Electronic Journal of Statistics
Vol. 16 (2022) 5437–5487
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2070

Improved estimators for semi-supervised high-dimensional regression model*

Ilan Livne

The Faculty of Industrial Engineering and Management, Technion, Israel
e-mail: ilan.livne@campus.technion.ac.il

David Azriel

The Faculty of Industrial Engineering and Management, Technion, Israel
e-mail: davidazr@technion.ac.il

Yair Goldberg

The Faculty of Industrial Engineering and Management, Technion, Israel
e-mail: yairgo@technion.ac.il

Abstract: We study a high-dimensional linear regression model in a semi-supervised setting, where for many observations only the vector of covariates X is given with no responses Y. We do not make any sparsity assumptions on the vector of coefficients, nor do we assume normality of the covariates. We aim at estimating the signal level, i.e., the amount of variation in the response that can be explained by the set of covariates. We propose an estimator, which is unbiased, consistent, and asymptotically normal. This estimator can be improved by adding zero-estimators arising from the unlabeled data. Adding zero-estimators does not affect the bias and potentially can reduce the variance. We further present an algorithm based on our approach that improves any given signal level estimator. Our theoretical results are demonstrated in a simulation study.

MSC2020 subject classifications: Primary 62J05; secondary 62F10.
Keywords and phrases: Linear regression, semi-supervised learning, U-statistics, variance estimation, zero-estimators.

Received August 2021.

1. Introduction

High-dimensional data analysis, where the number of predictors is larger than the sample size, is a topic of current interest. In such settings, an important goal is to estimate the signal level τ^2 and the noise level σ^2, i.e., to quantify how much variation in the response variable Y can be explained by the covariates X, versus how much of the variation is left unexplained. Formally, the variance of Y can be written as $\mathrm{Var}[E(Y|X)] + E[\mathrm{Var}(Y|X)] \equiv \tau^2 + \sigma^2$. For example, in disease classification using DNA microarray data, where the number of potential predictors, say the genotypes, is enormous per each individual, one may wish to

*Supported by BSF grant number 2018112.

understand how disease risk is associated with genotype versus environmental factors.

Estimating the signal and noise levels is important even in a low-dimensional setting. In particular, a statistical model partitions the total variability of the response variable into two components: the variance of the fitted model and the variance of the residuals. This partition is at the heart of techniques such as ANOVA and linear regression, where the signal and the noise levels might also be commonly referred to as explained versus unexplained variation, or between treatments versus within treatments variation. Moreover, in model selection problems, τ^2 and σ^2 may be required for computing popular statistics, such as Cp, AIC, BIC and R^2. Both τ^2 and σ^2 are also closely related to other important statistical problems, such as genetic heritability and signal to noise ratio [2, 7, 13, 25, 30]. Hence, developing good estimators for these quantities is a desirable goal.

When the number of covariates p is much smaller than the number of observations n, and a linear model is assumed, the ordinary least squares (henceforth, OLS) method provides us straightforward estimators for τ^2 and σ^2. However, when $p > n$, it becomes more challenging to perform inference on τ^2 and σ^2 without further assumptions. Under the assumption of sparse regression coefficients, several methods for estimating the signal level have been proposed. [10] introduced a refitted cross-validation method for estimating σ^2. Their method includes a two-stage procedure where a variable-selection technique is performed in the first stage, and OLS is used to estimate σ^2 in the second stage. [25] introduced the scaled lasso algorithm that jointly estimates the noise level and the regression coefficients by an iterative lasso procedure. A recent related work by [28] considers, as we do here, a semi-supervised setting. In their work, Cai and Guo proposed an estimator of τ^2, which integrates both labelled and unlabelled data and works well when the regression coefficient vector is sparse. For more related works, see the literature reviews of [28] and [30].

In practice, the sparsity assumption may not hold in some areas of interest such as genetic and chemical pollutants studies [5, 22]. In such cases, the effects of individual covariates tend to be weak and dense rather than strong and sparse. Hence, considering only a small number of significant coefficients can lead to biases and inaccuracies. One famous example is the problem of missing heritability, i.e., the gap between heritability estimates from genome-wide-association-studies (GWAS) and the corresponding estimates from twin studies [7, 33]. For example, by the year 2010, GWAS studies had identified a relatively small number of covariates that collectively explained around 10% of the total variations in the trait *height*, which is a small fraction compared to 80% of the total variations that were explained by twin studies [31]. Identifying all the GWAS covariates affecting a trait, and measuring how much variation they capture, is believed to bridge some of the heritability gap [33]. With that in mind, methods that heavily rely on the sparsity assumption may underestimate τ^2 by their nature.

Rather than assuming sparsity, or other structural assumptions on the coefficient vector β, a different approach for high-dimensional inference is to as-

sume some knowledge about the covariates distribution. [9] uses the method-of-moments to develop several asymptotically-normal estimators of τ^2 and σ^2, when the covariates are assumed to be Gaussian.

[17] proposed a procedure, which is based on singular value decomposition and convex optimization techniques, that provides estimates and confidence intervals under the assumption of Gaussian covariates. In both methods, the Gaussian assumption was needed to prove consistency and asymptotic-normality, and it is not clear how robust these methods are when the Gaussian assumption is violated.

We aim at relaxing the sparsity and the Gaussian assumptions under the semi-supervised setting. The term *semi-supervised setting* is used to describe a situation where a large amount of unlabeled data (covariate data without the corresponding responses) is available. For simplicity we generally assume that the distribution of the covariates X is known. In our simulation study (Section 5) we consider the situation where distribution of X is not known exactly but rather estimated from an unlabeled dataset.

We begin by introducing a naive estimator for the signal level τ^2. When the covariates are assumed Gaussian, we show that this estimator is asymptotically equivalent to an estimator suggested by [9]. We then show how the naive estimator can be improved using zero-estimators. Zero-estimators are introduced in the UMVUE literature [1, 21, 23], and are also used as a variance reduction technique in the Monte-Carlo simulation literature [3, 12, 19]. When the distribution of the covariates is known, an easy construction of many zero-estimators is feasible as shown in Section 4.

The contribution of this paper is threefold. First, we develop a notion of optimal oracle-estimators, which are served as benchmark for other estimators. Second, we propose two novel estimators that improve initial estimators of τ^2 and study their properties. Third, we provide an algorithm that in principle can improve any given estimator of τ^2.

The rest of this work in organized as follows. In Section 2 we describe the high-dimensional semi-supervised setting and introduce the naive estimator. In Section 3 we review the zero-estimator approach and suggest a new notion of optimality with respect to linear families of zero-estimators. An optimal oracle estimator of τ^2 is also presented. In Section 4 we apply the zero-estimator approach to improve the naive estimator. We then study some theoretical properties of the improved estimators. Simulation results are given in Section 5. Section 6 demonstrates how the zero-estimator approach can be generalized to other estimators. A discussion is given in Section 7, while the proofs are provided in the Appendix.

2. The naive estimator

2.1. Preliminaries

We begin with describing our setting and assumptions. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. observations drawn from some unknown distribution where $X_i \in \mathbb{R}^p$ and

$Y_i \in \mathbb{R}$. We consider a semi-supervised setting, where we have access to infinite i.i.d. observations of the covariates. Thus, we essentially assume we know the covariate distribution. Notice that the assumption of known covariate distribution has already been presented and discussed in the context of high-dimension regression without using the term "semi-supervised learning" [4, 17].

We consider the linear model

$$Y_i = \beta^T X_i + \epsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $E(\epsilon_i|X_i) = 0$ and $E(\epsilon_i^2|X_i) = \sigma^2$. As in [8] and [9], we assume that the intercept term β_0 is zero, which can be achieved in practice by centering the Y's. It is noteworthy that the theory presented in this paper can be developed without assuming $\beta_0 = 0$. However, it leads to cumbersome expressions which do not add any important insights to our current theoretical results and, therefore, are not included here. Let (X, Y) denote a generic observation and let σ_Y^2 denote the variance of Y. Notice that it can be decomposed into signal and noise components,

$$\sigma_Y^2 = \mathrm{Var}(X^T\beta + \epsilon) = \beta^T \mathrm{Cov}(X)\beta + \mathrm{Var}(\epsilon) = \beta^T \mathbf{\Sigma}\beta + \sigma^2, \tag{2}$$

where $\mathrm{Var}(\epsilon) = E(\epsilon^2) = \sigma^2$ and $\mathrm{Cov}(X) = \mathbf{\Sigma}$.

The *signal* component $\tau^2 \equiv \beta^T \mathbf{\Sigma}\beta$ can be thought of as the total variance explained by a linear function of the covariates, while the *noise* component σ^2 can be thought of as the variance left unexplained. We assume that $E(X) \equiv \mu$ are known and also that $\mathbf{\Sigma}$ is invertible. Therefore, we can apply the linear transformation $X \mapsto \mathbf{\Sigma}^{-1/2}(X - \mu)$ and assume w.l.o.g. that

$$E(X) = \mathbf{0} \quad \text{and} \quad \mathbf{\Sigma} = \mathbf{I}. \tag{3}$$

It follows by (2) that $\sigma_Y^2 = \|\beta\|^2 + \sigma^2$, which implies that in order to evaluate σ^2, it is enough to estimate both σ_Y^2 and $\|\beta\|^2$. The former can be easily evaluated from the sample, and the main challenge is to derive an estimator for $\|\beta\|^2$ in the high-dimensional setting.

2.2. A naive estimator

In order to find an unbiased estimator for $\|\beta\|^2 = \sum_{j=1}^p \beta_j^2$ we first consider the estimation of β_j^2 for each j. A straightforward approach is given as follows: Let $W_{ij} \equiv X_{ij}Y_i$ for $i = 1, ..., n$, and $j = 1, ..., p$. Notice that

$$E\left(W_{ij}\right) = E\left(X_{ij}Y_i\right) = E\left[X_{ij}\left(\beta^T X_i + \varepsilon_i\right)\right] = \beta_j,$$

Now, since $\{E(W_{ij})\}^2 = E(W_{ij}^2) - \mathrm{Var}(W_{ij})$, a natural unbiased estimator for β_j^2 is

$$\hat{\beta}_j^2 \equiv \frac{1}{n}\sum_{i=1}^n W_{ij}^2 - \frac{1}{n-1}\sum_{i=1}^n (W_{ij} - \overline{W}_j)^2 = \frac{1}{n(n-1)}\sum_{i_1 \neq i_2}^n W_{i_1 j} W_{i_2 j}, \tag{4}$$

where $\overline{W}_j = \frac{1}{n}\sum_{i=1}^{n} W_{ij}$. Thus, unbiased estimates of $\tau^2 \equiv \|\beta\|^2$ and σ^2 are given by

$$\hat{\tau}^2 = \sum_{j=1}^{p} \hat{\beta}_j^2 = \binom{n}{2}^{-1} \sum_{i_1<i_2} W_{i_1}^T W_{i_2}, \qquad \hat{\sigma}^2 = \hat{\sigma}_Y^2 - \hat{\tau}^2, \tag{5}$$

where $W_i = (W_{i1}, ..., W_{ip})^T$ and $\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. We use the term *naive* estimator to describe $\hat{\tau}^2$ since its construction is relatively simple and straightforward. The naive estimator was also discussed by [18]. A similar estimator was proposed by [9]. Specifically, let

$$\hat{\tau}^2_{Dicker} = \frac{\left\|\mathbf{X}^T\mathbf{Y}\right\|^2 - p\|\mathbf{Y}\|^2}{n\,(n+1)}$$

where $\mathbf{X}$ is the $n \times p$ design matrix and $\mathbf{Y} = (Y_1, ..., Y_n)^T$. The following lemma shows that $\hat{\tau}^2$ and $\hat{\tau}^2_{Dicker}$ are asymptotically equivalent under some conditions.

Lemma 1. *Assume the linear model in* (1) *and* $X_i \overset{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I})$*, and that* $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$*. When* $\tau^2 + \sigma^2$ *is bounded and* p/n *converges to a constant, then,*

$$\sqrt{n}\left(\hat{\tau}^2 - \hat{\tau}^2_{Dicker}\right) \xrightarrow{p} 0.$$

Note that in this paper we are interested in a high-dimensional regression setting and therefore we study the limiting behaviour when n and p go together to ∞. Using Corollary 1 from [9], which computes the asymptotic variance of $\hat{\tau}^2_{Dicker}$, and the above lemma, we obtain the following corollary.

Corollary 1. *Under the assumptions of Lemma 1,*

$$\sqrt{n}\left(\frac{\hat{\tau}^2 - \tau^2}{\psi}\right) \xrightarrow{D} N(0,1)\,,$$

where $\psi = 2\left\{\left(1 + \frac{p}{n}\right)\left(\sigma^2 + \tau^2\right)^2 - \sigma^4 + 3\tau^4\right\}$.

Let $\mathbf{A} = E\left(W_i W_i^T\right)$ and $\|\mathbf{A}\|_F^2$ denoted the Frobenius norm of $\mathbf{A}$. The variance of the naive estimator $\hat{\tau}^2$ under model (1), without assuming normality, is given by the following proposition.

Proposition 1. *Assume model* (1) *and additionally that* $\beta^T\mathbf{A}\beta$ *and* $\|\boldsymbol{A}\|_F^2$ *are finite. Then,*

$$Var\left(\hat{\tau}^2\right) = \frac{4\,(n-2)}{n\,(n-1)}\left[\beta^T\mathbf{A}\beta - \|\beta\|^4\right] + \frac{2}{n\,(n-1)}\left[\|\mathbf{A}\|_F^2 - \|\beta\|^4\right], \tag{6}$$

Notice that under the assumptions of Lemma 1, which included Gaussian covariates and noises, the expression in (6) reduces to ψ^2/n, approximately.

Proposition 1 is more general than Corollary 1 and holds without any Gaussian assumptions. Furthermore, the proof of Proposition 1 does not require homoscedasticity of ϵ.

The following proposition shows that the naive estimator is consistent under some minimal assumptions.

Proposition 2. *Assume model (1) and additionally that* $\tau^2 + \sigma^2 = O(1)$ *and* $\frac{\|A\|_F^2}{n^2} \to 0$. *Then,* $\hat{\tau}^2$ *is consistent. Moreover, when the columns of* $\mathbf{X}$ *are independent and both* p/n *and* $E(X_{ij}^4)$ *are bounded, then* $\frac{\|A\|_F^2}{n^2} \to 0$ *holds and* $\hat{\tau}^2$ *is* $\sqrt{n}$*-consistent.*

3. Oracle estimator

In this section we introduce the zero-estimator approach and study how it can be used to improve the naive estimator. In Section 3.1 we present the zero-estimator approach. An illustration of this approach is given in Section 3.2. Section 3.3 introduces a new notion of optimality with respect to linear families of zero-estimators. We then find an optimal oracle estimator of τ^2 and calculate its improvement over the naive estimator.

3.1. The zero-estimator approach

Before we describe the zero-estimator approach, we explain our motivation and discuss why this approach is useful in the semi-supervised setting. The naive estimator $\hat{\tau}^2$ is a symmetric unbiased U-statistic. For non-parametric distributions, if the vector of order statistic is sufficient and complete, there can exist at most one symmetric unbiased estimator, and this estimator is the UMVUE [20, Section 2.4]. However, when moments restriction exist, the order statistic is no longer complete (i.e., there are non-trivial zero-estimators) and hence the statement above no longer holds [11, 16]. Thus, by assumption (3), as the first and the second moments of X are restricted, $\hat{\tau}^2$ may not be a UMVUE and can be improved by using zero-estimators.

The idea of using zero-estimators to reduce variance is not new. Zero-estimators are introduced in the UMVUE literature [1, 21, 23]. When a complete and sufficient statistic is not available, zero-estimators can be used to reduce variance or to examine whether a particular estimator is a UMVUE [21, Theorem 1.7, p.85]. In the Monte-Carlo simulations literature, variance reduction using zero-estimators is referred to as the control variates method [3, 12, 19]. Notice that zero-estimators in the semi-supervised setting are natural since knowing the distribution of the covariates enables an easy construction of zero-estimators.

We now describe the approach in general terms. Consider a random variable $V \sim P$, where P belongs to a family of distributions $\mathcal{P}$. Let $g(V)$ be a zero-estimator, i.e., $E_P[g(V)] = 0$ for all $P \in \mathcal{P}$. Let $T(V)$ be an unbiased estimator of a certain quantity of interest θ. Then, the statistic $U_c(V)$, defined by $U_c(V) =$

$T(V) - cg(V)$ for a fixed constant c, is also an unbiased estimator of θ. The variance of $U_c(V)$ is

$$\text{Var}[U_c(V)] = \text{Var}[T(V)] + c^2\text{Var}[g(V)] - 2c \cdot \text{Cov}[T(V), g(V)]. \tag{7}$$

Minimizing $\text{Var}[U_c(V)]$ with respect to c yields the minimizer

$$c^* = \frac{\text{Cov}[T(V), g(V)]}{\text{Var}[g(V)]}. \tag{8}$$

Notice that $\text{Cov}[T(V), g(V)] \neq 0$ implies $\text{Var}[U_{c^*}(V)] < \text{Var}(T(V))$. In other words, by combining a correlated unbiased estimator of zero with the initial unbiased estimator of θ, one can lower the variance. Note that plugging c^* in (7) reveals how much variance can be potentially reduced,

$$\begin{aligned}\text{Var}[U_{c^*}(V)] =& \text{Var}[T(V)] - [c^*]^2\text{Var}[g(V)] \\ =& \text{Var}[T(V)] - \frac{\{\text{Cov}[T(V), g(V)]\}^2}{\text{Var}[g(V])} = (1-\rho^2)\text{Var}[T(V)],\end{aligned} \tag{9}$$

where ρ is the correlation coefficient between $T(V)$ and $g(V)$. Therefore, it is best to find an unbiased zero-estimator $g(V)$ which is highly correlated with $T(V)$, the initial unbiased estimator of θ. It is important to notice that c^* is an unknown quantity and, therefore, U_{c^*} is not a statistic. However, in practice, one can estimate c^* by some $\hat{c}^*$ and use the approximation $U_{\hat{c}^*}$ instead.

3.2. Illustration of the zero-estimator approach

The following example illustrates how the *zero-estimator approach* can be applied to improve the naive estimator $\hat{\tau}^2$ in the simple linear model setting.

Example 1 ($p = 1$)**.** *Assume model* (1) *with* $X \sim N(0,1)$*. By* (9)*, we wish to find a zero-estimator* $g(X)$ *which is correlated with* $\hat{\tau}^2$*. Consider the estimator* $U_c \equiv \hat{\tau}^2 + cg(X)$*, where* $g(X) \equiv \frac{1}{n}\sum_{i=1}^n (X_i^2 - 1)$ *and* c *is a fixed constant. The variance of* U_c *is minimized by* $c^* = -2\beta^2$ *and one can verify that* $Var(U_{c^*}) = Var(\hat{\tau}^2) - \frac{8}{n}\beta^4$*. For more details see Remark* 2 *in the Appendix.*

The above example illustrates the potential of using additional information that exists in the semi-supervised setting to lower the variance of the naive estimator $\hat{\tau}^2$. However, it also raises the question: *Can we achieve a lower variance by adding a different zero-estimator?* One might attempt to reduce the variance by adding a zero-estimator that is a linear combination of elements of the form $g_k(X) \equiv \frac{1}{n}\sum_{i=1}^n [X_i^k - E(X_i^k)]$, for $k \in \mathbb{N}$. Surprisingly, as shown in Theorem 2 below, the variance of the oracle-estimator $U_{c^*} \equiv \hat{\tau}^2 - 2\beta^2 g(X)$ cannot be further reduced by adding such a zero-estimator. Hence, the oracle-estimator U_{c^*} is optimal with respect to the family of zero-estimators constructed from elements of the form $g_k(X)$.

3.3. Optimal oracle estimator

We now define a new oracle unbiased estimator of τ^2 and prove that under some regularity assumptions this estimator is optimal with respect to a certain family of zero-estimators. Here, optimality means that the variance cannot be further reduced by including additional zero-estimators of *that* given family. We now specifically define our notion of optimality in a general setting.

