It is also well known that $\mathbb{E}\, T^{-1} = \gamma/(1-\gamma)$ (see e.g., Bai and Silverstein, 2009; Yao et al., 2015). Moreover, the difference between $\operatorname{tr} R_i^{-1}$ and $\operatorname{tr} R_j^{-1}$ can be bounded using the formula $A^{-1} - B^{-1} = B^{-1}(A - B)A^{-1}$. The details are omitted for brevity. It follows that we have the uniform convergence $\max_i |\operatorname{tr} R_i^{-1} - \gamma/(1-\gamma)| \to 0$.

Hence we obtain $\limsup \lambda_{\max}(T) \leq 1/[1 + \gamma/(1-\gamma)] = 1 - \gamma$, and also the lower bound $\liminf \lambda_{\min}(T) \geq (1-\gamma)(1-2\gamma)$. This finishes the argument.

A.4 Proof of Lemma A.2

We need to bound the smallest and largest eigenvalues of T. Now $T_{ij} = Q_{ij}^2 = (\delta_{ij} - x_i^\top R^{-1} x_j)^2$, where $R = X^\top X$. We will use the following well-known rank one perturbation formula:

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

We will also use a "leave-one-out" argument which has roots in random matrix theory (see e.g., Bai and Silverstein, 2009; Paul and Aue, 2014; Yao et al., 2015). Let

$$R_i = X^\top X - x_i x_i^\top = \sum_{j \neq i} x_j x_j^\top.$$

Then, $R^{-1} = R_i^{-1} - \frac{R_i^{-1} x_i x_i^\top R_i^{-1}}{1 + x_i^\top R_i^{-1} x_i}$.

We get that the quantity that is squared in the i, j-th entry of T is

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

Also, working on the diagonal, 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, 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), with $T = Q \odot Q$, we have

$$\lambda_{\max}(T) \leq \max_i (T_{ii} + \sum_{j \neq i} |T_{ij}|).$$

Thus, an upper bound on the operator norm of T is the maximum over all i of

$$\frac{1 + \sum_{j \neq i} (x_i^\top R_i^{-1} x_j)^2}{(1 + x_i^\top R_i^{-1} x_i)^2}.$$

Now, the sum in the numerator can be written as $x_i^\top R_i^{-1} (\sum_{j \neq i} x_j x_j^\top) R_i^{-1} x_i = x_i^\top R_i^{-1} x_i$. Therefore, there is an unexpected cancellation, which simplifies the analysis a great deal. Thus,

$$\lambda_{\max}(T) \leq \max_i \frac{1}{1 + x_i^\top R_i^{-1} x_i}.$$

Similarly, for the smallest eigenvalue, by the Gershgorin disk theorem, (Horn and Johnson, 1990, Thm 6.1.1), we have

$$\lambda_{\min}(T) \geq \min_i (T_{ii} - \sum_{j \neq i} |T_{ij}|).$$

We can express

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

This shows that

$$\lambda_{\min}(T) \geq \min_i \frac{1 - x_i^\top R_i^{-1} x_i}{(1 + x_i^\top R_i^{-1} x_i)^2}.$$

This finishes the proof.

A.5 Proof of Lemma A.3

We will use the following Trace Lemma quoted from Bai and Silverstein (2009).

Lemma A.4 (Trace Lemma, Lemma B.26 of Bai and Silverstein (2009))**.** *Let y be a p-dimensional random vector of i.i.d. elements with mean 0. Suppose that $\mathbb{E}\left[y_i^2\right] = 1$, and let A_p be a fixed $p \times p$ matrix. Then*

$$\mathbb{E}\left[|y^\top A_p y - \operatorname{tr} A_p|^q\right] \leq C_q \left\{ \left(\mathbb{E}\left[y_1^4\right] \operatorname{tr}[A_p A_p^\top]\right)^{q/2} + \mathbb{E}\left[y_1^{2q}\right] \operatorname{tr}[(A_p A_p^\top)^{q/2}] \right\},$$