Definition 1. Let T be an unbiased estimator of θ and let $g_1, g_2, ...$ be a sequence of zero-estimators, i.e., $E_\theta(g_i) = 0$ for $i \in \mathbb{N}$ and for all θ.

Let $\mathcal{G} = \{\sum_{k=1}^{m} c_k g_k : c_k \in \mathbb{R}, m \in \mathbb{N}\}$ be a family of zero-estimators. For a zero-estimator $g^* \in \mathcal{G}$, we say that $R^* \equiv T + g^*$ is an *optimal oracle estimator (OOE)* of θ with respect to $\mathcal{G}$, if $\mathrm{Var}_\theta[R^*] = \mathrm{Var}_\theta[T + g^*] \leqslant \mathrm{Var}_\theta[T + g]$ for all $g \in \mathcal{G}$ and for all θ.

We use the term oracle since $g^* \equiv \sum_{k=1}^{m} c_k^* g_k$ for some optimal coefficients $c_1^*, ..., c_m^*$, which are a function of the unknown parameter θ. The following theorem suggests a necessary and sufficient condition for obtaining an OOE.

Theorem 1. *Let $\mathbf{g}_m = (g_1, ..., g_m)^T$ be a vector of zero-estimators and assume the covariance matrix $M \equiv Var[\mathbf{g}_m]$ is positive definite for every m. Then, R^* is an optimal oracle estimator (OOE) with respect to the family of zero-estimators $\mathcal{G}$ iff R^* is uncorrelated with every zero-estimator $g \in \mathcal{G}$, i.e., $Cov_\theta[R^*, g] = 0$ for all $g \in \mathcal{G}$ and for all θ.*

Theorem 1 is closely related to Theorem 1.7 in Lehmann and Casella [21, p. 85]. While their gives a necessary and sufficient condition for obtaining a UMVUE estimator, our theorem provides the same condition for obtaining an optimal oracle estimator with respect to the family of zero-estimators $\mathcal{G}$.

Returning to our setting, define the following oracle estimator

$$T_{oracle} = \hat{\tau}^2 - 2\sum_{j=1}^{p}\sum_{j'=1}^{p} \psi_{jj'}, \tag{10}$$

where $\psi_{jj'} = \beta_j \beta_{j'} h_{jj'}$ and $h_{jj'} = \frac{1}{n}\sum_{i=1}^{n} [X_{ij}X_{ij'} - E(X_{ij}X_{ij'})]$, and let the $\mathcal{G}$ be the family of zero-estimators of the form $g_{k_1...k_p} = \frac{1}{n}\sum_{i=1}^{n}[X_{i1}^{k_1} \cdot ... \cdot X_{ip}^{k_p} - E(X_{i1}^{k_1} \cdot \ldots \cdot X_{ip}^{k_p})]$, where $(k_1, ..., k_p) \in \{0, 1, 2, 3, ...\}^p \equiv \mathbb{N}_0^p$. The following theorem shows that T_{oracle} is an OOE with respect to $\mathcal{G}$. We comment that the proof of Theorem 2 does not require homoscedasticity of ϵ.

Theorem 2 (General p)**.** *Assume model (1) and additionally that X has moments of all orders. Then, the oracle estimator T_{oracle} defined in (10) is an OOE of τ^2 with respect to $\mathcal{G}$.*

We now compute the variance reduction of T_{oracle} with respect to the naive estimator. The following statement is a corollary of Proposition 1.

Corollary 2. *Assume model (1) and additionally that the columns of* $\boldsymbol{X}$ *are independent. Then,*

$$Var\left(T_{oracle}\right) = Var\left(\hat{\tau}^2\right) - \frac{4}{n}\left\{\sum_{j=1}^{p}\beta_j^4\left[E\left(X_{1j}^4-1\right)\right] + 2\sum_{j\neq j'}\beta_j^2\beta_{j'}^2\right\}. \quad (11)$$

Moreover, in the special case where $X_i \overset{i.i.d}{\sim} N\left(\mathbf{0},\mathbf{I}\right)$. Then, Rewriting (11) yields

$$\mathrm{Var}\left(T_{oracle}\right) = \mathrm{Var}\left(\hat{\tau}^2\right) - \frac{4}{n}\left\{2\sum_{j=1}^{p}\beta_j^4 + 2\sum_{j\neq j'}\beta_j^2\beta_{j'}^2\right\} = \mathrm{Var}\left(\hat{\tau}^2\right) - \frac{8}{n}\tau^4. \quad (12)$$

Notice that by Cauchy–Schwarz inequality, since $E(X^2)=1$ then $E(X^4)\geq 1$, and therefore $\mathrm{Var}(T_{oracle}) < \mathrm{Var}(\hat{\tau}^2)$. The following example provides intuition about the improvement of $\mathrm{Var}(T_{oracle})$ over $\mathrm{Var}(\hat{\tau}^2)$.

Example 2. *Consider a setting where* $n=p$; $\tau^2=\sigma^2=1$ *and* $X_i \overset{i.i.d}{\sim} N\left(\mathbf{0},\mathbf{I}\right)$. *In this case, one can verify by (1) that* $Var(\hat{\tau}^2) = \frac{20}{n} + O(n^{-2})$ *and therefore* $Var(T_{oracle}) = \frac{12}{n} + O(n^{-2})$. *In other words: the optimal oracle estimator* T_{oracle} *reduces (asymptotically) the variance of the naive estimator by* 40%. *Moreover, when* p/n *converges to zero, the reduction is* 66%. *For more details and simulation results for this example, see Remark 3 in the Appendix.*

4. Proposed estimators

In this section we show how to use the zero-estimator approach to derive improved estimators over $\hat{\tau}^2$. In Section 4.1 we show that estimating all p^2 optimal coefficients given in (10) may introduce too much variance. Therefore, Sections 4.2 and 4.3 introduce alternative methods to reduce the number of zero-estimators used in estimation.

4.1. The cost of estimation

The optimal oracle estimator defined in (10) is based on adding p^2 zero-estimators. Therefore, it is reasonable to suggest and study the following estimator instead of the oracle one:

$$T = \hat{\tau}^2 - 2\sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj'},$$

where

$$\hat{\psi}_{jj'} = \frac{1}{n\left(n-1\right)\left(n-2\right)}\sum_{i_1\neq i_2\neq i_3} W_{i_1 j}W_{i_2 j'}\left[X_{i_3 j}X_{i_3 j'} - E\left(X_{i_3 j}X_{i_3 j'}\right)\right],$$

is a U-statistics estimator of $\psi_{jj} \equiv \beta_j \beta_{j'} h_{jj'}$. Notice that $E\left(\hat{\psi}_{jj'}\right) = 0$ and that for $i_1 \neq i_2$ we have $E(W_{i_1 j} W_{i_2 j'}) = \beta_j \beta_{j'}$. Thus, T is an unbiased estimator of τ^2 and we wish to check it reduces the variance of naive estimator $\hat{\tau}^2$. This is described in the following proposition.

Proposition 3. *Assume model (1) and additionally that* $\tau^2 + \sigma^2 = O(1)$; $E(X_{ij}^4) \leq C$ *for some positive constant* C, *and* $p/n = O(1)$. *Then,*

$$\begin{aligned} Var(T) &= Var(T_{oracle}) + \frac{8p^2\sigma_Y^4}{n^3} + O(n^{-2}) \\ &= Var\left(\hat{\tau}^2\right) - \frac{4}{n}\left\{\sum_{j=1}^{p} \beta_j^4 \left[E\left(X_{1j}^4 - 1\right)\right] + 2\sum_{j \neq j'} \beta_j^2 \beta_{j'}^2\right\} \\ &+ \frac{8p^2\sigma_Y^4}{n^3} + O(n^{-2}), \end{aligned} \tag{13}$$

where $\sigma_Y^2 \equiv \tau^2 + \sigma^2$.

Note that the second equation in (13) follows from (11). To build some intuition, consider the case when $X_i \overset{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I})$ and $p = n$. Then, the last equation can be rewritten as

$$\mathrm{Var}(T) = \mathrm{Var}\left(\hat{\tau}^2\right) + \frac{8}{n}\left(2\tau^2\sigma^2 + \sigma^4\right) + O(n^{-2}). \tag{14}$$

Notice that the term $\frac{8}{n}\left(2\tau^2\sigma^2 + \sigma^4\right)$ in (14) reflects the additional variability that comes with the attempt at estimating all p^2 optimal coefficients. Therefore, the estimator T fails to improve the naive estimator $\hat{\tau}^2$ and a similar result holds for $p/n \to c$ for some positive constant c. Thus, alternative ways that improve the naive estimator are warranted, which are discussed next.

4.2. Improvement with a single zero-estimator

A simple way to improve the naive estimator is by adding only a single zero-estimator. More specifically, let $U_{c^*} = \hat{\tau}^2 - c^* g_n$ where $c^* = \frac{\mathrm{Cov}[\hat{\tau}^2, g_n]}{\mathrm{Var}[g_n]}$ and g_n is some zero-estimator. By (9) we have

$$\mathrm{Var}[U_{c^*}] = \mathrm{Var}(\hat{\tau}^2) - \frac{\{\mathrm{Cov}[\hat{\tau}^2, g_n]\}^2}{\mathrm{Var}[g_n]}. \tag{15}$$

Notice that U_{c^*} is an oracle estimator and thus c^* needs to be estimated in order to eventually construct a non-oracle estimator. Let $g_n = \frac{1}{n}\sum_{i=1}^{n} g_i$ be the sample mean of some zero estimators $g_1, ..., g_n$. By (8), it can be shown that

$$c^* = \frac{2\sum_{j=1}^{p} \beta_j \theta_j}{\mathrm{Var}(g_i)}, \tag{16}$$

where $\theta_j \equiv E(S_{ij})$ and $S_{ij} = W_{ij}g_i$. Notice that $\mathrm{Var}(g_i)$ does not depend on i. Derivation of (16) can be found in Remark 4 in the Appendix. Here, we specifically chose $g_i = \sum_{j<j'} X_{ij}X_{ij'}$ as it worked well in the simulations but we do not argue that this is the best choice. Let $T_{c^*} = \hat{\tau}^2 - c^* g_n$ denote the oracle estimator for the specific choice of g_n, and where c^* is given in (16). Notice that by (15) we have

$$\mathrm{Var}(T_{c^*}) = \mathrm{Var}\left(\hat{\tau}^2\right) - \frac{\left[2\sum_{j=1}^{p}\beta_j\theta_j\right]^2}{n\mathrm{Var}(g)}, \tag{17}$$

where g is just a generic g_i for some i. The following example demonstrates the improvement of $\mathrm{Var}(T_{c^*})$ over $\mathrm{Var}(\hat{\tau}^2)$.

Example 3 (Example 2 - continued)**.** *Consider a setting where $n = p$; $\tau^2 = \sigma^2 = 1$; $X_i \overset{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I})$ and $\beta_j = \frac{1}{\sqrt{p}}$ for $j = 1, ..., p$. Notice that this is an extreme non-sparse settings since the signal level τ^2 is uniformly distributed across all p covariates. In this case one can verify that $Var(T_{c^*}) = \frac{12}{n} + O(n^{-2})$, which is approximately 40% improvement over the naive estimator variance (asymptotically). For more details see Remark 5 in the Appendix.*

In the view of (16), a straightforward U-statistic estimator for c^* is

$$\hat{c}^* = \frac{\frac{2}{n(n-1)}\sum_{i_1 \neq i_2}\sum_{j=1}^{p} W_{i_1 j}S_{i_2 j}}{\mathrm{Var}(g)}, \tag{18}$$

where $\mathrm{Var}(g)$ is assumed known as it depends only on the marginal distribution of X. Thus, we suggest the following estimator

$$T_{\hat{c}^*} = \hat{\tau}^2 - \hat{c}^* g_n, \tag{19}$$

and prove that T_{c^*} and $T_{\hat{c}^*}$ are asymptotically equivalent under some conditions.

Proposition 4. *Assume model (1) and additionally that $\tau^2 + \sigma^2$ and p/n are $O(1)$. Also, for every j_1, j_2, j_3, j_4 assume that $E\left(X_{1j_1}^2 X_{1j_2}^2 X_{1j_3}^2 X_{1j_4}^2\right)$ is bounded and that the columns of the design matrix $\boldsymbol{X}$ are independent. Then, $\sqrt{n}\left[T_{c^*} - T_{\hat{c}^*}\right] \xrightarrow{p} 0$.*

We note that the requirement that the columns of $\mathbf{X}$ be independent holds, for example, when X is Gaussian, and this requirement can be relaxed to some form of weak dependence.

4.3. Improvement by selecting small number of covariates

Rather than using a single zero-estimator to improve the naive estimator, we now consider estimating a small number of coefficients of T_{oracle}. Recall that T_{oracle} is based on adding p^2 zero estimators to the naive estimator. This estimation comes

with high cost in terms of additional variability as shown is (14). Therefore, it is reasonable to use only a small number of zero estimators. Specifically, let $\mathbf{B} \subset \{1, ..., p\}$ be a fixed set of some indices such that $|\mathbf{B}| \ll p$ and consider the estimator

$$T_{\mathbf{B}} = \hat{\tau}^2 - 2 \sum_{j,j' \in \mathbf{B}} \hat{\psi}_{jj'}. \tag{20}$$

By the same argument as in Proposition 3 we now have

$$\mathrm{Var}\left(T_{\mathbf{B}}\right) = \mathrm{Var}\left(\hat{\tau}^2\right) - \frac{4}{n}\left\{\sum_{j \in \mathbf{B}} \beta_j^4 \left[E\left(X_{ij}^4\right) - 1\right] + 2 \sum_{j \neq j' \in \mathbf{B}} \beta_j^2 \beta_{j'}^2\right\} + O\left(n^{-2}\right). \tag{21}$$

Also notice that when $X_i \overset{i.i.d}{\sim} N\left(\mathbf{0}, \mathbf{I}\right)$, (21) can be rewritten as

$$\mathrm{Var}\left(T_{\mathbf{B}}\right) = \mathrm{Var}\left(\hat{\tau}^2\right) - \frac{8}{n}\tau_{\mathbf{B}}^4 + O(n^{-2}). \tag{22}$$

where $\tau_{\mathbf{B}}^2 = \sum_{j \in \mathbf{B}} \beta_j^2$. Thus, if $\tau_{\mathbf{B}}^2$ is sufficiently large, one can expect a significant improvement over the naive estimator by using a small number of zero-estimators. For example, when $\tau_{\mathbf{B}}^2 = 0.5$; $p = n$; $\tau^2 = \sigma^2 = 1$, then $T_{\mathbf{B}}$ reduces the $\mathrm{Var}(\hat{\tau}^2)$ by 10%. For more details see Remark 6 in the Appendix.

Notice that we do not assume sparsity of the coefficients. The sparsity assumption essentially ignores covariates that do not belong to the set $\mathbf{B}$. When β_j's for $j \notin \mathbf{B}$ contribute much to the signal level $\tau^2 \equiv \|\beta\|^2$, the sparse approach leads to disregarding a significant portion of the signal, while our estimators do account for this as all p covariates are used in $\hat{\tau}^2$.

The following example illustrates some key aspects of our proposed estimators.

Example 4 (Example 3 - continued)**.** *Let $n = p$; $\tau^2 = \sigma^2 = 1$ and $X_i \overset{i.i.d}{\sim} N\left(\mathbf{0}, \mathbf{I}\right)$. Consider the following two extreme scenarios:*

1. *non-sparse setting: The signal level τ^2 is uniformly distributed over all p covariates, i.e., $\beta_j^2 = \frac{1}{p}$ for all $j = 1, ..., p$.*
2. *Sparse setting: the signal level τ^2 is "point mass" distributed over the set $\boldsymbol{B}$, i.e., $\tau_{\boldsymbol{B}}^2 = \tau^2$.*

Two interesting key points:

1. *In the first scenario the estimator $T_{\boldsymbol{B}}$ has the same asymptotic variance as $\hat{\tau}^2$, while the estimator T_{c^*} reduces the variance by approximately 40%.*
2. *In the second scenario the variance reduction of $T_{\boldsymbol{B}}$ is approximately 40%, while T_{c^*} has the same asymptotic variance as $\hat{\tau}^2$.*

Interestingly, in this example, the OOE estimator T_{oracle} asymptotically improves the naive by 40% regardless of the scenario choice, as shown by (12). For more details see Remark 7 in the Appendix.

A desirable set of indices $\mathbf{B}$ contains relatively small amount of covariates that capture a significant part of the signal level τ^2. There are different methods to choose the covariates that will be included in $\mathbf{B}$, but these are not a primary focus of this work. For more information about covariate selection methods see [24] and [32] and references therein. In Section 5 below we work with a certain selection algorithm defined there. We call δ a covariate *selection algorithm* if for every dataset $(\mathbf{X}_{n\times p}, \mathbf{Y}_{n\times 1})$ it chooses a subset of indices $\mathbf{B}_\delta$ from $\{1, ..., p\}$. Our proposed estimator for τ^2, which is based on selecting small number of covariates, is given in Algorithm 1.

Algorithm 1: Proposed Estimator based on covariate selection

Input: A dataset $(\mathbf{X}_{n\times p}, \mathbf{Y}_{n\times 1})$ and a selection algorithm γ.

1. Calculate the naive estimator $\hat{\tau}^2 = \frac{1}{n(n-1)} \sum_{j=1}^{p} \sum_{i_1 \neq i_2}^{n} W_{i_1 j} W_{i_2 j}$, where $W_{ij} = X_{ij} Y_i$.
2. Apply algorithm γ to $(\mathbf{X}, \mathbf{Y})$ to construct $\mathbf{B}_\gamma$.
3. Calculate the zero-estimator terms:

$$\hat{\psi}_{jj'} \equiv \frac{2}{n(n-1)(n-2)} \sum_{i_1 \neq i_2 \neq i_3} W_{i_1 j} W_{i_2 j'} \left[X_{i_3 j} X_{i_3 j'} - E\left(X_{i_3 j} X_{i_3 j'} \right) \right],$$

for all $j, j' \in \mathbf{B}_\gamma$.

Result: Return $T_\gamma = \hat{\tau}^2 - \sum_{jj' \in \mathbf{B}_\gamma} \hat{\psi}_{jj'}$.

Some asymptotic properties of T_γ are given by the following proposition.

Proposition 5. *Assume there is a set* $\mathbf{B} \equiv \left\{ j : \beta_j^2 > b \right\}$ *where* b *is a positive constant, such that* $|\mathbf{B}| = p_0$ *where* p_0 *is a fixed constant. Also assume that*

$$\lim_{n\to\infty} n \left[P\left(\{ \mathbf{B}_\gamma \neq \mathbf{B} \} \right) \right]^{1/2} = 0,$$

and that $E\left(T_\gamma^4\right)$ *and* $E(T_{\mathbf{B}}^4)$ *are bounded. Then,*

$$\sqrt{n}(T_\gamma - T_{\mathbf{B}}) \xrightarrow{p} 0.$$

Notice that the condition $\lim_{n\to\infty} n \left[P\left(\{ \mathbf{B}_\gamma \neq \mathbf{B} \} \right) \right]^{1/2} = 0$ is stronger than the standard definition of consistency, $\lim_{n\to\infty} P\left(\{ \mathbf{B}_\gamma \neq \mathbf{B} \} \right) = 0$, which is used in the variable-selection literature; see [6] and references therein. However, the convergence rate of many practical selection procedures is exponential, which is much faster than is required for the above condition to hold. For example, the lasso algorithm asymptotically selects the support of β at an exponential rate under some assumptions [15, Theorem 11.3].