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 $\operatorname{tr}[(A_p A_p^\top)^{q/2}] \leq pC^q$ and $\operatorname{tr}[A_p A_p^\top] \leq pC^2$. Consider now a random vector x with the properties assumed in the present lemma. For $y = \sqrt{p}x/\sigma$ and $q = 2 + \eta/2$, using that $\mathbb{E}\left[y_i^{2q}\right] \leq C$ and the other the conditions in Lemma A.3, Lemma A.4 thus yields

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

or equivalently $\mathbb{E}\left[|x^\top A_p x - \frac{\sigma^2}{p}\operatorname{tr} A_p|^{2+\eta}\right] \le Cp^{-(1+\eta/4)}$.

By Markov's inequality applied to the $2+\eta$-th moment of $\varepsilon_p = x^\top A_p x - \frac{\sigma^2}{p}\operatorname{tr} A_p$, we obtain as required

$$P(|\varepsilon_p|^{2+\eta} > C) \le Cp^{-(1+\eta/4)}.$$

□

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

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

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, the (ij)-element of $\mathbb{E}(\widehat{\varepsilon}\widehat{\varepsilon}^\top)^{\odot 2}$ is

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

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

$$[(Q\odot Q)1_n]\,[(Q\odot Q)1_n]^\top = (Q\odot Q)1_n1_n^\top(Q\odot Q),$$

and $\left(\sum_{l=1}^n Q_{il}Q_{jl}\right)^2$ is the (ij)-element of

$$Q^2\odot Q^2 = Q\odot Q.$$

Summarizing the calculation above, we obtain

$$\mathbb{E}(\widehat{\varepsilon}\widehat{\varepsilon}^\top)^{\odot 2} = \sigma^4(Q\odot Q)1_n1_n^\top(Q\odot Q) + 2\sigma^4 Q\odot Q,$$

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. Denoting

$$E_2 = \operatorname{diag}\left[(X^\top X)^{-1}\right] \odot \operatorname{diag}\left[(X^\top X)^{-1}\right],$$

we get the degrees of freedom as a vector for all j is

$$d = \frac{2E_2}{\operatorname{diag}\left[(S \odot S)1_n 1_n^\top (S \odot S)^\top\right] + 2\operatorname{diag}\left[(S \odot S)(Q \odot Q)^{-1}(S \odot S)^\top\right] - E_2},$$

where the division is understood to be entrywise. This finishes 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 $X_j^2/(1 - 2X_j^2)$. 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, we find

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

Since $X^\top X = 1$, we have $E = 1$. Finally, we need to find

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

Since $S = X^\top$, $u = X \odot X$, and $Q \odot Q = D + uu^\top$, so that $u^\top 1_n = 1$, this simplifies to

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

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,

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

As we saw,

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

Thus,

$$\mathrm{diag}\,\mathbb{E}\,\widehat{C}_W = \mathrm{diag}[S\,\mathrm{diag}[(Q\odot Q)\Sigma_{vec}]S^\top] = (S\odot S)(Q\odot Q)\Sigma_{vec}$$

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

$$b_W = (S\odot S)[(Q\odot Q) - I_n]\Sigma_{vec}.$$

This is the desired result.

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

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

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

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

Thus the bias is

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

This is the desired result, finishing the proof.

A.9 Proof of Theorem 3

We would like to bound $\|\widehat{V} - V\|$, where by $\|\cdot\|$ denotes usual Euclidean vector norm. Recall that

$$V = (S\odot S)\Sigma_{vec}$$

abd

$$\widehat{V} = (S\odot S)(Q\odot Q)^{-1}(\widehat{\varepsilon}\odot\widehat{\varepsilon})$$

So, we can bound by the definition of operator norms,

$$\|\widehat{V} - V\| \le \|S\odot S\|_{op}\|(Q\odot Q)^{-1}\|_{op}\|(\widehat{\varepsilon}\odot\widehat{\varepsilon}) - (Q\odot Q)\Sigma_{vec}\|$$