Remark 1 (Practical considerations)**.** *Some cautions regarding the estimator* T_γ *need to be considered in practice. When* n *is insufficiently large, then* $\mathbf{B}_\gamma$ *might be different than* $\mathbf{B}$ *and Proposition 5 no longer holds. Specifically, let*

$\mathbf{S} \cap \mathbf{B}_\gamma$ *and* $\mathbf{B} \cap \mathbf{S}_\gamma$ *be the set of false positive and false negative errors, respectively, where* $\mathbf{S} = \{1, ..., p\} \setminus \mathbf{B}$ *and* $\mathbf{S}_\gamma = \{1, ..., p\} \setminus \mathbf{B}_\gamma$. *While false negatives merely result in not including some potential zero-estimator terms in our proposed estimator, false positives can lead to a substantial bias. This is true since the expected value of a post-selected zero-estimator is not necessarily zero anymore. A common approach to overcome this problem is to randomly split the data into two parts where the first part is used for covariate selection and the second part is used for evaluation of the zero-estimator terms.*

4.4. Estimating the variance of the proposed estimators

We now suggest estimators for $\mathrm{Var}(\hat{\tau}^2)$, $\mathrm{Var}(T_\gamma)$ and $\mathrm{Var}(T_{\hat{c}^*})$. Let

$$\widehat{\mathrm{Var}\left(\hat{\tau}^2\right)} = \frac{4}{n}\left[\frac{(n-2)}{(n-1)}\left[\hat{\sigma}_Y^2\hat{\tau}^2 + \hat{\tau}^4\right] + \frac{1}{2\,(n-1)}\left(p\hat{\sigma}_Y^4 + 4\hat{\sigma}_Y^2\hat{\tau}^2 + 3\hat{\tau}^4\right)\right],$$

where $\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2$, and $\hat{\sigma}_Y^4 = \left(\hat{\sigma}_Y^2\right)^2$. The following proposition shows that $\widehat{\mathrm{Var}\left(\hat{\tau}^2\right)}$ is consistent under some conditions.

Proposition 6. *Assume model* (1) *and additionally that* $\tau^2 + \sigma^2 = O(1)$, $X_i \overset{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I})$ *and* $p/n = O(1)$. *Then,*

$$n\left[\widehat{Var\left(\hat{\tau}^2\right)} - Var\left(\hat{\tau}^2\right)\right] \xrightarrow{p} 0.$$

Consider now $\mathrm{Var}(T_\gamma)$ and let $\widehat{\mathrm{Var}\left(T_\gamma\right)} = \widehat{\mathrm{Var}\left(\hat{\tau}^2\right)} - \frac{8}{n}\hat{\tau}^4_{\mathbf{B}_\gamma}$, where $\hat{\tau}^2_{\mathbf{B}_\gamma} = \sum_{j\in\mathbf{B}_\gamma}\hat{\beta}_j^2$ and $\hat{\tau}^4_{\mathbf{B}_\gamma} = \left(\hat{\tau}^2_{\mathbf{B}_\gamma}\right)^2$. The following propositions shows that $\widehat{\mathrm{Var}\left(T_\gamma\right)}$ is consistent.

Proposition 7. *Under the assumptions of Propositions 5 and 6,*

$$n\left[\widehat{Var\left(T_\gamma\right)} - Var\left(T_\gamma\right)\right] \xrightarrow{p} 0.$$

When normality of the covariates is not assumed, we suggest the following estimators:

$$\widetilde{\mathrm{Var}\left(\hat{\tau}^2\right)} = \frac{4\,(n-2)}{n\,(n-1)}\left[\widehat{\beta^T\mathbf{A}\beta} - \widehat{\|\beta\|^4}\right] + \frac{2}{n\,(n-1)}\left[\widehat{\|\mathbf{A}\|_F^2} - \widehat{\|\beta\|^4}\right];$$

$$\widetilde{\mathrm{Var}\left(T_\gamma\right)} = \widetilde{\mathrm{Var}\left(\hat{\tau}\right)} - \frac{4}{n}\left\{\sum_{j\in\mathbf{B}_\gamma}^{p}\hat{\beta}_j^4\left[E\left(X_{1j}^4\right) - 1\right] + 2\sum_{j\neq j'\in\mathbf{B}_\gamma}\hat{\beta}_j^2\hat{\beta}_{j'}^2\right\};$$

and

$$\widetilde{\mathrm{Var}\left(T_{\hat{c}^*}\right)} = \widetilde{\mathrm{Var}\left(\hat{\tau}^2\right)} - \frac{\left[\frac{2}{n(n-1)}\sum\limits_{i_1\neq i_2}\sum\limits_{j=1}^{p} W_{i_1 j}S_{i_2 j}\right]^2}{\mathrm{Var}\left(g_i\right)},$$

where

$$\widehat{\beta^T \mathbf{A} \beta} = \frac{1}{n\left(n-1\right)\left(n-2\right)} \sum_{i_1=i_2 \neq i_3} \mathbf{W}_{i_1} \left(\mathbf{W}_{i_2} \mathbf{W}_{i_2}^T\right) \mathbf{W}_{i_3},$$

$\widehat{\|\mathbf{A}\|_F^2} = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \left(\mathbf{W}_{i_1}^T \mathbf{W}_{i_2}\right)^2$, $\widehat{\|\beta\|^4} = (\frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \mathbf{W}_{i_1}^T \mathbf{W}_{i_2})^2$ are all U-statistics estimators, and $\hat{\beta}_j^2$ is given by (4). The proofs of Propositions 6 and 7 are given in the Appendix. We do not provide consistency proof for the estimators $\widehat{\mathrm{Var}(\hat{\tau}^2)}$, $\widehat{\mathrm{Var}(\hat{\tau}_\gamma^2)}$ and $\widehat{\mathrm{Var}(T_{\hat{c}^*})}$. However, our simulations support the consistency claim when the assumptions of Proposition 3 hold.

5. Simulations results

In this section, we illustrate the performance of the proposed estimators using a simulation study. Specifically, the following estimators are compared:

- The naive estimator $\hat{\tau}^2$, which is given in (5).
- The optimal oracle estimator T_{oracle}, which is given in (10).
- The estimator $T_{\hat{c}^*}$, which is based on adding a single zero-estimator and is given in (19).
- The estimator T_γ, which is based on selecting a small number of covariates and is given by Algorithm 1. Details about the specific selection algorithm we used can be found in Remark 8 in the Appendix.

The above estimators are compared to two additional estimators that were suggested previously:

- The PSI procedure (Post Selective Inference), which was calculated using the `estimateSigma` function from the `selectiveInference` R package [26]. The PSI procedure is based on the LASSO method which assumes sparsity of the coefficients and therefore ignores small coefficients [27].
- Ridge estimator is well-known technique for estimating the regression coefficient vector β. Since the parameter of interest here is τ^2 rather than β, we consider a plug-in ridge estimator which is constructed by taking the sum of squares of ridge regression estimated coefficients calculated by the `glmnet` R package [14].

It is noteworthy that unlike the PSI estimator, the ridge estimator does not require sparsity. However, since the goal of ridge regression is to estimate the coefficient vector β rather than τ^2, a naive ridge plug-in estimator of τ^2 is not expected to perform well, as shown in the simulations below.

We simulated data from the linear model (1). We fixed $\beta_j^2 = \frac{\tau_{\mathbf{B}}^2}{5}$ for $j = 1, \ldots, 5$, and $\beta_j^2 = \frac{\tau^2 - \tau_{\mathbf{B}}^2}{p-5}$ for $j = 6, \ldots, p$, where τ^2 and $\tau_{\mathbf{B}}^2$ vary among different scenarios. The covariates were generated from the centered exponential distribution, i.e., $X_{ij} \overset{iid}{\sim} \mathrm{Exp}(1) - 1$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. The noise ϵ was generated from the standard normal distribution. The number of observations and covariates is $n = p = 300$, and the residual variance σ^2 equals to 1. For

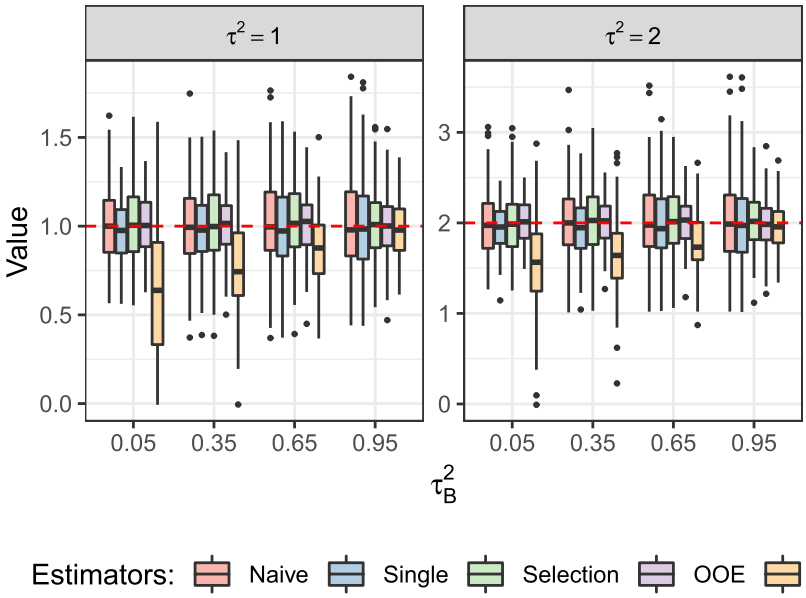

FIG 1. *Boxplots representing the estimators' distribution. The x-axis stands for* $\tau^2_{\mathbf{B}}$. *The red dashed is the true value of* τ^2.

each scenario, we generated 100 independent datasets and estimated τ^2 by using the different estimators. Boxplots of the estimates are plotted in Figure 1 and results of the RMSE are given in Table 1. Since the ridge-based estimator of τ^2 is highly biased, it is not comparable to the other proposed estimators and we omit it from the figures below. Code for reproducing the results is available at https://git.io/Jt6bC.

Table 1 shows the mean, the root mean square error (RMSE), and the relative improvement with respect to the naive $\hat{\tau}^2$ for the different estimators. Standard errors are given in parenthesis. Important points to notice:

- Both of the proposed estimators $T_{\hat{c}^*}$ and T_γ and the oracle estimator T_{oracle} improve the naive estimator in all scenarios. When $\tau^2 = 2$, these improvements are more substantial than for the case of $\tau^2 = 1$.
- The improved estimators are complementary to each other, i.e., for small values of $\tau^2_{\mathbf{B}}$ the Single estimator $T_{\hat{c}^*}$ performs better than the Selection estimator T_γ, and the opposite occurs for large values of $\tau^2_{\mathbf{B}}$. For example, when $\tau^2 = 1$ and $\tau^2_{\mathbf{B}} = 5\%$, the Single estimator $T_{\hat{c}^*}$ improves the naive estimator by 26% and when $\tau^2_{\mathbf{B}} = 95\%$, the Selection estimator T_γ improves the naive by 23%. This aligns with the result shown in Example 4.
- The PSI and ridge estimators perform poorly in a non-sparse setting. For example, when $\tau^2 = 1$ and $\tau^2_{\mathbf{B}} = 35\%$ their RMSE are larger than the RMSE of the naive estimator by 47% and 226%, respectively. Notice that the RMSE of the ridge estimator is large also under the sparse setting.

TABLE 1

Summary statistics for the proposed estimators; $n = p = 300$. Mean, root mean square error (RMSE) and percentage change from the naive estimator (in terms of RMSE) are shown. Simulation standard errors are shown in parenthesis. The table results were computed over 100 simulated datasets for each setting. An estimate for the standard error of RMSE was calculated using the delta method. The estimator with the lowest RMSE (excluding the oracle) is in bold.

τ_B^2	τ^2	Estimator	Mean	RMSE	% Change
5%	1	naive	1.01 (0.051)	0.226 (0.035)	0
5%	1	OOE	1.01 (0.037)	0.165 (0.023)	−26.99
5%	1	PSI	0.63 (0.08)	0.514 (0.064)	127.43
5%	1	Selection	1.01 (0.051)	0.225 (0.034)	−0.44
5%	1	**Single**	0.97 (0.037)	0.168 (0.024)	**−25.66**
5%	1	Ridge	0.21 (0.003)	0.789 (0.003)	213.1
35%	1	naive	1.02 (0.054)	0.242 (0.041)	0
35%	1	OOE	1.01 (0.039)	0.173 (0.027)	−28.51
35%	1	PSI	0.77 (0.061)	0.356 (0.052)	47.11
35%	1	Selection	1.02 (0.052)	0.231 (0.035)	−4.55
35%	1	**Single**	0.98 (0.044)	0.194 (0.033)	**−19.83**
35%	1	Ridge	0.18 (0.003)	0.825 (0.003)	226.09
65%	1	naive	1.02 (0.057)	0.256 (0.046)	0
65%	1	OOE	1.01 (0.042)	0.185 (0.03)	−27.73
65%	1	PSI	0.87 (0.047)	0.246 (0.036)	−3.91
65%	1	**Selection**	1.02 (0.05)	0.224 (0.034)	**−12.5**
65%	1	Single	1 (0.053)	0.235 (0.038)	−8.2
65%	1	Ridge	0.13 (0.002)	0.868 (0.002)	213.36
95%	1	naive	1.02 (0.062)	0.278 (0.048)	0
95%	1	OOE	1.01 (0.045)	0.202 (0.033)	−27.34
95%	1	**PSI**	0.98 (0.035)	0.157 (0.023)	**−43.53**
95%	1	Selection	1.02 (0.048)	0.214 (0.034)	−23.02
95%	1	Single	1.01 (0.063)	0.278 (0.047)	0
95%	1	Ridge	0.11 (0.001)	0.894 (0.001)	190.26
5%	2	naive	2.01 (0.039)	0.39 (0.027)	0
5%	2	OOE	2.01 (0.025)	0.245 (0.014)	−37.18
5%	2	PSI	1.54 (0.052)	0.692 (0.052)	77.44
5%	2	Selection	2.01 (0.039)	0.386 (0.027)	−1.03
5%	2	**Single**	1.94 (0.026)	0.264 (0.02)	**−32.31**
5%	2	Ridge	0.33 (0.004)	1.672 (0.004)	282.61
35%	2	naive	2.02 (0.043)	0.427 (0.035)	0
35%	2	OOE	2.01 (0.025)	0.254 (0.017)	−40.52
35%	2	PSI	1.66 (0.044)	0.554 (0.04)	29.74
35%	2	Selection	2.02 (0.04)	0.397 (0.027)	−7.03
35%	2	**Single**	1.96 (0.033)	0.326 (0.023)	**−23.65**
35%	2	Ridge	0.25 (0.004)	1.748 (0.004)	306.51
65%	2	naive	2.03 (0.046)	0.456 (0.04)	0
65%	2	OOE	2 (0.027)	0.272 (0.019)	−40.35
65%	2	PSI	1.78 (0.034)	0.403 (0.028)	−11.62
65%	2	**Selection**	2.03 (0.037)	0.372 (0.024)	**−18.42**
65%	2	Single	1.99 (0.041)	0.411 (0.03)	−9.87
65%	2	Ridge	0.18 (0.002)	1.819 (0.002)	287.02
95%	2	naive	2.04 (0.049)	0.494 (0.042)	0
95%	2	OOE	2 (0.03)	0.296 (0.022)	−40.08
95%	2	**PSI**	1.96 (0.025)	0.255 (0.018)	**−48.38**
95%	2	Selection	2.03 (0.033)	0.329 (0.024)	−33.4
95%	2	Single	2.01 (0.05)	0.494 (0.041)	0
95%	2	Ridge	0.15 (0.002)	1.855 (0.002)	246.08

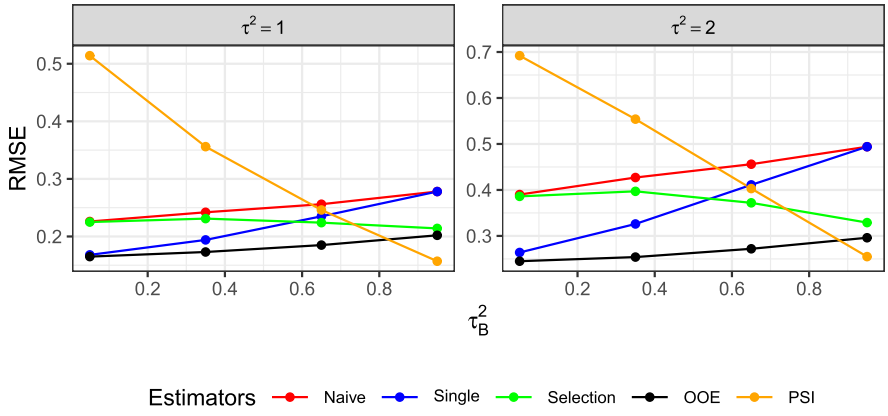

FIG 2. *Root mean square error (RMSE) for the different estimators. The x-axis stands for the sparsity level* $\tau^2_{\mathbf{B}}$.

TABLE 2
Summary statistics similar to Table 1; $n = p = 300$*;* $\tau^2_{\mathbf{B}} = 0.65$*;* $\tau^2 = 1$.

N	Estimator	Mean	RMSE	% Change
5000	naive	0.96 (0.025)	0.255 (0.021)	0
5000	OOE	0.98 (0.018)	0.179 (0.011)	−29.8
5000	Selection	0.99 (0.025)	0.247 (0.018)	−3.14
5000	Single	0.94 (0.023)	0.238 (0.021)	−6.67
10000	naive	0.98 (0.024)	0.242 (0.017)	0
10000	OOE	0.98 (0.02)	0.2 (0.013)	−17.36
10000	Selection	0.98 (0.022)	0.219 (0.013)	−9.5
10000	Single	0.95 (0.023)	0.232 (0.016)	−4.13
20000	naive	1.03 (0.027)	0.267 (0.015)	0
20000	OOE	1.03 (0.02)	0.201 (0.017)	−24.72
20000	Selection	1.01 (0.023)	0.233 (0.016)	−12.73
20000	Single	1.01 (0.026)	0.256 (0.015)	−4.12

Figure 2 plots the RMSE of each estimator as a function of the sparsity level $\tau^2_{\mathbf{B}}$ and the signal level τ^2. It is demonstrated that the Single and Selection estimators estimators improve (i.e., lower or equal RMSE) the naive estimator in all settings.

In the following we consider the case that the distribution of the covariates is only partially known. It is assumed that a large amount of unlabeled data is available and the distribution of the covariates is estimated based on this data. Specifically, we assume that additional sample of N i.i.d observations $X_{n+1}, ..., X_{n+N}$ are given while the responses $Y_{n+1}, ..., Y_{n+N}$ are not. Rather than treating $\mu \equiv E(X)$ and $\mathbf{\Sigma} \equiv \mathrm{Cov}(X)$ as known, we estimate these parameters by their plug-in estimators, $\hat{\mu} \equiv \frac{1}{N}\sum_{i=n+1}^{n+N} X_i$ and $\hat{\mathbf{\Sigma}} \equiv \frac{1}{N}\sum_{i=n+1}^{n+N}(X_i - \mathbf{1}\hat{\mu})(X_i - \mathbf{1}\hat{\mu})^T$, where $\mathbf{1} \equiv (1, \ldots, 1)^T$. We then apply the linear transformation, $X \mapsto \hat{\mathbf{\Sigma}}^{-1/2}(X - \hat{\mu})$, which corresponds the transformation shown in Section 2.1, and apply our estimators to the transformed X. We repeated the simulation study above for different values of N.

Table 2 is similar to Table 1 but includes different values of N rather than different values of $\tau^2_{\mathbf{B}}$. For simplicity we present only the scenario of $\tau^2 = 1$ and $\tau^2_{\mathbf{B}} = 0.35$ but the results for other scenarios are similar. It can be observed from the table that for large values of N the results are fairly similar to those in Table 1.

6. Generalization to other estimators

The suggested methodology in this paper is not limited to improving only the naive estimator, but can also be generalized to other estimators. The key is to add zero-estimators that are highly correlated with our initial estimator of τ^2; see Equation (9). Unlike the naive estimator, which is represented by a closed-form expression, other common estimators, such as the EigenPrism estimator [17], are computed numerically by solving a convex optimization problem. For a given zero-estimator, this makes the task of estimating the optimal-coefficient c^* more challenging than before. To overcome this challenge, we approximate the optimal coefficient c^* using bootstrap samples. We present a general algorithm that achieves improvement without claiming optimality. The algorithm is based on adding a single zero-estimator as in Section 4.2. A different version of the above algorithm, in which only a subset of covariates are used as for zero-estimators terms, was also used and is referred below to as the Selection estimator. See details in Remark 9 in the Appendix.

We illustrate the improvement obtained by Algorithm 2 by choosing $\tilde{\tau}^2$ to be the EigenPrism procedure [17], but other estimators can be used as well. We

Algorithm 2: Empirical Estimators

Input: A dataset $(\mathbf{X}, \mathbf{Y})$, an initial estimator $\tilde{\tau}^2$.

1. Calculate an initial estimator $\tilde{\tau}^2$ of τ^2.
2. **Bootstrap step:**
   - Resample with replacement n observations from $(\mathbf{X}, \mathbf{Y})$.
   - Calculate the initial estimator $\tilde{\tau}^2$ of τ^2.
   - Calculate the zero-estimator $g_n = \frac{1}{n}\sum_{i=1}^{n} g_i$ where $g_i = \sum_{j<j'} X_{ij}X_{ij'}$.

   This procedure is repeated B times in order to produce $(\tilde{\tau}^2)^{*1}, ..., (\tilde{\tau}^2)^{*B}$ and $g_n^{*1}, ..., g_n^{*B}$.
3. Approximate the coefficient c^* by

$$\tilde{c}^* = \frac{\widehat{\mathrm{Cov}\,(\tilde{\tau}^2, g_n)}}{\mathrm{Var}\,(g_n)},$$

   where $\widehat{\mathrm{Cov}\,(\cdot)}$ denotes the empirical covariance from the bootstrap samples, and $\mathrm{Var}(g_n)$ is known by the semi-supervised setting.

Result: Return the empirical estimator $T_{emp} = \tilde{\tau}^2 - \tilde{c}^* g_n$.

TABLE 3
Summary statistics similar to Table 1.

$\tau_{\mathbf{B}}^2$	τ^2	Estimator	Mean	RMSE	% Change
5%	1	Eigenprism	0.98 (0.019)	0.195 (0.012)	0
5%	1	Single	0.98 (0.018)	0.183 (0.012)	−6.15
5%	1	Selection	0.98 (0.02)	0.195 (0.013)	0
35%	1	Eigenprism	0.99 (0.02)	0.198 (0.013)	0
35%	1	Single	0.99 (0.019)	0.193 (0.013)	−2.53
35%	1	Selection	0.99 (0.02)	0.197 (0.013)	−0.51
65%	1	Eigenprism	1 (0.021)	0.206 (0.013)	0
65%	1	Single	1 (0.021)	0.205 (0.013)	−0.49
65%	1	Selection	1 (0.02)	0.199 (0.013)	−3.4
95%	1	Eigenprism	1.01 (0.022)	0.215 (0.014)	0
95%	1	Single	1.01 (0.022)	0.215 (0.014)	0
95%	1	Selection	1.01 (0.02)	0.201 (0.013)	−6.51
5%	2	EigenPrism	2.05 (0.029)	0.292 (0.018)	0
5%	2	Single	2.03 (0.026)	0.262 (0.016)	−10.27
5%	2	Selection	2.05 (0.029)	0.292 (0.017)	0
35%	2	EigenPrism	2.02 (0.029)	0.287 (0.02)	0
35%	2	Single	2.01 (0.027)	0.272 (0.02)	−5.23
35%	2	Selection	2.01 (0.028)	0.281 (0.02)	−2.09
65%	2	EigenPrism	2.01 (0.03)	0.296 (0.023)	0
65%	2	Single	2 (0.029)	0.292 (0.023)	−1.35
65%	2	Selection	1.99 (0.028)	0.28 (0.021)	−5.41
95%	2	EigenPrism	1.99 (0.031)	0.31 (0.025)	0
95%	2	Single	1.99 (0.031)	0.31 (0.025)	0
95%	2	Selection	1.97 (0.028)	0.279 (0.021)	−10

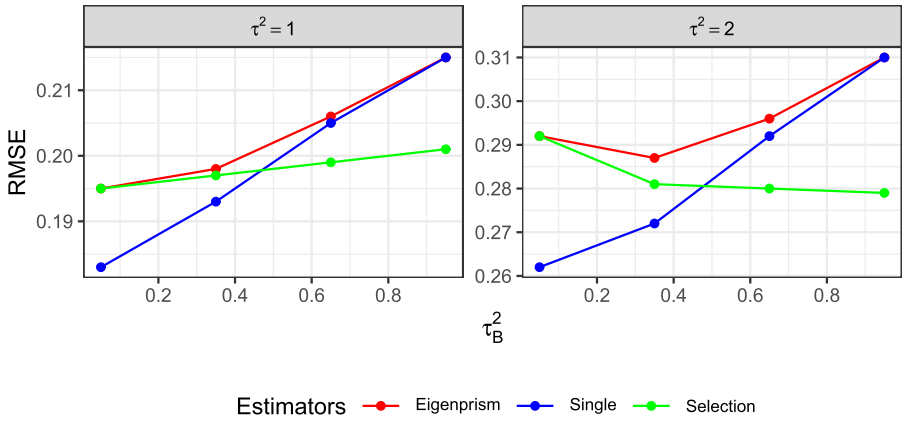

FIG 3. *Root mean square error (RMSE) for the proposed estimators. The x-axis stands for the sparsity level* $\tau_{\mathbf{B}}^2$.

consider the same setting as in Section 5. The number of bootstrap samples is $M = 100$. Results are given in Table 3 and the code for reproducing the results is available at https://git.io/Jt6bC.

The simulation results appear in Table 3 and in Figure 3. Both empirical estimators show an improvement over the EigenPrism estimator $\tilde{\tau}^2$. The results

here are fairly similar to the results shown for the naive estimator in Section 5, with just a smaller degree of improvement. As before, the Single and Selection estimators are complementary to each other, i.e., for small values of $\tau^2_{\mathbf{B}}$ the Single estimator performs better than the Selection estimator and the opposite occurs for large values of $\tau^2_{\mathbf{B}}$. This highlights the fact that the zero-estimator approach is not limited to improving only the naive estimator but rather has the potential to improve other estimators as well.

7. Discussion and future work

This paper presents a new approach for improving estimation of the explained variance τ^2 of a high-dimensional regression model in a semi-supervised setting without assuming sparsity. The key idea is to use a zero-estimator that is correlated with the initial unbiased estimator of τ^2 in order to lower its variance without introducing additional bias. The semi-supervised setting, where the number of observations is much greater than the number of responses, allows us to construct such zero-estimators. We introduced a new notion of optimality with respect to zero-estimators and presented an oracle-estimator that achieves this type of optimality. We proposed two different (non-oracle) estimators that showed a significant reduction, but not optimal, in the asymptotic variance of the naive estimator. Our simulations showed that our approach can be generalized to other types of initial estimators other than the naive estimator.

Many open questions remain for future research. While our proposed estimators improved the naive estimator, it did not achieve the optimal improvement of the oracle estimator. Thus, it remains unclear if and how one can achieve optimal improvement. Moreover, in this work, strong assumption was made about the unsupervised data size, i.e., $N = \infty$. Thus, generalizing the suggested approach by relaxing this assumption to allow for a more general setting with a *finite* $N \gg n$ is a natural direction for future work. A more ambitious future goal would be to extend the suggested approach to generalized linear models (GLM), and specifically to logistic regression. In this case, the concepts of signal and noise levels are less clear and are more challenging to define.

Appendix

Proof of Lemma 1.
Notice that $\mathbf{X}^T\mathbf{Y} = \left(\sum_{i=1}^n W_{i1}, ..., \sum_{i=1}^n W_{ip}\right)^T$ where $\mathbf{X}$ is the $n \times p$ design matrix and $\mathbf{Y} = (Y_1, ..., Y_n)^T$. Thus, the naive estimator can be also written as

$$\hat{\tau}^2 = \frac{1}{n\,(n-1)} \sum_{i_1 \neq i_2} \sum_{j=1}^{p} W_{i_1 j} W_{i_2 j} = \frac{\left\|\mathbf{X}^T\mathbf{Y}\right\|^2 - \sum_{j=1}^{p} \sum_{i=1}^{n} W_{ij}^2}{n\,(n-1)}.$$

The Dicker estimate for τ^2 ia given by $\hat{\tau}^2_{Dicker} \equiv \frac{\left\|\mathbf{X}^T\mathbf{Y}\right\|^2 - p\|\mathbf{Y}\|^2}{n(n+1)}$. We need to prove that root-n times the difference between the estimators converges in

probability to zero, i.e., $\sqrt{n}\left(\hat{\tau}^2_{Dicker} - \hat{\tau}^2\right) \xrightarrow{p} 0$. We have,

$$\begin{aligned}\sqrt{n}\left(\hat{\tau}^2_{Dicker} - \hat{\tau}^2\right) &= \sqrt{n}\left(\frac{\left\|\mathbf{X}^T\mathbf{Y}\right\|^2 - p\|\mathbf{Y}\|^2}{n(n+1)} - \frac{\left\|\mathbf{X}^T\mathbf{Y}\right\|^2 - \sum_{j=1}^{p}\sum_{i=1}^{n} W_{ij}^2}{n(n-1)}\right) \\ &= \sqrt{n}\left(\frac{\sum_{j=1}^{p}\sum_{i=1}^{n} W_{ij}^2}{n(n-1)} - \frac{p\|\mathbf{Y}\|^2}{n(n+1)} - \frac{2\left\|\mathbf{X}^T\mathbf{Y}\right\|^2}{n(n-1)(n+1)}\right).\end{aligned} \tag{23}$$

It is enough to prove that:

1. $n^{-1.5}(\sum_{j=1}^{p}\sum_{i=1}^{n} W_{ij}^2 - p\|\mathbf{Y}\|^2) \xrightarrow{p} 0$,
2. $n^{-2.5}(\|\mathbf{X}^T\mathbf{Y}\|^2) \xrightarrow{p} 0$.

We start with the first term,

$$\begin{aligned}n^{-1.5}\left(\sum_{j=1}^{p}\sum_{i=1}^{n} W_{ij}^2 - p\|\mathbf{Y}\|^2\right) &= n^{-1.5}\left(\sum_{j=1}^{p}\sum_{i=1}^{n} Y_i^2 X_{ij}^2 - p\sum_{i=1}^{n} Y_i^2\right) \\ = n^{-1.5}\left(\sum_{i=1}^{n} Y_i^2 \sum_{j=1}^{p} X_{ij}^2 - p\sum_{i=1}^{n} Y_i^2\right) &= n^{-0.5}\sum_{i=1}^{n} Y_i^2\left[\frac{1}{n}\sum_{j=1}^{p}\left(X_{ij}^2 - 1\right)\right] \\ &\equiv n^{-0.5}\sum_{i=1}^{n}\omega_i\end{aligned} \tag{24}$$

where $\omega_i = Y_i^2\left[\frac{1}{n}\sum_j\left\{X_{ij}^2 - 1\right\}\right]$. Notice that ω_i depends on n but this is suppressed in the notation. In order to show that $n^{-0.5}\sum_{i=1}^{n}\omega_i \xrightarrow{p} 0$, it is enough to show that $E\left(n^{-0.5}\sum_{i=1}^{n}\omega_i\right) \to 0$ and $\mathrm{Var}\left(n^{-0.5}\sum_{i=1}^{n}\omega_i\right) \to 0$. Moreover, since $E\left(n^{-0.5}\sum_{i=1}^{n}\omega_i\right) = \sqrt{n}E(\omega_i)$ and $\mathrm{Var}\left(n^{-0.5}\sum_{i=1}^{n}\omega_i\right) = \mathrm{Var}(\omega_i) = E\left(\omega_i^2\right) - [E(\omega_i)]^2$, it is enough to show that $\sqrt{n}E(\omega_i)$ and $E\left(\omega_i^2\right)$ converge to zero.

Consider now $\sqrt{n}E(\omega_i)$. By (24) we have

$$\sum_{i=1}^{n}\omega_i = \frac{1}{n}\left[\sum_{j=1}^{p}\sum_{i=1}^{n} W_{ij}^2 - p\|\mathbf{Y}\|^2\right].$$

Taking expectation of both sides,

$$\sum_{i=1}^{n} E(\omega_i) = \frac{1}{n}\left[\sum_{j=1}^{p}\sum_{i=1}^{n} E\left(W_{ij}^2\right) - pE\left(\|\mathbf{Y}\|^2\right)\right].$$

Now, notice that

$$E(W_{ij}^2) = E[X_{ij}^2(\beta^T X + \epsilon)^2] = \|\beta\|^2 + \sigma^2 + \beta_j^2[E(X_{ij}^4) - 1] = \tau^2 + \sigma^2 + 2\beta_j^2. \tag{25}$$

Also notice that $Y_i^2/\left(\sigma_\varepsilon^2 + \tau^2\right) \sim \chi_1^2$, and hence $E\left(\|\mathbf{Y}\|^2\right) = n\left(\tau^2 + \sigma^2\right)$. Therefore,

$$\begin{aligned} nE\left(\omega_i\right) &= \frac{1}{n}\left[\sum_{j=1}^{p}\sum_{i=1}^{n}\left(\tau^2 + \sigma^2 + 2\beta_j^2\right) - pn\left(\tau^2 + \sigma^2\right)\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}\left[p\left(\tau^2 + \sigma^2\right) + 2\tau^2\right] - pn\left(\tau^2 + \sigma^2\right)\right] = 2\tau^2 \end{aligned}$$

which implies that $\sqrt{n}E\left(\omega_i\right) = \frac{2\tau^2}{\sqrt{n}} \to 0$.

Consider now $E(\omega_i^2)$. By Cauchy-Schwartz,

$$\begin{aligned} E\left(\omega_i^2\right) &= E\left(Y_i^4\left[n^{-1}\sum_{j=1}^{p}\left\{X_{ij}^2 - 1\right\}\right]^2\right) \\ &\leq \left\{E\left(Y_i^8\right)\right\}^{1/2}\left\{E\left(\left[n^{-1}\sum_{j=1}^{p}\left\{X_{ij}^2 - 1\right\}\right]^4\right)\right\}^{1/2}. \end{aligned}$$

Notice that $Y_i \sim N\left(0, \tau^2 + \sigma^2\right)$ by construction and therefore $E(Y_i^8) = O(1)$ as n and p go to infinity. Let $V_j = X_{ij}^2 - 1$ and notice that $E(V_j) = 0$. We have

$$\begin{aligned} E\left(\left[n^{-1}\sum_{j=1}^{p}\left\{X_{ij}^2 - 1\right\}\right]^4\right) &= E\left(\left[n^{-1}\sum_{j=1}^{p}V_j\right]^4\right) \\ &= n^{-4}\sum_{j_1,j_2,j_3,j_4} E\left(V_{j_1}V_{j_2}V_{j_3}V_{j_4}\right). \end{aligned}$$

The expectation $\sum_{j_1,j_2,j_3,j_4} E\left(V_{j_1}V_{j_2}V_{j_3}V_{j_4}\right)$ is not 0 when $j_1 = j_2$ and $j_3 = j_4$ (up to permutations) or when all terms are equal. In the first case we have

$$\sum_{j\neq j'} E\left(V_j^2V_{j'}^2\right) = \sum_{j\neq j'}\left[E\left(V_j^2\right)\right]^2 = p\left(p-1\right)\left[E\left\{\left(X_{ij}^2 - 1\right)^2\right\}\right]^2 \leq C_1p^2,$$

for a positive constant C_1. In the second case we have

$$\sum_{j=1}^{p} E\left(V_j^4\right) = pE\left[\left(X_{ij}^2 - 1\right)^4\right] \leq C_2p,$$

for a positive constant C_2. Hence, as p and n have the same order of magnitude, we have

$$\left\{E\left[\left(n^{-1}\sum_{j=1}^{p}\left\{X_{ij}^2-1\right\}\right)^4\right]\right\}^{1/2}=\left\{n^{-4}\sum_{j_1,j_2,j_3,j_4}E\left(V_{j_1}V_{j_2}V_{j_3}V_{j_4}\right)\right\}^{1/2}$$
$$\leq\left\{n^{-4}\cdot O\left(p^2\right)\right\}^{1/2}\leq K/n,$$

which implies $E(\omega_i^2)\leq K_1/n\to 0$, where K and K_1 are positive constants. This completes the proof that

$$n^{-1.5}\left(\sum_{j=1}^{p}\sum_{i=1}^{n}W_{ij}^2-p\|\mathbf{Y}\|^2\right)\xrightarrow{p}\ 0.$$

We now move to prove that $n^{-2.5}\left(\left\|\mathbf{X}^T\mathbf{Y}\right\|^2\right)\xrightarrow{p}0$. By Markov's inequality, for $\epsilon>0$

$$P\left(n^{-2.5}\left\|\mathbf{X}^T\mathbf{Y}\right\|^2>\varepsilon\right)\leq n^{-2.5}E\left(\left\|\mathbf{X}^T\mathbf{Y}\right\|^2\right)/\varepsilon.$$

Thus, it is enough to show that $n^{-2}E\left(\left\|\mathbf{X}^T\mathbf{Y}\right\|^2\right)$ is bounded. Notice that

$$\begin{aligned}E\left(\left\|\mathbf{X}^T\mathbf{Y}\right\|^2\right)&=\sum_{i_1,i_2}\sum_{j=1}^{p}E\left(W_{i_1j}W_{i_2j}\right)=\sum_{i=1}^{n}\sum_{j=1}^{p}E\left(W_{ij}^2\right)\\&+\sum_{i_1\neq i_2}\sum_{j=1}^{p}E\left(W_{i_1j}W_{i_2j}\right)=\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\tau^2+\sigma^2+2\beta_j^2\right)+\sum_{i_1\neq i_2}\sum_{j=1}^{p}\beta_j^2\\&=n\left[p\left(\tau^2+\sigma^2\right)+2\tau^2\right]+n\left(n-1\right)\tau^2\\&=n\left[p\left(\tau^2+\sigma^2\right)+\left(n+1\right)\tau^2\right],\end{aligned}$$

where we used (25) in the third equality. Therefore,

$$n^{-2}E\left(\left\|\mathbf{X}^T\mathbf{Y}\right\|^2\right)=n^{-1}\left[p\left(\tau^2+\sigma^2\right)+\left(n+1\right)\tau^2\right].$$

Since p and n have the same order of magnitude and $\tau^2+\sigma^2$ is bounded by assumption, then $n^{-2}E\left(\left\|\mathbf{X}^T\mathbf{Y}\right\|^2\right)$ is also bounded. This completes the proof of $n^{-2.5}\left(\left\|\mathbf{X}^T\mathbf{Y}\right\|^2\right)\xrightarrow{p}0$ and hence $\sqrt{n}\left(\hat{\tau}_{Dicker}^2-\hat{\tau}^2\right)\xrightarrow{p}\ 0$. □

Proof of Corollary 1.
According to Corollary 1 in [9], we have

$$\frac{\sqrt{n}\left(\hat{\tau}_{Dicker}-\tau^2\right)}{\psi}\xrightarrow{D}N\left(0,1\right),$$

where $\psi = 2\left\{\left(1+\frac{p}{n}\right)\left(\sigma^2+\tau^2\right)^2 - \sigma^4 + 3\tau^4\right\}$, given that p/n converges to a constant. Therefore we can write

$$\frac{\sqrt{n}\left(\hat{\tau}^2-\tau^2\right)}{\psi} = \frac{1}{\psi}\left[\sqrt{n}\left(\hat{\tau}^2-\hat{\tau}_{Dicker}\right)+\sqrt{n}\left(\hat{\tau}_{Dicker}-\tau^2\right)\right],$$

and obtain $\sqrt{n}\left(\frac{\hat{\tau}^2-\tau^2}{\psi}\right) \xrightarrow{D} N(0,1)$ by Slutsky's theorem. □

Proof of Proposition 1.
Let $\mathbf{W}_i = (W_{i1},...,W_{ip})^T$ and notice that $\hat{\tau}^2 = \frac{1}{n(n-1)}\sum_{i_1\neq i_2}^{n}\sum_{j=1}^{p} W_{i_1 j}W_{i_2 j}$ is a U-statistic of order 2 with the kernel $h(\mathbf{w}_1,\mathbf{w}_2) = \mathbf{w}_1^T\mathbf{w}_2 = \sum_{j=1}^{p} w_{1j}w_{2j}$, where $\mathbf{w}_i \in \mathbb{R}^p$.