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

1. Bounding $\|S\odot S\|_{op}$.
   Schur's inequality (e.g., Horn and Johnson, 1994, Thm. 5.5.1), states that
   $$\|S\odot S\|_{op} \le \|S\|_{op}^2.$$
   Moreover, $\|S\|_{op} = 1/\sigma_{\min}(X)$.
   By the Bai-Yin law, (Bai and Yin, 1993), $\sigma_{\min}(X) \ge n^{1/2} - p^{1/2} - c$, for any constant $c > 0$ almost surely (a.s.). The meaning of constants can change from line to line.
   Assuming that there is a constant $c < 1$ such that $p/n < c$, we also get $\sigma_{\min}(X) \ge c'(n^{1/2} - p^{1/2})$ for any constant $c' < 1$ (whp).
   Thus, we get the bound
   $$n\|S\odot S\|_{op} \le nc\frac{1}{(n^{1/2} - p^{1/2})^2} \le c\frac{1}{(1-\gamma^{1/2})^2}$$
   a.s., for any constant $c > 1$.

2. Bounding $\|(Q \odot Q)^{-1}\|_{op}$.
   This follows from Theorem 2, see Section A.3. That argument shows that

$$\|(Q \odot Q)^{-1}\|_{op} \leq c\frac{1}{(1-\gamma)(1-2\gamma)}$$

   a.s., for any constant $c > 1$, under high-dimensional asymptotics.

3. Bounding $\alpha = \|(\widehat{\varepsilon} \odot \widehat{\varepsilon}) - (Q \odot Q)\Sigma_{vec}\|$.
   We can express $\alpha^2 = \sum_i \alpha_i^2$, where

$$\alpha_i^2 = (\widehat{\varepsilon}_i^2 - (q_i \odot q_i)^\top \Sigma_{vec})^2.$$

   Since $\mathbb{E}\,\widehat{\varepsilon}_i^2 = (q_i \odot q_i)^\top \Sigma_{vec}$, which follows from the earlier unbiasedness argument, we have $\mathbb{E}\,\alpha_i^2 = \mathrm{Var}\,\widehat{\varepsilon}_i^2$.
   An easy calculation shows that, with $\Gamma_k = \mathbb{E}\,\varepsilon_k^4$, we have

$$\mathrm{Var}\,\widehat{\varepsilon}_i^2 = \sum_k q_{ik}^4(\Gamma_k - 3\Sigma_k^4) + 2[(q_i \odot q_i)^\top \Sigma_{vec}]^2.$$

   Now the kurtosis is zero by assumption, so $\Gamma_k - 3\Sigma_k^4 = 0$. Therefore, we can bound by Markov's inequality:

$$P(\alpha \geq t) \leq \frac{\sum_i \mathbb{E}\,\alpha_i^2}{t^2} = \frac{2\sum_i[(q_i \odot q_i)^\top \Sigma_{vec}]^2}{t^2} = \frac{2 \cdot \|(Q \odot Q)\Sigma_{vec}\|^2}{t^2}.$$

   Using a similar approach to above, the bound for $\|Q \odot Q\|_{op}$ follows from Theorem 2, see Section A.3. That argument shows that

$$\|Q \odot Q\|_{op} \leq c(1-\gamma)$$

   a.s., for any constant $c > 1$, under high-dimensional asymptotics. Hence a.s. under high-dimensional asymptotics

$$P(\alpha \geq t) \leq \frac{2c(1-\gamma)\|\Sigma_{vec}\|^2}{t^2}.$$

In conclusion, under high-dimensional asymptotics

$$P\left(\frac{\|\widehat{V} - V\|}{\|\Sigma_{vec}\|} \geq \frac{t}{n}\right) \leq \frac{2c}{t^2} \cdot \frac{1}{(1-\gamma^{1/2})^2 \cdot (1-2\gamma)}$$

a.s., for any constant $c > 1$. This proves the required result.

A.10 Proof of Theorem 4

Since we assumed Gaussian noise, we have

$$\widehat{\varepsilon} = Q\varepsilon \sim \mathcal{N}(0, Q\Sigma Q).$$

So, we can write $\widehat{\varepsilon} = Q\Sigma^{1/2}Z$, where $Z \sim \mathcal{N}(0, I_n)$. Let us denote $M = Q\Sigma^{1/2}$.