By Theorem 12.3 in [29],

$$\mathrm{Var}\left(\hat{\tau}^2\right) = \frac{4(n-2)}{n(n-1)}\zeta_1 + \frac{2}{n(n-1)}\zeta_2, \tag{26}$$

where

$$\zeta_1 = \mathrm{Cov}\left[h(\mathbf{W}_1,\mathbf{W}_2), h\left(\mathbf{W}_1,\widetilde{\mathbf{W}}_2\right)\right]$$

and

$$\zeta_2 = \mathrm{Cov}\left[h(\mathbf{W}_1,\mathbf{W}_2), h(\mathbf{W}_1,\mathbf{W}_2)\right];$$

where $\widetilde{\mathbf{W}}_2$ is an independent copy of $\mathbf{W_2}$. Now, let $\mathbf{A} = E\left(\mathbf{W}_i\mathbf{W}_i^T\right)$ be a $p\times p$ matrix and notice that

$$\begin{aligned}\zeta_1 &= \mathrm{Cov}\left[h(\mathbf{W}_1,\mathbf{W}_2), h\left(\mathbf{W}_1,\widetilde{\mathbf{W}}_2\right)\right]\\ &= \sum_{j,j'}^{p}\mathrm{Cov}\left(W_{1j}W_{2j}, W_{1j'}\widetilde{W}_{2j'}\right) = \sum_{j,j'}^{p}\left(\beta_j\beta_{j'}E\left[W_{1j}W_{1j'}\right] - \beta_j^2\beta_{j'}^2\right)\\ &= \beta^T\mathbf{A}\beta - \|\beta\|^4\end{aligned}$$

and

$$\begin{aligned}\zeta_2 &= \mathrm{Cov}\left[h(\mathbf{W}_1,\mathbf{W}_2), h(\mathbf{W}_1,\mathbf{W}_2)\right]\\ &= \sum_{j,j'}\mathrm{Cov}\left(W_{1j}W_{2j}, W_{1j'}W_{2j'}\right) = \sum_{j,j'}\left(\left(E\left[W_{1j}W_{1j'}\right]\right)^2 - \beta_j^2\beta_{j'}^2\right)\\ &= \|\mathbf{A}\|_F^2 - \|\beta\|^4,\end{aligned}$$

where $\|\mathbf{A}\|_F^2$ is the Frobenius norm of $\mathbf{A}$. Thus, by rewriting (26) the variance of the naive estimator is given by

$$\mathrm{Var}\left(\hat{\tau}^2\right) = \frac{4(n-2)}{n(n-1)}\left[\beta^T\mathbf{A}\beta - \|\beta\|^4\right] + \frac{2}{n(n-1)}\left[\|\mathbf{A}\|_F^2 - \|\beta\|^4\right]. \tag{27}$$

□

Proof of Proposition 2.
Notice that $\hat{\tau}^2$ is consistent if $\mathrm{Var}[\hat{\tau}^2] \xrightarrow{n\to\infty} 0$ since $\hat{\tau}^2$ is unbiased. Thus, by (27) it is enough to require that $\frac{\beta^T\mathbf{A}\beta}{n} \xrightarrow{n\to\infty} 0$ and $\frac{\|\mathbf{A}\|_F^2}{n^2} \xrightarrow{n\to\infty} 0$. The latter is assumed and we now show that the former also holds true. Let $\lambda_1 \geq \ldots \geq \lambda_p$ be the eigenvalues of $\mathbf{A}$ and notice that $\mathbf{A}$ is symmetric. We have that $n^{-2}\lambda_1^2 \leq n^{-2}\sum_{j=1}^p \lambda_j^2 = n^{-2}tr(\mathbf{A}^2) = n^{-2}\|\mathbf{A}\|_F^2$ and therefore (iii) implies that $\frac{\lambda_1}{n} \xrightarrow{n\to\infty} 0$. Now, $\frac{1}{n}\beta^T\mathbf{A}\beta \equiv \frac{1}{n}\|\beta\|^2[(\frac{\beta}{\|\beta\|})^T\mathbf{A}\frac{\beta}{\|\beta\|}] \leq \frac{1}{n}\|\beta\|^2\lambda_1 \xrightarrow{n\to\infty} 0$, where the last limit follows from the assumption that $\tau^2 = O(1)$, and from the fact that $\frac{\lambda_1}{n} \xrightarrow{n\to\infty} 0$. We conclude that $\mathrm{Var}[\hat{\tau}^2] \xrightarrow{n\to\infty} 0$.

We now prove the moreover part, that is, independence of the columns of $\mathbf{X}$ implies that $\frac{\|\mathbf{A}\|_F^2}{n^2} \xrightarrow{n\to\infty} 0$. By definition we have $\|\mathbf{A}\|_F^2 = \sum_{j,j'}[E(W_{ij}W_{ij'})]^2$. Notice that when $j = j'$ we have,

$$
\begin{aligned}
E\left(W_{ij}^2\right) &= E\left(X_{ij}^2 Y_i^2\right) = E\left(X_{ij}^2\left[\beta^T X_i + \varepsilon_i\right]^2\right) \\
&= E\left(X_{ij}^2\left[\sum_{k,k'}\beta_k\beta_{k'}X_{ik}X_{ik'} + 2\beta^T X_i\varepsilon_i + \varepsilon_i^2\right]\right) \\
&= E\left(X_{ij}^2\sum_{k,k'}\beta_k\beta_{k'}X_{ik}X_{ik'}\right) + 0 + E\left(X_{ij}^2\varepsilon_i^2\right) \\
&= E\left(X_{ij}^2\sum_{k=1}^p\beta_k^2X_{ik}^2\right) + \underbrace{E\left(X_{ij}^2\sum_{k\neq k'}\beta_k\beta_{k'}X_{ik}X_{ik'}\right)}_{0} + \sigma^2E\left(X_{ij}^2\right) \\
&= \beta_j^2E\left(X_{ij}^4\right) + \sum_{k\neq j}^p\beta_k^2\underbrace{E\left(X_{ik}^2X_{ij}^2\right)}_{1} + \sigma^2 \\
&= \beta_j^2E\left(X_{ij}^4\right) + \|\beta\|^2 - \beta_j^2 + \sigma^2 = \|\beta\|^2 + \sigma^2 + \beta_j^2\left[E\left(X_{ij}^4 - 1\right)\right].
\end{aligned}
$$

Notice that $E\left(X_{ij}^2\sum_{k\neq k'}\beta_k\beta_{k'}X_{ik}X_{ik'}\right) = 0$ follows from the assumptions that the columns of $\mathbf{X}$ are independent and $E(X_{ij}) = 0$ for each j. Also notice that in the third row we used the assumption that $E(\epsilon_i^2|X_i) = \sigma^2$.

Similarly, when $j \neq j'$,

$$
\begin{aligned}
E\left(W_{ij}W_{ij}\right) &= E\left(X_{ij}X_{ij'}Y_i^2\right) = E\left[X_{ij}X_{ij'}\left(\beta^T X_i + \varepsilon_i\right)^2\right] \\
&= E\left[X_{ij}X_{ij'}\left(\beta^T X_i + \varepsilon_i\right)^2\right] \\
&= E\left[X_{ij}X_{ij'}\left(\sum_{k,k'}\beta_k\beta_{k'}X_{ik}X_{ik'} + 2\beta^T X_i\varepsilon_i + \varepsilon_i^2\right)\right] \\
&= E\left[X_{ij}X_{ij'}\sum_{k,k'}\beta_k\beta_{k'}X_{ik}X_{ik'}\right] + 0 + E\left(X_{ij}X_{ij'}\varepsilon_i^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&= 2\beta_j\beta_{j'}E\left(X_{ij}^2X_{ij'}^2\right) + 0 + \underbrace{E\left(X_{ij}X_{ij'}\right)}_{0}E\left(\varepsilon_i^2\right)\\
&= 2\beta_j\beta_{j'}E\left(X_{ij}^2\right)E\left(X_{ij'}^2\right)\\
&= 2\beta_j\beta_{j'}.
\end{aligned}
$$

This can be written more compactly as

$$
E(W_{ij}W_{ij'}) = \begin{cases} 2\beta_j\beta_{j'}, & j \neq j' \\ \sigma_Y^2 + \beta_j^2[E(X_{ij}^4) - 1], & j = j', \end{cases} \tag{28}
$$

where $\sigma_Y^2 = \|\beta\|^2 + \sigma^2$. Therefore,

$$
\begin{aligned}
\|\mathbf{A}\|_F^2 &= 4\sum_{j\neq j'}\beta_j^2\beta_{j'}^2 + \sum_j\left(\sigma_Y^2 + \beta_j^2[E(X_{ij}^4) - 1]\right)^2\\
&\leq 4\|\beta\|^4 + \sum_j\left(\sigma_Y^4 + \beta_j^4[E(X_{ij}^4) - 1]^2 + 2\sigma_Y^2\beta_j^2[E(X_{ij}^4) - 1]\right)\\
&= p\sigma_Y^4 + O(1). \qquad (29)
\end{aligned}
$$

where the last equality holds since $\sigma_Y^2 \equiv \tau^2 + \sigma^2 = O(1)$, $E(X_{ij}^4) = O(1)$ and by the Cauchy–Schwarz inequality we have $\sum_j \beta_j^4 \leq \sum_{j,j'}\beta_j^2\beta_{j'}^2 = \|\beta\|^4 = O(1)$. Now since $p/n = O(1)$ then $\frac{\|\mathbf{A}\|_F^2}{n^2} \to 0$ and we conclude that $\mathrm{Var}(\hat{\tau}^2) = O(\frac{1}{n})$, i.e., $\hat{\tau}^2$ is $\sqrt{n}$-consistent. □

Remark 2. ***Calculations for Example 1:***

$$
\begin{aligned}
Cov[\hat{\tau}^2, g(X)] &\equiv Cov\left(\frac{2}{n\,(n-1)}\sum_{i_1<i_2}W_{i_1}W_{i_2}, \frac{1}{n}\sum_{i=1}^n[X_i^2 - 1]\right)\\
&= \frac{2}{n^2\,(n-1)}\sum_{i_1<i_2}\sum_{i=1}^n Cov\left(X_{i_1}Y_{i_1}X_{i_2}Y_{i_2}, X_i^2\right)\\
&= \frac{2}{n^2\,(n-1)}\sum_{i_1<i_2}\sum_{i=1}^n[E\left(X_{i_1}Y_{i_1}X_{i_2}Y_{i_2}X_i^2\right) - \beta^2]\\
&= \frac{4}{n^2\,(n-1)}\sum_{i_1<i_2}[E\left(X_{i_1}^3Y_{i_1}\right)\beta - \beta^2] \qquad (30)\\
&= \frac{4\beta}{n^2\,(n-1)}\sum_{i_1<i_2}[E\left(X_{i_1}^3Y_{i_1}\right) - \beta]\\
&= \frac{4\beta}{n^2\,(n-1)}\frac{n\,(n-1)}{2}[E\left(X_{i_1}^3Y_{i_1}\right) - \beta]\\
&= \frac{2\beta}{n}[E\left(X^3Y\right) - \beta],
\end{aligned}
$$

where in the third equality we used $E(X^2) = 1$ *and* $E(XY) \equiv \beta$. *In the fourth equality the expectation is zero for all* $i \neq i_1, i_2$. *Now, since* $X \sim N(0,1)$ *and* $E(\epsilon|X) = 0$, *then*

$$E\left(X^3 Y\right) = E\left(X^3\left(\beta X + \varepsilon\right)\right) = \beta E\left(X^4\right) = 3\beta.$$

Therefore, $Cov[\hat{\beta}^2, g(X)] = \frac{4\beta^2}{n}$. *Notice that*

$$Var[g] = \text{var}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 1\right)\right] = \frac{1}{n}\left[E\left(X^4\right) - E\left(X^2\right)\right] = \frac{2}{n}.$$

Therefore, by (8) *we get* $c^* = -2\beta^2$. *Plugging-in* c^* *back in* (9) *yields*

$$Var(U_{c^*}) = Var(\hat{\tau}^2) - \frac{8}{n}\beta^4.$$

Proof of Theorem 1.

1. We now prove the first direction: OOE $\Rightarrow \text{Cov}[R^*, g] = 0$ for all $g \in \mathcal{G}$.

Let $R^* \equiv T + g^*$ be an OOE for θ with respect to the family of zero-estimators $\mathcal{G}$. By definition, $\text{Var}[R^*] \leqslant \text{Var}[T + g]$ for all $g \in \mathcal{G}$. For every $g = \sum_{k=1}^{m} c_k g_k$, define $\tilde{g} \equiv g - g^* = \sum_{k=1}^{m}(c_k - c_k^*)g_k = \sum_{k=1}^{m} \tilde{c}_k g_k$ for some fixed m, and note that $\tilde{g} \in \mathcal{G}$. Then,

$$\begin{aligned}\text{Var}[R^*] \leqslant \text{Var}[T + g] = \text{Var}[T + g^* + \tilde{g}] &= \text{Var}\left[R^* + \sum_{k=1}^{m} \tilde{c}_k g_k\right] \\ &= \text{Var}[R^*] + 2\sum_{k=1}^{m} \tilde{c}_k \cdot \text{Cov}[R^*, g_k] + \text{Var}\left[\sum_{k=1}^{m} \tilde{c}_k g_k\right].\end{aligned}$$

Therefore, for all $(\tilde{c}_1, ..., \tilde{c}_m)$,

$$0 \leqslant 2\sum_{k=1}^{m} \tilde{c_k} \cdot \text{Cov}[R^*, g_k] + \text{Var}\left[\sum_{k=1}^{m} \tilde{c_k} g_k\right],$$

which can be represented compactly as

$$0 \leqslant -2\tilde{\mathbf{c}}^T \mathbf{b} + \text{Var}[\tilde{\mathbf{c}}^T \mathbf{g_m}] = -2\tilde{\mathbf{c}}^T \mathbf{b} + \tilde{\mathbf{c}}^T M \tilde{\mathbf{c}} \equiv f(\tilde{\mathbf{c}}), \tag{31}$$

where $\mathbf{b} \equiv -\left(\text{Cov}[R^*, g_1], ..., \text{Cov}[R^*, g_m]\right)^T$, $\mathbf{g_m} \equiv (g_1, ..., g_m)^T$, $M = \text{Cov}[\mathbf{g_m}]$ and $\tilde{\mathbf{c}} \equiv (\tilde{c}_1, ..., \tilde{c}_m)^T$. Notice that $f(\tilde{\mathbf{c}})$ is a convex function in $\tilde{\mathbf{c}}$ that satisfies $f(\tilde{\mathbf{c}}) \geq 0$ for all $\tilde{\mathbf{c}}$. Differentiate $f(\tilde{\mathbf{c}})$ in order to find its minimum

$$\nabla f(\tilde{\mathbf{c}}) = -2\mathbf{b} + 2M\tilde{\mathbf{c}} = 0.$$

Assuming M is positive definite and solving for $\tilde{\mathbf{c}}$ yields the minimizer

$$\tilde{\mathbf{c}}_{\min} = M^{-1}\mathbf{b}.$$

Plug-in $\tilde{\mathbf{c}}_{\min}$ in the (31) yields

$$f(\tilde{\mathbf{c}}_{\min}) \equiv -2(M^{-1}\mathbf{b})^T\mathbf{b} + (M^{-1}\mathbf{b})^T M(M^{-1}\mathbf{b}) = -\mathbf{b}^T M^{-1}\mathbf{b} \geq 0. \tag{32}$$

Since, by assumption, M is positive definite, so is M^{-1}, i.e.,

$$\mathbf{b}^T M^{-1}\mathbf{b} > 0$$

for all non-zero $\mathbf{b} \in \mathbb{R}^m$. Thus, (32) is satisfied only if $\mathbf{b} \equiv \mathbf{0}$, i.e.,

$$\mathrm{Cov}[R^*, \mathbf{g_m}] = \mathbf{0}$$

which also implies $\mathrm{Cov}[R^*, \sum_{k=1}^m c_k g_k] = 0$ for any $c_1, ..., c_m \in \mathbb{R}$. Therefore,

$$\mathrm{Cov}[R^*, g] = 0,$$

for all $g \in \mathcal{G}$.

2. We now prove the other direction: if R^* is uncorrelated with all zero-estimators of a given family $\mathcal{G}$ then it is an OOE.

Let $R^* = T + g^*$ and $R \equiv T + g$ be unbiased estimators of θ, where $g^*, g \in \mathcal{G}$. Define $\tilde{g} \equiv R^* - R = g^* - g$ and notice that $\tilde{g} \in \mathcal{G}$. Since by assumption R^* is uncorrelated with $\tilde{g}$,

$$0 = \mathrm{Cov}[R^*, \tilde{g}] \equiv \mathrm{Cov}[R^*, R^* - R] = \mathrm{Var}[R^*] - \mathrm{Cov}[R^*, R],$$

and hence $\mathrm{Var}[R^*] = \mathrm{Cov}[R^*, R]$. By the Cauchy–Schwarz inequality,

$$(\mathrm{Cov}[R^*, R])^2 \leq \mathrm{Var}[R^*]\mathrm{Var}[R],$$

we conclude that $\mathrm{Var}[R^*] \leq \mathrm{Var}[R] = \mathrm{Var}[T + g]$ for all $g \in \mathcal{G}$. □

Proof of Theorem 2.

We start by proving Theorem 2 for the special case of $p = 2$ and then generalize for $p > 2$. By Theorem 1 we need to show that $\mathrm{Cov}\left(T_{oracle}, g_{k_1k_2}\right) = 0$ for all $(k_1, k_2) \in \mathbb{N}_0^2$ where $g_{k_1k_2} = \frac{1}{n}\sum_{i=1}^n \left[X_{i1}^{k_1} X_{i2}^{k_2} - E\left(X_{i1}^{k_1} X_{i2}^{k_2}\right)\right]$. Write,

$$\begin{aligned}\mathrm{Cov}\left(T_{oracle}, g_{k_1k_2}\right) &= \mathrm{Cov}\left(\hat{\tau}^2 - 2\sum_{j=1}^2\sum_{j'=1}^2 \psi_{jj'}, g_{k_1k_2}\right) \\ &= \mathrm{Cov}\left(\hat{\tau}^2, g_{k_1k_2}\right) - 2\sum_{j=1}^2\sum_{j'=1}^2 \mathrm{Cov}\left(\psi_{jj'}, g_{k_1k_2}\right).\end{aligned}$$

Thus, we need to show that

$$\mathrm{Cov}\left(\hat{\tau}^2, g_{k_1k_2}\right) = 2\sum_{j=1}^2\sum_{j'=1}^2 \mathrm{Cov}\left(\psi_{jj'}, g_{k_1k_2}\right). \tag{33}$$

We start with calculating the LHS of (33), namely $\mathrm{Cov}\left(\hat{\tau}^2, g_{k_1k_2}\right)$. Recall that $\hat{\tau}^2 \equiv \hat{\beta}_1^2 + \hat{\beta}_2^2$ and therefore $\mathrm{Cov}[\hat{\tau}^2, g_{k_1k_2}] = \mathrm{Cov}(\hat{\beta}_1^2, g_{k_1k_2}) + \mathrm{Cov}(\hat{\beta}_2^2, g_{k_1k_2})$. Now, for all $(k_1, k_2) \in \mathbb{N}_0^2$ we have

$$
\begin{aligned}
\mathrm{Cov}[\hat{\beta}_1^2, g_{k_1k_2}] &\equiv \mathrm{Cov}\left(\frac{2}{n(n-1)}\sum_{i_1<i_2} W_{i_1 1}W_{i_2 1}, \frac{1}{n}\sum_{i=1}^{n}(X_{i1}^{k_1}X_{i2}^{k_2} - E[X_{i1}^{k_1}X_{i2}^{k_2}])\right) \\
&= \frac{2}{n^2(n-1)}\sum_{i_1<i_2}\sum_{i=1}^{n}\mathrm{Cov}\left(X_{i_1 1}Y_{i_1}X_{i_2 1}Y_{i_2}, X_{i1}^{k_1}X_{i2}^{k_2}\right) \\
&= \frac{2}{n^2(n-1)}\sum_{i_1<i_2}\sum_{i=1}^{n}\left(E[X_{i_1 1}Y_{i_1}X_{i_2 1}Y_{i_2}X_{i1}^{k_1}X_{i2}^{k_2}] - \beta_1^2 E[X_{i1}^{k_1}X_{i2}^{k_2}]\right) \\
&= \frac{4}{n^2(n-1)}\sum_{i_1<i_2}\left(E[X_{i_1 1}Y_{i_1}X_{i_2 1}Y_{i_2}X_{i_1 1}^{k_1}X_{i_1 2}^{k_2}] - \beta_1^2 E[X_{i_1 1}^{k_1}X_{i_1 2}^{k_2}]\right) \\
&= \frac{4}{n^2(n-1)}\sum_{i_1<i_2}\left(E[X_{i_1 1}^{k_1+1}Y_{i_1}X_{i_1 2}^{k_2}]E[X_{i_2 1}Y_{i_2}] - \beta_1^2 E[X_{i_1 1}^{k_1}X_{i_1 2}^{k_2}]\right) \\
&= \frac{4}{n^2(n-1)}\sum_{i_1<i_2}\left(E[X_{i_1 1}^{k_1+1}Y_{i_1}X_{i_1 2}^{k_2}]\beta_1 - \beta_1^2 E[X_{i_1 1}^{k_1}X_{i_1 2}^{k_2}]\right) \\
&= \frac{4}{n^2(n-1)}\frac{n(n-1)}{2}\left(E[X_{11}^{k_1+1}Y_1X_{12}^{k_2}]\beta_1 - \beta_1^2 E[X_{11}^{k_1}X_{12}^{k_2}]\right) \\
&= \frac{2}{n}\left(E[X_{11}^{k_1+1}Y_1X_{12}^{k_2}]\beta_1 - \beta_1^2 E[X_{11}^{k_1}X_{12}^{k_2}]\right), \qquad (34)
\end{aligned}
$$

where the calculations can be justified by similar arguments to those presented in (30). We shall use the following notation:

$$
\begin{aligned}
A &\equiv E\left[X_{11}^{k_1+2}X_{12}^{k_2}\right] \\
B &\equiv E\left[X_{11}^{k_1+1}X_{12}^{k_2+1}\right] \\
C &\equiv E\left[X_{11}^{k_1}X_{12}^{k_2}\right] \\
D &\equiv E\left[X_{11}^{k_1}X_{12}^{k_2+2}\right].
\end{aligned}
$$

Notice that A, B, C and D are functions of (k_1, k_2) but this is suppressed in the notation. Write,

$$
\begin{aligned}
E[X_{11}^{k_1+1}X_{12}^{k_2}Y_1] &= E[X_{11}^{k_1+1}X_{12}^{k_2}(\beta_1 X_{11} + \beta_2 X_{12} + \epsilon_1)] \\
&= \beta_1 E[X_{11}^{k_1+2}X_{12}^{k_2}] + \beta_2 E[X_{11}^{k_1+1}X_{12}^{k_2+1}] = \beta_1 A + \beta_2 B.
\end{aligned}
$$

Thus, rewrite (34) and obtain

$$
\mathrm{Cov}[\hat{\beta}_1^2, g_{k_1k_2}] = \frac{2}{n}\left([\beta_1 A + \beta_2 B]\,\beta_1 - \beta_1^2 C\right). \qquad (35)
$$

Similarly, by symmetry,

$$\mathrm{Cov}[\hat{\beta}_2^2, g_{k_1k_2}] = \frac{2}{n}\left([\beta_2 D + \beta_1 B]\,\beta_2 - \beta_2^2 C\right). \tag{36}$$

Using (35) and (36) we get

$$\begin{aligned}
\mathrm{Cov}[\hat{\tau}^2, g_{k_1k_2}] &= \mathrm{Cov}(\hat{\beta}_1^2, g_{k_1k_2}) + \mathrm{Cov}(\hat{\beta}_2^2, g_{k_1k_2}) \\
&= \frac{2}{n}\left([\beta_1 A + \beta_2 B]\,\beta_1 - \beta_1^2 C + [\beta_2 D + \beta_1 B]\,\beta_2 - \beta_2^2 C\right) \\
&= \frac{2}{n}\left[\overbrace{\beta_1^2 A + \beta_2^2 D}^{L_1} + \overbrace{2\beta_1\beta_2 B}^{L_2} - \overbrace{C\left(\beta_1^2 + \beta_2^2\right)}^{L_3}\right] \\
&= \frac{2}{n}\left(L_1 + L_2 - L_3\right).
\end{aligned} \tag{37}$$

We now move to calculate the RHS of (33), namely $\sum_{j=1}^{2}\sum_{j'=1}^{2}\mathrm{Cov}(\psi_{jj'}, g_{k_1k_2})$. First, recall that

$$h_{jj} \equiv \frac{1}{n}\sum_{i=1}^{n}\left[X_{ij}X_{ij'} - E\left(X_{ij}X_{ij'}\right)\right]$$

and

$$g_{k_1k_2} \equiv \frac{1}{n}\sum_{i=1}^{n}\left[X_{i1}^{k_1}X_{i2}^{k_2} - E\left(X_{i1}^{k_1}X_{i2}^{k_2}\right)\right],$$

where $(k_1, k_2) \in \mathbb{N}_0^2$. Hence, $h_{11} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_{i1}^2 - 1\right)$ which by definition is also equal to g_{20}. Similarly, we have $h_{12} = h_{21} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_{i1}X_{i2}\right) = g_{11}$ and $h_{22} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_{i2}^2 - 1\right) = g_{02}$. Thus,

$$\begin{aligned}
\sum_{j=1}^{2}\sum_{j'=1}^{2}\mathrm{Cov}\left(\psi_{jj'}, g_{k_1k_2}\right) &= \sum_{j=1}^{2}\sum_{j'=1}^{2}\beta_j\beta_{j'}\mathrm{Cov}\left(h_{jj'}, g_{k_1k_2}\right) \\
&= \beta_1^2\mathrm{Cov}\left(h_{11}, g_{k_1k_2}\right) + 2\beta_1\beta_2\mathrm{Cov}\left(h_{12}, g_{k_1k_2}\right) + \beta_2^2\mathrm{Cov}\left(h_{22}, g_{k_1k_2}\right) \\
&= \beta_1^2\mathrm{Cov}\left(g_{20}, g_{k_1k_2}\right) + 2\beta_1\beta_2\mathrm{Cov}\left(g_{11}, g_{k_1k_2}\right) + \beta_2^2\mathrm{Cov}\left(g_{02}, g_{k_1k_2}\right).
\end{aligned} \tag{38}$$

Now, observe that for every $(k_1, k_2, d_1, d_2) \in \mathbb{N}_0^4$,

$$\begin{aligned}
&\mathrm{Cov}[g_{k_1k_2}, g_{d_1d_2}] \\
&= \mathrm{Cov}\Big(\frac{1}{n}\sum_{i=1}^{n}[X_{i1}^{k_1}X_{i2}^{k_2} - E(X_{i1}^{k_1}X_{i2}^{k_2})], \frac{1}{n}\sum_{i=1}^{n}[X_{i1}^{d_1}X_{i2}^{d_2} - E(X_{i1}^{d_1}X_{i2}^{d_2})]\Big) \\
&= n^{-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}(E[X_{i_11}^{k_1}X_{i_12}^{k_2}X_{i_21}^{d_1}X_{i_22}^{d_2}] - E[X_{i_11}^{k_1}X_{i_12}^{k_2}]E[X_{i_21}^{d_1}X_{i_22}^{d_2}]) \\
&= \frac{1}{n}\Big(E[X_{11}^{k_1+d_1}X_{12}^{k_2+d_2}] - E[X_{11}^{k_1}X_{12}^{k_2}]E[X_{11}^{d_1}X_{12}^{d_2}]\Big),
\end{aligned} \tag{39}$$

where the third equality holds since the terms with $i_1 \neq i_2$ vanish. It follows from (39) that

$$\begin{aligned}
\mathrm{Cov}[g_{k_1k_2}, g_{20}] &= \tfrac{1}{n}\left(E[X_{11}^{k_1+2}X_{12}^{k_2}] - E[X_{11}^{k_1}X_{12}^{k_2}]\right) = \tfrac{1}{n}(A-C),\\
\mathrm{Cov}[g_{k_1k_2}, g_{11}] &= \tfrac{1}{n}E[X_{11}^{k_1+1}X_{12}^{k_2+1}] = \tfrac{B}{n},\\
\mathrm{Cov}[g_{k_1k_2}, g_{02}] &= \tfrac{1}{n}\left(E[X_{11}^{k_1}X_{12}^{k_2+2}] - E[X_{11}^{k_1}X_{12}^{k_2}]\right) = \tfrac{1}{n}(D-C)
\end{aligned}$$

Therefore, rewrite (38) to get

$$\begin{aligned}
2\sum_{j=1}^{2}\sum_{j'=1}^{2}\mathrm{Cov}\left(\psi_{jj'}, g_{k_1k_2}\right) &= \frac{2}{n}\left[\overbrace{\beta_1^2A+\beta_2^2D}^{L_1} + \overbrace{2\beta_1\beta_2B}^{L_2} - \overbrace{C\left(\beta_1^2+\beta_2^2\right)}^{L_3}\right]\\
&= \frac{2}{n}\left(L_1+L_2-L_3\right), \qquad (40)
\end{aligned}$$

which is exactly the same expression as in (37). Hence, equation (33) follows which completes the proof of Theorem 2 for $p = 2$.

We now generalize the proof for $p > 2$. Similarly to (33) we want to show that

$$\mathrm{Cov}\left(\hat{\tau}^2, g_{k_1\ldots k_p}\right) = 2\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathrm{Cov}\left(\psi_{jj'}, g_{k_1\ldots k_p}\right). \qquad (41)$$

We begin by calculating the LHS of (41), i.e., the covariance between $\hat{\tau}^2$ and $g_{k_1\ldots k_p}$. By the same type of calculations as in (34), for all $(k_1, \ldots, k_p) \in \mathbb{N}_0^p$ we have

$$\begin{aligned}
&\mathrm{Cov}\left[\hat{\beta}_j^2, g_{k_1,\ldots,k_p}\right] =\\
&\tfrac{2}{n}\left\{\left[\beta_j E\left(X_{1j}^{k_j+2}\prod_{m\neq j}X_{1m}^{k_m}\right) + \sum_{j\neq j'}\beta_{j'}E\left(X_{1j}^{k_j+1}X_{1j'}^{k_{j'}+1}\prod_{m\neq j,j'}X_{1m}^{k_m}\right)\right]\beta_j - \beta_j^2 E\left(\prod_{m=1}^{p}X_{1m}^{k_m}\right)\right\}
\end{aligned}$$

Summing the above expressions for $j = 1, \ldots, p$, yields

$$\begin{aligned}
\mathrm{Cov}\left[\hat{\tau}^2, g_{k_1,\ldots,k_p}\right] &= \sum_{j=1}^{p}\mathrm{Cov}\left[\hat{\beta}_j^2, g_{k_1,\ldots,k_p}\right]\\
&= \frac{2}{n}\sum_{j=1}^{p}\beta_j^2 E\left(X_{1j}^{k_j+2}\prod_{m\neq j}X_{1m}^{k_m}\right)\\
&+ \frac{2}{n}\sum_{j\neq j'}\beta_j\beta_{j'}E\left(X_{1j}^{k_j+1}X_{1j'}^{k_{j'}+1}\prod_{m\neq j,j'}X_{1m}^{k_m}\right) \qquad (42)\\
&- \frac{2}{n}\sum_{j=1}^{p}\beta_j^2 E\left(\prod_{m=1}^{p}X_{1m}^{k_m}\right)\\
&\equiv \frac{2}{n}\left(L_1+L_2-L_3\right),
\end{aligned}$$

where L_1, L_2 and L_3 are just a generalization of the notation given in (37). Again, notice that L_1, L_2 and L_3 are functions of $k_1, ..., k_p$ but this is suppressed in the notation.

We now move to calculate the RHS of (41), namely $2\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathrm{Cov}\left(\psi_{jj'}, g_{k_1...k_p}\right)$. Since $\psi_{jj'} = \beta_j\beta_{j'}h_{jj'}$ we have,

$$\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathrm{Cov}\left(\psi_{jj'}, g_{k_1...k_p}\right) = \sum_{j=1}^{p}\sum_{j'=1}^{p}\beta_j\beta_{j'}\mathrm{Cov}\left(h_{jj'}, g_{k_1...k_p}\right). \tag{43}$$

Again, notice the relationship between $h_{jj'}$ and $g_{k_1...k_p}$: when $j = j'$ we have $h_{jj} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_{ij}^2 - 1\right) = g_{0...2...0}$, (i.e., the j-th entry is 2 and all others are 0), and for $j \neq j'$ we have $h_{jj'} \equiv \frac{1}{n}\sum_{i=1}^{n}X_{ij}X_{ij'} = g_{0...1...1...0}$, (i.e., the j-th and j'-th entries are 1 and all other entries are 0). Hence,

$$\begin{aligned}\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathrm{Cov}\left(\psi_{jj'}, g_{k_1...k_p}\right) &= \sum_{j=1}^{p}\sum_{j'=1}^{p}\beta_j\beta_{j'}\mathrm{Cov}\left(h_{jj'}, g_{k_1...k_p}\right)\\ &= \sum_{j=1}^{p}\beta_j^2\mathrm{Cov}\left(g_{0...2...0}, g_{k_1...k_p}\right) + \sum_{j\neq j'}\beta_j\beta_{j'}\mathrm{Cov}\left(g_{0...1...1...0}, g_{k_1...k_p}\right).\end{aligned} \tag{44}$$

Similarly to (39), for all pairs of index vectors $(k_1, ..., k_p) \in \mathbb{N}_0^p$, and $(k_1', ..., k_p') \in \mathbb{N}_0^p$ we have,

$$\begin{aligned}&\mathrm{Cov}\left(g_{k_1,...,k_p}, g_{k_1',...,k_p'}\right)\\ &= \frac{1}{n}\left\{E\left(\prod_{j=1}^{p}X_{1j}^{k_j+k_j'}\right) - E\left(\prod_{j=1}^{p}X_{1j}^{k_j}\right)E\left(\prod_{j=1}^{p}X_{1j}^{k_j'}\right)\right\}.\end{aligned}$$

This implies that

$$\mathrm{Cov}\left[g_{0...2...0}, g_{k_1,...,k_p}\right] = \frac{1}{n}\left[E\left(X_{1j}^{k_j+2}\prod_{m\neq j}X_{1m}^{k_m}\right) - E\left(\prod_{m=1}^{p}X_{1m}^{k_m}\right)\right]$$

and

$$\mathrm{Cov}\left[g_{0...1...1...0}, g_{k_1,...,k_p}\right] = \frac{1}{n}E\left(X_{1j}^{k_j+1}X_{1j'}^{k_{j'}+1}\prod_{m\neq j,j'}X_{1m}^{k_m}\right).$$

Hence, rewrite (44) to see that

$$\begin{aligned}&2\sum_{j=1}^{p}\sum_{j'=1}^{p}\mathrm{Cov}\left(\psi_{jj'}, g_{k_1...k_p}\right)\\ &= 2\sum_{j=1}^{p}\beta_j^2\mathrm{Cov}\left(g_{0...2...0}, g_{k_1...k_p}\right) + 2\sum_{j\neq j'}\beta_j\beta_{j'}\mathrm{Cov}\left(g_{0...1...1...0}, g_{k_1...k_p}\right)\end{aligned}$$

$$= \frac{2}{n}\sum_{j=1}^{p}\beta_j^2\left[E\left(X_{1j}^{k_j+2}\prod_{m\neq j}X_{1m}^{k_m}\right)-E\left(\prod_{m=1}^{p}X_{1m}^{k_m}\right)\right]+$$

$$\frac{2}{n}\sum_{j\neq j'}\beta_j\beta_{j'}E\left(X_{1j}^{k_j+1}X_{1j'}^{k_{j'}+1}\prod_{m\neq j,j'}X_{1m}^{k_m}\right)=\frac{2}{n}\big(L_1-L_3+L_2\big), \quad (45)$$

which is exactly the same expression as in (42). Hence, equation (41) follows which completes the proof of Theorem 2. □

Proof of Corollary 2.
Write,

$$\mathrm{Var}\left(T_{oracle}\right)=\mathrm{Var}\left(\hat{\tau}^2-2\sum_{j,j'}\psi_{jj'}\right)$$

$$=\mathrm{Var}\left(\hat{\tau}^2\right)-4\sum_{j,j'}\beta_j\beta_{j'}\mathrm{Cov}\left(\hat{\tau}^2,h_{jj'}\right)+4\,\mathrm{Var}\left(\sum_{j,j'}\psi_{jj'}\right). \quad (46)$$

Consider $\sum_{j,j'}\beta_j\beta_{j'}\mathrm{Cov}\left(\hat{\tau}^2,h_{jj'}\right)$. We have

$$\sum_{j,j'}\beta_j\beta_{j'}\mathrm{Cov}\left(\hat{\tau}^2,h_{jj'}\right)$$

$$=\sum_{j=1}^{p}\beta_j^2\mathrm{Cov}\left(\hat{\tau}^2,h_{jj}\right)+\sum_{j\neq j'}\beta_j\beta_{j'}\mathrm{Cov}\left(\hat{\tau}^2,h_{jj'}\right)$$

$$=\sum_{j=1}^{p}\beta_j^2\mathrm{Cov}\left(\hat{\tau}^2,g_{0\ldots2\ldots0}\right)+\sum_{j\neq j'}\beta_j\beta_{j'}\mathrm{Cov}\left(\hat{\tau}^2,g_{0\ldots1\ldots1\ldots0}\right)$$

$$=\sum_{j=1}^{p}\beta_j^2\left[\frac{2\beta_j^2}{n}\left(E\left(X_{1j}^4\right)-1\right)\right]+\sum_{j\neq j'}\beta_j\beta_{j'}\left[\frac{4}{n}\beta_j\beta_{j'}\right]$$

$$=\frac{2}{n}\sum_{j=1}^{p}\beta_j^4\left[\left(E\left(X_{1j}^4\right)-1\right)\right]+\frac{4}{n}\sum_{j\neq j'}\beta_j^2\beta_{j'}^2$$

where the second and third equality are justified by (44) and (42) respectively. Consider now $\mathrm{Var}\left(\sum_{j,j'}\psi_{jj'}\right)$. Let $\beta_{1234}\equiv\beta_{j_1}\beta_{j_2}\beta_{j_3}\beta_{j_4}$. Write,

$$\mathrm{Var}\left(\sum_{j,j'}\psi_{jj'}\right)=\mathrm{Cov}\left(\sum_{j,j'}\beta_j\beta_{j'}h_{jj'},\sum_{j,j'}\beta_j\beta_{j'}h_{jj'}\right)$$

$$=\sum_{j_1,j_2,j_3j_4}\beta_{1234}\mathrm{Cov}\left(h_{j_1j_2},h_{j_3j_4}\right)$$

$$=\frac{1}{n^2}\sum_{j_1,j_2,j_3,j_4}\beta_{1234}\sum_{i_1,i_2}\mathrm{Cov}\left(X_{i_1j_1}X_{i_1j_2},X_{i_2j_3}X_{i_2j_4}\right)$$

$$= \frac{1}{n^2} \sum_{j_1,j_2,j_3,j_4} \beta_{1234} \sum_{i_1,i_2} \left[E\left(X_{i_1 j_1} X_{i_1 j_2} X_{i_2 j_3} X_{i_2 j_4}\right) - E\left(X_{i_1 j_1} X_{i_1 j_2}\right) E\left(X_{i_2 j_3} X_{i_2 j_4}\right) \right]$$

$$= n^{-2} \sum_{j_1,j_2,j_3,j_4} \beta_{1234} \sum_{i=1}^{n} \left[E\left(X_{ij_1} X_{ij_2} X_{ij_3} X_{ij_4}\right) - E\left(X_{ij_1} X_{ij_2}\right) E\left(X_{ij_3} X_{ij_4}\right) \right]$$

$$= \frac{1}{n} \sum_{j_1,j_2,j_3,j_4} \beta_{1234} \left[E\left(X_{1j_1} X_{1j_2} X_{1j_3} X_{1j_4}\right) - E\left(X_{1j_1} X_{1j_2}\right) E\left(X_{1j_3} X_{1j_4}\right) \right],$$

where the fifth equality holds since the summand is 0 for all $i_1 \neq i_2$. The summation is not zero in only three cases:

1) $j_1 = j_4 \neq j_2 = j_3$
2) $j_1 = j_3 \neq j_2 = j_4$
3) $j_1 = j_2 = j_3 = j_4$.