Now, we have $\widehat{V}_i = A_i^\top(\widehat{\varepsilon} \odot \widehat{\varepsilon})$, where $A_i^\top$ is the i-th row of $A = (S \odot S)(Q \odot Q)^{-1}$. So,

$$\widehat{V}_i = \sum_j A_{ij}\widehat{\varepsilon}_j^2 = \sum_j A_{ij}(\sum_k M_{jk}Z_k)^2 = \sum_{k,l} Z_k Z_k(\sum_j A_{ij}M_{jk}M_{jl})$$

This shows that

$$\widehat{V}_i = Z^\top W_i Z,$$

where W_i is the $n \times n$ matrix

$$W_i = M^\top \operatorname{diag}(A_i) M.$$

Letting Λ_i be the diagonal matrix of eigenvalues of W_i, we obtain that the distribution of $\widehat{V}_i$ is a weighted mixture of chi-squared random variables with weights λ_j, $j = 1, \ldots, n$.

We will use the second order Poincare inequality, see Chatterjee (2009), Theorem 2.2. This states that the total variation we need to bound is at most

$$d_{TV}(\widehat{V}_i, B_i) \le 2 \cdot 5^{1/2} \cdot \frac{\kappa_1 \kappa_2}{\sigma^2},$$

where

$$\kappa_1 = [\mathbb{E}\, \|\nabla g(Z)\|^4]^{1/4}$$

and

$$\kappa_2 = [\mathbb{E}\, \|\nabla^2 g(Z)\|_{op}^4]^{1/4},$$

while $g(x) = x^\top W_i x$ is the function mapping the normal random vector Z into $\widehat{V}_i$, so that $\widehat{V}_i = g(Z)$. In addition, σ^2 is the variance of $g(Z)$.

Now, it can be checked that

$$\nabla g(Z) = 2 W_i Z,$$

so, for another normal random vector Z', denoting $L = \sum_{j=1}^n (\lambda_j Z_j')^2$,

$$2^{-4}\, \mathbb{E}\, \|\nabla g(Z)\|^4 = \mathbb{E}[\sum_{j=1}^n (\lambda_j Z_j')^2]^2 = \operatorname{Var} L + (\mathbb{E}\, L)^2.$$

Next,

$$\operatorname{Var} L = \sum_{j=1}^n \operatorname{Var}[(\lambda_j Z_j')^2] = 2 \sum_{j=1}^n \lambda_j^4.$$

Meanwhile, $\mathbb{E}\, L = \sum_{j=1}^n \lambda_j^2$, and thus

$$2^{-4}\, \mathbb{E}\, \|\nabla g(Z)\|^4 = 2 \sum_{j=1}^n \lambda_j^4 + (\sum_{j=1}^n \lambda_j^2)^2 \le 3 (\sum_{j=1}^n \lambda_j^2)^2.$$

We obtain that $\kappa_1 \le 2 \cdot 3^{1/4} (\sum_j \lambda_j^2)^{1/2}$.

Continuing,

$$\nabla^2 g(Z) = 2 W_i,$$

is non-random, hence

$$\kappa_2 = 2 \|W_i\|_{op} = 2 \lambda_{\max}.$$

Finally, we can calculate σ^2. Since $\widehat{V}_i = Z^\top W_i Z$, as we have already noticed, the distribution of $\widehat{V}_i$ is a weighted mixture of chi-squared random variables with weights λ_j, $j = 1, \ldots, n$. Hence

$$\sigma^2 = \operatorname{Var} \widehat{V}_i = 2 \sum_{j=1}^n \lambda_j^2.$$

Putting everything together, we obtain that

$$d_{TV}(\widehat{V}_i, B_i) \le 2 \cdot 5^{1/2} \cdot \frac{2 \cdot 3^{1/4} (\sum_j \lambda_j^2)^{1/2} \cdot 2 \lambda_{\max}}{2 \sum_{j=1}^n \lambda_j^2},$$

which simplifies to the desired result. This finishes the proof.