For the first two cases the summation equals $\frac{1}{n} \sum_{j \neq j'} \beta_j^2 \beta_{j'}^2$. For the third case the summation equals to $\frac{1}{n} \sum_{j=1}^{n} \beta_j^4 \left[E\left(X_{1j}^4 - 1\right) \right]$. Overall we have

$$\mathrm{Var}\left(\sum_{j,j'} \psi_{jj'} \right) = \overbrace{\frac{1}{n} \sum_{j \neq j'} \beta_j^2 \beta_{j'}^2}^{case\,1} + \overbrace{\frac{1}{n} \sum_{j \neq j'} \beta_j^2 \beta_{j'}^2}^{case\,2} + \overbrace{\frac{1}{n} \sum_{j=1}^{n} \beta_j^4 \left[E\left(X_{1j}^4 - 1\right) \right]}^{case\,3}.$$

Rewrite (46) to get

$$\begin{aligned}
\mathrm{Var}\left(T_{oracle}\right) &= \mathrm{Var}\left(\hat{\tau}^2\right) - 4 \sum_{j,j'} \beta_j \beta_{j'} \mathrm{Cov}\left(\hat{\tau}^2, h_{jj'}\right) + 4\,\mathrm{Var}\left(\sum_{j,j'} \psi_{jj'} \right) \\
&= \mathrm{Var}\left(\hat{\tau}^2\right) - 4 \left[\frac{2}{n} \sum_{j=1}^{p} \beta_j^4 \left[\left(E\left(X_{1j}^4\right) - 1 \right) \right] + \frac{4}{n} \sum_{j \neq j'} \beta_j^2 \beta_{j'}^2 \right] \\
&\quad + \frac{4}{n} \left\{ 2 \sum_{j \neq j'} \beta_j^2 \beta_{j'}^2 + \sum_{j=1}^{p} \beta_j^4 \left[E\left(X_{1j}^4 - 1\right) \right] \right\} \\
&= \mathrm{Var}\left(\hat{\tau}^2\right) - \frac{4}{n} \left\{ \sum_{j=1}^{p} \beta_j^4 \left[E\left(X_{1j}^4 - 1\right) \right] + 2 \sum_{j \neq j'} \beta_j^2 \beta_{j'}^2 \right\}.
\end{aligned}$$

□

Remark 3. ***Calculations for Example 2.***

Recall that by (6) we have

$$Var\left(\hat{\tau}^2\right) = \frac{4\left(n-2\right)}{n\left(n-1\right)} \left[\beta^T \mathbf{A} \beta - \|\beta\|^4 \right] + \frac{2}{n\left(n-1\right)} \left[\|\mathbf{A}\|_F^2 - \|\beta\|^4 \right].$$

Now, when we assume standard Gaussian covariates, one can verify that $\beta^T \mathbf{A} \beta - \|\beta\|^4 = \sigma_Y^2 \tau^2 + \tau^4$ *and* $\|\mathbf{A}\|_F^2 - \|\beta\|^4 = p\sigma_Y^4 + 4\sigma_Y^2 \tau^2 + 3\tau^4$, *where* $\sigma_Y^2 = \sigma^2 + \tau^2$.

Thus, in this case we can write

$$Var\left(\hat{\tau}^2\right) = \frac{4}{n}\left[\frac{(n-2)}{(n-1)}\left[\sigma_Y^2\tau^2+\tau^4\right] + \frac{1}{2\,(n-1)}\left(p\sigma_Y^4 + 4\sigma_Y^2\tau^2 + 3\tau^4\right)\right]. \quad (47)$$

Plug-in $\tau^2 = \sigma^2 = 1$ *to get*

$$Var(\hat{\tau}^2) = \frac{20}{n} + O(n^{-2}), \quad (48)$$

and $Var(T_{oracle}) = Var(\hat{\tau}^2) - \frac{8}{n}\tau^4 = \frac{12}{n} + O(n^{-2})$ *by* (12). *More generally, the asymptotic improvement of* T_{oracle} *over the naive estimator is:*

$$\begin{aligned}
&\lim_{n,p\to\infty} \frac{Var\left(\hat{\tau}^2\right) - Var\left(T_{oracle}\right)}{Var\left(\hat{\tau}^2\right)} \\
&= \lim_{n,p\to\infty} \frac{8\tau^4/n}{\frac{4}{n}\left[\frac{(n-2)}{(n-1)}\left(\sigma_Y^2\tau^2+\tau^4\right) + \frac{1}{2(n-1)}\left(p\sigma_Y^4 + 4\sigma_Y^2\tau^2 + 3\tau^4\right)\right]} \\
&= \frac{2\tau^4}{3\tau^4 + \frac{4p\tau^4 + 4\sigma_Y^2\tau^2 + 3\tau^4}{2n}} = \frac{2}{3+2\frac{p}{n}},
\end{aligned}$$

where we used the fact that $\sigma_Y^2 = \tau^2 + \sigma^2 = 2\tau^2$ *in the second equality. Now, notice that when* $p = n$ *then the reduction is* $\frac{2}{3+2} = 40\%$ *and when* p/n *converges to zero, the reduction is* 66%.

In order to verify the above results, we repeated the simulation study from Section 5 but with Gaussian covariates, considering only the naive estimator $\hat{\tau}^2$ *and the OOE estimator* T_{oracle}. *For the low-dimensional case, we fixed* $p = 3$ *and considered* $\beta_j^2 = \frac{\tau_{\mathbf{B}}^2}{2}$ *for* $j = 1, \ldots, 2$, *and* $\beta_3^2 = \tau^2 - \tau_{\mathbf{B}}^2$. *Table 4 suggests that the OOE estimator achieves similar reduction in variance as claimed in theory, namely that for the low-dimensional case, the reduction is about 66% and when* $n = p$ *the reduction is about 40%.*

Proof of Proposition 3.

Write,

$$\begin{aligned}
\mathrm{Var}\left(T\right) &= \mathrm{Var}\left[\hat{\tau}^2 - 2\sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj'}\right] \\
&= \mathrm{Var}\left(\hat{\tau}^2\right) - 4\mathrm{Cov}\left(\hat{\tau}^2, \sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj'}\right) + 4\mathrm{Var}\left(\sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj'}\right). \quad (49)
\end{aligned}$$

We start with calculating the middle term. Let

$$p_n(k) \equiv n(n-1)(n-2)\cdots(n-k).$$

TABLE 4

Summary statistics for the naive and the OOE estimators; $n = 300$; $\tau^2 = \sigma^2 = 1$. Mean, mean square error (MSE) and percentage change from the naive estimator (in terms of MSE) are shown. Simulation standard errors are shown in parenthesis. The table results were computed over 100 simulated datasets for each setting.

p	$\tau^2_{\mathbf{B}}$	Estimator	Mean	MSE	% Change
300	5%	Naive	1 (0.015)	0.068 (0.106)	0
300	5%	OOE	1.01 (0.012)	0.042 (0.07)	−38.24
300	35%	Naive	1.01 (0.015)	0.068 (0.095)	0
300	35%	OOE	1.02 (0.011)	0.039 (0.059)	−42.65
300	65%	Naive	1.01 (0.015)	0.069 (0.101)	0
300	65%	OOE	1.01 (0.011)	0.038 (0.056)	−44.93
300	95%	Naive	1.01 (0.015)	0.072 (0.11)	0
300	95%	OOE	1.01 (0.012)	0.04 (0.058)	−44.44
3	5%	Naive	0.99 (0.011)	0.033 (0.05)	0
3	5%	OOE	1.01 (0.007)	0.014 (0.022)	−57.58
3	35%	Naive	1.01 (0.012)	0.043 (0.069)	0
3	35%	OOE	1.01 (0.007)	0.016 (0.025)	−62.79
3	65%	Naive	1.01 (0.013)	0.052 (0.087)	0
3	65%	OOE	1.01 (0.008)	0.018 (0.029)	−65.38
3	95%	Naive	1.01 (0.013)	0.049 (0.083)	0
3	95%	OOE	1.01 (0.007)	0.017 (0.029)	−65.31

Write,

$$\begin{aligned}
&\operatorname{Cov}\left(\hat{\tau}^2, \sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj}\right)\\
&= \operatorname{Cov}\Bigg(\frac{1}{p_n(1)}\sum_{i_1\neq i_2}\sum_{j=1}^{p} W_{i_1 j}W_{i_2 j}\\
&\qquad\qquad , \frac{1}{p_n(2)}\sum_{j,j'}\sum_{i_1\neq i_2\neq i_3} W_{i_1 j}W_{i_2 j'}\left[X_{i_3 j}X_{i_3 j'} - E\left(X_{i_3 j}X_{i_3 j'}\right)\right]\Bigg)\\
&= C_n \sum_{I}\sum_{J} \operatorname{Cov}\left(W_{i_1 j_1}W_{i_2 j_1}, W_{i_3 j_2}W_{i_4 j_3}\left[X_{i_5 j_2}X_{i_5 j_3} - E\left(X_{i_5 j_2}X_{i_5 j_3}\right)\right]\right), \quad (50)
\end{aligned}$$

where $C_n \equiv \frac{1}{p_n(1)\cdot p_n(2)}$, I is the set of all quintuples of indices $(i_1, i_2, i_3, i_4, i_5)$ such that $i_1 \neq i_2$ and $i_3 \neq i_4 \neq i_5$, and J is the set of all triples of indices (j_1, j_2, j_3). For the set I, there are $\binom{2}{1}\cdot 3 = 6$ different cases to consider when one of $\{i_1, i_2\}$ is equal to one of $\{i_3, i_4, i_5\}$, and an additional $\binom{2}{2}\cdot 3! = 6$ cases to consider when two of $\{i_1, i_2\}$ are equal to two of $\{i_3, i_4, i_5\}$. Similarly, for the set J there are three cases to consider when only two indices of $\{j_1, j_2, j_3\}$ are equal to each other, (e.g., $j_1 = j_2 \neq j_3$); one case to consider when no pair of indices is equal to each other and, one case to consider when all three indices are equal. Thus, there are total of $(6+6)\times(3+1+1) = 60$ cases to consider. Here we demonstrate only one such case. Let $I_1 = \{(i_1, \ldots, i_5) : i_1 = i_5 \neq i_2 \neq i_3 \neq i_4\}$

and $J_1 = \{(j_1, j_2, j_3) : j_1 = j_2 = j_3\}$. Write,

$$\begin{aligned}
&C_n \sum_{I_1} \sum_{J_1} \mathrm{Cov}\left(W_{i_1 j_1} W_{i_2 j_1}, W_{i_3 j_2} W_{i_4 j_3} \left[X_{i_5 j_2} X_{i_5 j_3} - E\left(X_{i_5 j_2} X_{i_5 j_3}\right)\right]\right) \\
&= C_n \sum_{I_1} \sum_{j=1}^{p} \mathrm{Cov}\left(W_{i_1 j} W_{i_2 j}, W_{i_3 j} W_{i_4 j} \left[X_{i_1 j}^2 - 1\right]\right) \\
&= C_n \sum_{I_1} \sum_{j=1}^{p} E(W_{i_2 j}) E(W_{i_3 j}) E(W_{i_4 j}) E\left(W_{i_1 j} \left[X_{i_1 j}^2 - 1\right]\right) \\
&= C_n \sum_{I_1} \sum_{j=1}^{p} \beta_j^3 E\left(W_{ij} \left[X_{ij}^2 - 1\right]\right). \qquad (51)
\end{aligned}$$

Now, notice that

$$\begin{aligned}
E\left[W_{ij}\left(X_{ij}^2 - 1\right)\right] &= E\left[X_{ij} Y_i \left(X_{ij}^2 - 1\right)\right] \\
&= E\left[X_{ij}^3 \left(\beta^T X + \varepsilon_i\right)\right] - \beta_j \\
&= \beta_j E\left(X_{ij}^4\right) - \beta_j \\
&= \beta_j [E(X_{ij}^4) - 1]. \qquad (52)
\end{aligned}$$

Rewrite (51) to get

$$\begin{aligned}
C_n \sum_{I_1} \sum_{j=1}^{p} \beta_j^3 E\left[W_{ij}\left(X_{ij}^2 - 1\right)\right] &= C_n \sum_{I_1} \sum_{j=1}^{p} \beta_j^3 \left(\beta_j [E(X_{ij}^4) - 1]\right) \\
&= \frac{p_n(3)}{p_n(1) \cdot p_n(2)} \sum_{j=1}^{p} \beta_j^4 [E(X_{ij}^4) - 1] \\
&= \frac{(n-3)}{n(n-1)} \sum_{j=1}^{p} \beta_j^4 [E(X_{ij}^4) - 1] \\
&= \frac{1}{n} \sum_{j=1}^{p} \beta_j^4 [E(X_{ij}^4) - 1] + O(n^{-2}),
\end{aligned}$$

where we used (52) to justify the first equality. By the same type of calculation, one can compute the covariance in (50) over all 60 and obtain that

$$\mathrm{Cov}\left(\hat{\tau}^2, \sum_{j=1}^{p} \sum_{j'=1}^{p} \hat{\psi}_{jj'}\right) = \frac{2}{n} \left\{ \sum_{j=1}^{p} \beta_j^4 \left[E\left(X_{1j}^4 - 1\right)\right] + 2 \sum_{j \neq j'} \beta_j^2 \beta_{j'}^2 \right\} + O\left(n^{-2}\right). \qquad (53)$$

We now move to calculate the last term of (49). Recall that

$$\hat{\psi}_{jj'} = \frac{1}{n(n-1)(n-2)} \sum_{i_1 \neq i_2 \neq i_3} W_{i_1 j} W_{i_2 j'} \left[X_{i_3 j} X_{i_3 j'} - E\left(X_{i_3 j} X_{i_3 j'}\right)\right].$$

Therefore,

$$\operatorname{Var}\left(\sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj'}\right)=\sum_{J}\operatorname{Cov}\left(\hat{\psi}_{j_1j_2},\hat{\psi}_{j_3j_4}\right) \tag{54}$$

$$=p_n^{-2}(2)\sum_{J}\operatorname{Cov}\left(\sum_{i_1\neq i_2\neq i_3}W_{i_1j_1}W_{i_2j_2}X_{i_3j_1}X_{i_3j_2},\sum_{i_1\neq i_2\neq i_3}W_{i_1j_3}W_{i_2j_4}X_{i_3j_3}X_{i_3j_4}\right)$$

$$=p_n^{-2}(2)\sum_{J}\sum_{I}\operatorname{Cov}\left(W_{i_1j_1}W_{i_2j_2}X_{i_3j_1}X_{i_3j_2},W_{i_4j_3}W_{i_5j_4}X_{i_6j_3}X_{i_6j_4}\right),$$

where J is now defined to be the set of all quadruples (j_1,j_2,j_3,j_4), and I is now defined to be the set of all sextuples $(i_1,...,i_6)$ such that $i_1\neq i_2\neq i_3$ and $i_4\neq i_5\neq i_6$. For the set I, there are three different cases to consider: (1) when one of $\{i_1,i_2,i_3\}$ is equal to one of $\{i_4,i_5,i_6\}$; (2) when two of $\{i_1,i_2,i_3\}$ are equal to two of $\{i_4,i_5,i_6\}$; and (3) when $\{i_1,i_2,i_3\}$ are equal to $\{i_4,i_5,i_6\}$. There are $\binom{3}{1}\cdot 3=9$ options for the first case, $\binom{3}{2}\cdot 3!=18$ for the second case, and $\binom{3}{3}\cdot 3!=6$ options for the third case. For the set J, there are five different cases to consider: (1) when there is only *one* pair of equal indices (e.g., $j_1=j_2\neq j_3\neq j_4$); (2) when there are *two* pairs of equal indices (e.g., $j_1=j_2\neq j_3=j_4$); (3) when only three indices are equal (e.g., $j_1=j_2=j_3\neq j_4$); (4) when all four indices are equal and; (5) all four indices are different from each other. Note that there are $\binom{4}{2}=6$ combinations for the first case, $\binom{4}{2}=6$ for the second case, $\binom{4}{3}=4$ combinations for the third case, and a single combination for each of the last two cases. Thus, there are total of $(9+18+6)\times(6+6+4+1+1)=594$. Again we demonstrate only one such calculation. Let $I_2=\{(i_1,...,i_6):i_1=i_4,i_2=i_5,i_3=i_6\}$ and $J_2=\{(j_1,j_2,j_3,j_4):j_1=j_3\neq j_2=j_4\}$. In the view of (54),

$$p_n^{-2}(2)\sum_{J_2}\sum_{I_2}\operatorname{Cov}\left(W_{i_1j_1}W_{i_2j_2}X_{i_3j_1}X_{i_3j_2},W_{i_4j_3}W_{i_5j_4}X_{i_6j_3}X_{i_6j_4}\right)=$$

$$=p_n^{-2}(2)\sum_{J_2}\sum_{I_2}\operatorname{Cov}\left(W_{i_1j_1}W_{i_2j_2}X_{i_3j_1}X_{i_3j_2},W_{i_1j_1}W_{i_2j_2}X_{i_3j_1}X_{i_3j_2}\right)$$

$$=p_n^{-2}(2)\sum_{J_2}\sum_{I_2}E\left(W_{i_1j_1}^2\right)E\left(W_{i_2j_2}^2\right)E\left(X_{i_3j_1}^2\right)E\left(X_{i_3j_2}^2\right)$$

$$=p_n^{-2}(2)\sum_{J_2}\sum_{I_2}\left(\sigma_Y^2+\beta_{j_1}^2\left\{E\left(X_{ij_1}^4\right)-1\right\}\right)\left(\sigma_Y^2+\beta_{j_2}^2\left\{E\left(X_{ij_2}^4\right)-1\right\}\right)$$

$$\leq p_n^{-2}(2)\sum_{J_2}\sum_{I_2}\left(\sigma_Y^2+\beta_{j_1}^2\left(C-1\right)\right)\left(\sigma_Y^2+\beta_{j_2}^2\left(C-1\right)\right)$$

$$=p_n^{-1}(2)\sum_{j_1\neq j_2}\left[\sigma_Y^4+\sigma_Y^2\left(C-1\right)\left(\beta_{j_1}^2+\beta_{j_2}^2\right)+\left(C-1\right)^2\beta_{j_1}^2\beta_{j_2}^2\right]$$

$$=p_n^{-1}(2)\left[p\left(p-1\right)\sigma_Y^4+\sigma_Y^2\left(C-1\right)\sum_{j_1\neq j_2}\left(\beta_{j_1}^2+\beta_{j_2}^2\right)+\left(C-1\right)^2\sum_{j_1\neq j_2}\beta_{j_1}^2\beta_{j_2}^2\right]$$

$$\leq p_n^{-1}(2)\left[p\left(p-1\right)\sigma_Y^4+\sigma_Y^2\left(C-1\right)\left(2p\tau^2\right)+\left(C-1\right)^2\tau^4\right],$$

where the fourth equality we use $E\left(W_{ij}^2\right)=\sigma_Y^2+\beta_j^2[E(X_{ij}^2)-1]$, which is given by (28), and in the fifth equality we used the assumption that $E(X_{ij}^4)\leq C$ for some positive C. Since we assume $p/n=O(1)$, the above expression can be further simplified to $\frac{p^2\sigma_Y^4}{n^3}+O\left(n^{-2}\right)$.

By the same type of calculation, one can compute the covariance in (54) over all 594 cases and obtain that

$$\operatorname{Var}\Big(\sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj'}\Big)=\frac{1}{n}\left\{\sum_{j=1}^{p}\beta_j^4\left[E\left(X_{1j}^4-1\right)\right]+2\sum_{j\neq j'}\beta_j^2\beta_{j'}^2\right\}+\frac{2p^2\sigma_Y^4}{n^3}+O\left(n^{-2}\right). \quad (55)$$

Lastly, plug-in (53) and (55) into (49) to get

$$\begin{aligned}\operatorname{Var}(T)&=\operatorname{Var}\left(\hat{\tau}^2\right)-4\operatorname{Cov}\left(\hat{\tau}^2,\sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj'}\right)+4\operatorname{Var}\left(\sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\psi}_{jj'}\right)\\&=\operatorname{Var}\left(\hat{\tau}^2\right)-4\left(\frac{2}{n}\left\{\sum_{j=1}^{p}\beta_j^4\left[E\left(X_{1j}^4-1\right)\right]+2\sum_{j\neq j'}\beta_j^2\beta_{j'}^2\right\}\right)\\&+4\left(\frac{1}{n}\left\{\sum_{j=1}^{p}\beta_j^4\left[E\left(X_{1j}^4-1\right)\right]+2\sum_{j\neq j'}\beta_j^2\beta_{j'}^2\right\}+\frac{2p^2\sigma_Y^4}{n^3}\right)+O\left(n^{-2}\right)\\&=\operatorname{Var}\left(\hat{\tau}^2\right)-\frac{4}{n}\left\{\sum_{j=1}^{p}\beta_j^4\left[E\left(X_{1j}^4-1\right)\right]+2\sum_{j\neq j'}\beta_j^2\beta_{j'}^2\right\}\\&\qquad+\frac{8p^2\sigma_Y^4}{n^3}+O(n^{-2})\\&=\operatorname{Var}\left(T_{oracle}\right)+\frac{4p^2\sigma_Y^4}{n^3}+O\left(n^{-2}\right),\end{aligned}$$

where the last equality holds by (12). □

Remark 4. ***Calculations for equation*** (16)*:*

Write,

$$\begin{aligned}Cov\left(\hat{\tau}^2,g_n\right)&=Cov\left(\frac{1}{n\left(n-1\right)}\sum_{i_1\neq i_2}\sum_{j=1}^{p}W_{i_1j}W_{i_2j},\frac{1}{n}\sum_{i=1}^{n}g_i\right)\\&=\frac{1}{n^2\left(n-1\right)}\sum_{i_1\neq i_2}\sum_{j=1}^{p}\sum_{i=1}^{n}E\left(W_{i_1j}W_{i_2j}g_i\right)\\&=\frac{2}{n^2\left(n-1\right)}\sum_{i_1\neq i_2}\sum_{j=1}^{p}E\left(W_{i_1j}g_{i_1}\right)E\left(W_{i_2j}\right)\end{aligned}$$

$$= \frac{2}{n^2(n-1)} \sum_{i_1 \neq i_2} \sum_{j=1}^{p} E(W_{i_1 j} g_{i_1}) \beta_j$$
$$= \frac{2}{n} \sum_{j=1}^{p} E(S_{ij}) \beta_j,$$

where $S_{ij} \equiv W_{ij} g_i$. *Also notice that* $Var(g_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} g_i\right) = \frac{Var(g_i)}{n}$. *Thus, by* (8) *we get*

$$c^* = \frac{Cov\left(\hat{\tau}^2, g_n\right)}{Var(g_n)} = \frac{2\sum_{j=1}^{p} E(S_{ij})\beta_j}{Var(g_i)}.$$

Remark 5. ***Calculations for Example 3:***

In order to calculate $Var(T_{c^*})$ *we need to calculate the numerator and denominator of* (17). *Consider first* $\theta_j \equiv E(S_{ij})$. *Write,*

$$\theta_j \equiv E(S_{ij}) = E(X_{ij} Y_i g_i) = E\left(X_{ij}\left(\beta^T X_i + \varepsilon_i\right) g_i\right)$$
$$= E\left(X_{ij}\left(\sum_{m=1}^{p} \beta_m X_{im} + \varepsilon_i\right) \sum_{k<k'} X_{ik} X_{ik'}\right)$$
$$= \sum_{m=1}^{p} \sum_{k<k'} \beta_m E(X_{ij} X_{im} X_{ik} X_{ik'}),$$

where in the last equality we used the assumption that $E(\epsilon|X) = 0$. *Since the columns of* $\mathbf{X}$ *are independent, the summation is not zero (up to permutations) when* $j = k$ *and* $m = k'$. *In this case we have*

$$\theta_j = \sum_{m=1}^{p} \sum_{k<k'} \beta_m E(X_{ij} X_{im} X_{ik} X_{ik'}) = \sum_{m \neq j}^{p} \beta_m E\left(X_{ij}^2 X_{im}^2\right)$$
$$= \sum_{m \neq j}^{p} \beta_m E\left(X_{ij}^2\right) E\left(X_{im}^2\right) = \sum_{m \neq j}^{p} \beta_m.$$

Notice that in the fourth equality we used the assumption that $E(X_{ij}^2) = 1$ *for all* $j = 1, ..., p$. *Thus,*

$$\sum_{j=1}^{p} \beta_j E(S_{ij}) = \sum_{j=1}^{p} \beta_j \sum_{m \neq j}^{p} \beta_m = \sum_{j=1}^{p} \beta_j \left(\sum_{m=1}^{p} \beta_m - \beta_j\right)$$
$$= \left(\sum_{j=1}^{p} \beta_j\right)^2 - \sum_{j=1}^{p} \beta_j^2 = \left(\sum_{j=1}^{p} \beta_j\right)^2 - \tau^2. \quad (56)$$

Plug-in $\tau^2 = 1$ *and* $\beta_j = \frac{1}{\sqrt{p}}$ *to get the numerator of* (17)*:*

$$\left[2\sum_{j=1}^{p}\beta_j E\left(S_{ij}\right)\right]^2 = 4\left[\left(\sum_{j=1}^{p}\beta_j\right)^2 - \tau^2\right]^2 = 4\left[\left(p\frac{1}{\sqrt{p}}\right)^2 - 1\right]^2 = 4(p-1)^2.$$

Consider now the denominator of (17)*. Write,*

$$Var(g_i) = E\left(g_i^2\right) = E\left[\left(\sum_{j<j'} X_{ij}X_{ij'}\right)^2\right] = \sum_{j_1<j_2}\sum_{j_3<j_4} E\left(X_{ij_1}X_{ij_2}X_{ij_3}X_{ij_4}\right).$$

Since we assume that the columns of $\boldsymbol{X}$ *are independent, the summation is not zero when* $j_1 = j_3$ *and* $j_2 = j_4$*. Thus,*

$$Var(g_i) = \sum_{j_1<j_2} E\left(X_{ij_1}^2 X_{ij_2}^2\right) = \sum_{j_1<j_2} E\left(X_{ij_1}^2\right) E\left(X_{ij_2}^2\right) = p\left(p-1\right)/2. \quad (57)$$

Notice that we used the assumption that since we assume that $\boldsymbol{\Sigma} = \mathbf{I}$ *in the last equality. Now, recall by* (48) *that* $Var\left(\hat{\tau}^2\right) = \frac{20}{n} + O\left(\frac{1}{n^2}\right)$*. Therefore, we have*

$$\begin{aligned} Var\left(T_{c^*}\right) &= Var\left(\hat{\tau}^2\right) - \frac{\left[2\sum_{j=1}^{p}\beta_j E\left(S_{ij}\right)\right]^2}{n\,Var\left(g_i\right)} \\ &= \frac{20}{n} + O\left(\frac{1}{n^2}\right) - \frac{4(p-1)^2}{n\cdot\left[p\left(p-1\right)/2\right]} = \frac{12}{n} + O\left(\frac{1}{n^2}\right), \end{aligned} \quad (58)$$

where we used the assumption that $n = p$ *in the last equality.*

Proof of Proposition 4.

We need to prove that $\sqrt{n}\left[T_{c^*} - T_{\hat{c}^*}\right] \xrightarrow{p} 0$. Write,

$$\sqrt{n}\left[T_{c^*} - T_{\hat{c}^*}\right] = \sqrt{n}\left[\hat{\tau}^2 - c^* g_n - \left(\hat{\tau}^2 - \hat{c}^* g_n\right)\right] = \sqrt{n} g_n (\hat{c}^* - c^*).$$

By Markov and Cauchy-Schwarz inequalities, it is enough to show that

$$\begin{aligned} P\left\{\left|\sqrt{n}g_n\left(\hat{c}^* - c^*\right)\right| > \varepsilon\right\} &\leq \frac{E\left\{\left|\sqrt{n}g_n\left(\hat{c}^* - c^*\right)\right|\right\}}{\varepsilon} \\ &\leq \frac{\sqrt{nE\left(g_n^2\right)E\left[\left(\hat{c}^* - c^*\right)^2\right]}}{\varepsilon} = \frac{\sqrt{\mathrm{Var}(g)\mathrm{Var}(\hat{c}^*)}}{\varepsilon} \to 0. \end{aligned} \quad (59)$$

Notice that by (18) we have

$$\mathrm{Var}\left(g\right)\mathrm{Var}\left(\hat{c}^*\right) = \frac{\mathrm{Var}\left(U\right)}{\mathrm{Var}\left(g\right)}, \quad (60)$$

where $U \equiv \frac{2}{n(n-1)} \sum_{i_1 \neq i_2} \sum_{j=1}^{p} W_{i_1 j} W_{i_2 j} g_{i_2}$. The variance of g is

$$\begin{aligned}\mathrm{Var}(g) = E(g^2) = E\left[\left(\sum_{j<j'} X_{ij} X_{ij'}\right)^2\right] &= \sum_{j_1<j_2} \sum_{j_3<j_4} E\left(X_{ij_1} X_{ij_2} X_{ij_3} X_{ij_4}\right) \\ = \sum_{j_1<j_2} E\left(X_{ij_1}^2 X_{ij_2}^2\right) &= \sum_{j_1<j_2} E\left(X_{ij_1}^2\right) E\left(X_{ij_2}^2\right) = p\left(p-1\right)/2, \end{aligned} \tag{61}$$

where the equation above holds since we assume that $\mathbf{\Sigma} = \mathbf{I}$ and that the columns of $\mathbf{X}$ are independent. Hence, by (59)–(61) it enough to prove $\frac{\mathrm{Var}(U)}{p^2} \to 0$.

The variance of U is

$$\begin{aligned}\mathrm{Var}(U) &= \mathrm{Var}\left[\frac{2}{n\left(n-1\right)} \sum_{i_1 \neq i_2} \sum_{j=1}^{p} W_{i_1 j} W_{i_2 j} g_{i_2}\right] \\ &= \frac{4}{n^2 (n-1)^2} \sum_{j,j'}^{p} \sum_{i_1 \neq i_2, i_3 \neq i_4} \mathrm{Cov}\left[W_{i_1 j} W_{i_2 j} g_{i_2}, W_{i_3 j'}. W_{i_4 j'} g_{i_4}\right] \end{aligned} \tag{62}$$

The covariance in (62) is different from zero in the two following cases:

1. When $\{i_1, i_2\}$ equals to $\{i_3, i_4\}$.
2. When one of $\{i_1, i_2\}$ equals to $\{i_3, i_4\}$ while the other is different.

The first condition includes two different sub-cases and each of those consists $n(n-1)$ quadruples (i_1, i_2, i_3, i_4) that satisfy the condition. Similarly, the second condition above includes four different sub-cases and each of those consists $n(n-1)(n-2)$ quadruples that satisfy the condition.

We now calculate the covariance for all these six sub-cases.

(1) The covariance when $i_1 = i_3, i_2 = i_4$ is

$$\begin{aligned}\delta_1 &\equiv \mathrm{Cov}\left[W_j \tilde{W}_j \tilde{g}, W_{j'} \tilde{W}_{j'} \tilde{g}\right] \\ &= E\left(W_j W_{j'}\right) E\left[\tilde{W}_j \tilde{W}_{j'} \tilde{g}^2\right] - E\left(W_j\right) E\left[\tilde{W}_j \tilde{g}\right] E\left(\tilde{W}_{j'}\right) E\left[\tilde{W}_{j'} \tilde{g}\right] \\ &= E\left(W_j W_{j'}\right) E\left[\tilde{W}_j \tilde{W}_{j'} \tilde{g}^2\right] - \beta_j \beta_{j'} \theta_j \theta_{j'}, \end{aligned} \tag{63}$$

where $\tilde{W}$ and $\tilde{g}$ are independent copies of W and g respectively.

(2) The covariance when and $i_1 = i_4, i_2 = i_3$ is

$$\begin{aligned}\delta_2 &\equiv \mathrm{Cov}\left[W_j \tilde{W}_j \tilde{g}, \tilde{W}_{j'} W_{j'} g\right] \\ &= E\left[W_j W_{j'} g\right] E\left[\tilde{W}_j \tilde{W}_{j'} \tilde{g}\right] - E\left(W\right) E\left[\tilde{W}_j \tilde{g}\right] E\left(\tilde{W}_{j'}\right) E\left[W_{j'} g\right] \\ &= \left\{E\left[W_j W_{j'} g\right]\right\}^2 - \beta_j \beta_{j'} \theta_j \theta_{j'}. \end{aligned} \tag{64}$$

(3) The covariance when $i_1 = i_3, i_2 \neq i_4$ is

$$\delta_3 \equiv \mathrm{Cov}\left[W_j \tilde{W}_j \tilde{g}, W_{j'} \tilde{\tilde{W}}_{j'} \tilde{\tilde{g}}\right] = E\left(W_j W_{j'}\right) \theta_j \theta_{j'} - \beta_j \beta_{j'} \theta_j \theta_{j'}, \tag{65}$$

where $\tilde{\tilde{W}}$ and $\tilde{\tilde{g}}$ are another independent copies of W and g respectively.

(4) The covariance when $i_1 = i_4, i_2 \neq i_3$ is

$$\delta_4 \equiv \mathrm{Cov}\left[W_j\tilde{W}_j\tilde{g}, \tilde{\tilde{W}}_{j'}W_{j'}g\right] = \theta_j\beta_{j'}E\left[W_jW_{j'}g\right] - \beta_j\beta_{j'}\theta_j\theta_{j'}. \tag{66}$$

(5) The covariance when $i_2 = i_3, i_1 \neq i_4$ is similar to δ_4, i.e.,

$$\delta_5 \equiv \mathrm{Cov}\left[W_j\tilde{W}_j\tilde{g}, \tilde{W}_{j'}\tilde{\tilde{W}}_{j'}\tilde{\tilde{g}}\right] = \beta_j\theta_{j'}E\left[\tilde{W}_j\tilde{W}_{j'}\tilde{g}\right] - \beta_j\beta_{j'}\theta_j\theta_{j'}. \tag{67}$$

(6) The covariance when $i_2 = i_4, i_1 \neq i_3$ is

$$\delta_6 \equiv \mathrm{Cov}\left[W_j\tilde{W}_j\tilde{g}, \tilde{\tilde{W}}_{j'}\tilde{W}_{j'}\tilde{g}\right] = \beta_j\beta_{j'}E\left[\tilde{W}_j\tilde{W}_{j'}\tilde{g}^2\right] - \beta_j\beta_{j'}\theta_j\theta_{j'} \tag{68}$$

Thus, plugging-in (63)–(68) into (62) gives

$$\mathrm{Var}\,(U) = 4\sum_{j,j'}\left\{\frac{1}{n\,(n-1)}\,(\delta_1+\delta_2) + \frac{(n-2)}{n\,(n-1)}\,(\delta_3+\delta_4+\delta_5+\delta_6)\right\}. \tag{69}$$

Recall that we wish to show that $\frac{\mathrm{Var}(U)}{p^2} \to 0$. Since we assume that $n/p = O(1)$, it is enough to show that

$\sum_{j,j'}\frac{(\delta_1+\delta_2)}{n^4} \to 0$ and $\sum_{j,j'}\frac{(\delta_3+\delta_4+\delta_5+\delta_6)}{n^3} \to 0$. Careful calculations, which are not presented here, show that under the linear model when the covariates are independent

$$\sum_{j,j'}\delta_1 \leq C^2\tau^2\sigma^2p^2\frac{p-1}{2},\ \sum_{j,j'}\delta_2 \leq C^2\tau^3p^3 \quad \text{and,}$$

$$\sum_{j,j'}\delta_3 \leq p^2\tau^2, \sum_{j,j'}\delta_4 \leq C\tau^4p^2,\ \sum_{j,j'}\delta_5 \leq C\tau^2(\tau^2+\sigma^2)\frac{p(p-1)}{2},$$

where C is a bound on $E(X_{j_1}^2X_{j_2}^2X_{j_3}^2X_{j_4}^2)$. It follows that $\frac{\mathrm{Var}(U)}{p^2} \to 0$ because $n/p = O(1)$, which completes the proof of the proposition. □

Remark 6. *We now calculate the asymptotic improvement of $T_{\mathbf{B}}$ over the naive estimator. For simplicity, consider the case when $\tau^2 = \sigma^2 = 1$. Recall the variance of $\hat{\tau}^2$ and $T_{\mathbf{B}}$ given in (6) and (21), respectively. Write,*

$$\begin{aligned}
&\lim_{n,p\to\infty}\frac{Var\left(\hat{\tau}^2\right) - Var\left(T_{\mathbf{B}}\right)}{Var\left(\hat{\tau}^2\right)} \\
&= \lim_{n,p\to\infty}\frac{8\tau_{\mathbf{B}}^4/n}{\frac{4}{n}\left[\frac{(n-2)}{(n-1)}\left(\sigma_Y^2\tau^2+\tau^4\right) + \frac{1}{2(n-1)}\left(p\sigma_Y^4 + 4\sigma_Y^2\tau^2 + 3\tau^4\right)\right]} \\
&= \frac{2\tau_{\mathbf{B}}^4}{3\tau^4 + \frac{4p\tau^4+4\sigma_Y^2\tau^2+3\tau^4}{2n}} = \frac{0.5}{3+2\frac{p}{n}},
\end{aligned}$$

where we used (6) in the first equality, and the fact that $\sigma_Y^2 = 2\tau^2 = 2$ in the second equality. Now, notice that when $p = n$ and $\tau_{\mathbf{B}}^2 = 0.5$ then the reduction is $\frac{0.5}{3+2} = 10\%$ and when p/n converges to zero, the reduction is 16%.

Remark 7. ***Calculations for Example 4:***

Consider the first scenario where $\beta_j^2 = \frac{1}{p}$. *Recall that we assume that the set* $\boldsymbol{B}$ *is a fixed set of indices such that* $|\mathbf{B}| \ll p$. *Therefore, we have* $\tau_{\mathbf{B}}^2 = \sum_{j\in\mathbf{B}} \beta_j^2 = O\left(\frac{1}{p}\right)$. *Now, by (22) we have* $Var(T_{\boldsymbol{B}}) = Var(\hat{\tau}^2) - \frac{8}{n}\tau_{\boldsymbol{B}}^2 + O(n^{-2})$ *and by Remark 5 we have* $Var(\hat{\tau}^2) = \frac{20}{n} + O(n^{-2})$. *Using the assumption that* $n = p$ *we can conclude that* $Var(T_{\mathbf{B}}) = \frac{20}{n} + O\left(\frac{1}{n^2}\right)$. *Hence, in this scenario,* $T_{\boldsymbol{B}}$ *and the naive estimator have the same asymptotic variance. In contrast, recall that in Example 3 we showed that the asymptotic variance of* T_{c^*} *is* 40% *lower than the variance of the naive estimator.*

Consider now the second scenario where $\hat{\tau}_{\mathbf{B}}^2 = \tau^2 = 1$. *By (22) we have*

$$Var(T_{\mathbf{B}}) = Var\left(\hat{\tau}^2\right) - \frac{8}{n}\tau_{\mathbf{B}}^4 + O(n^{-2}) = \frac{12}{n} + O(n^{-2}).$$

Hence, in this scenario the asymptotic variance of $T_{\boldsymbol{B}}$ *is* 40% *smaller than the variance of the naive estimator. Consider now* $Var(T_{c^*})$. *By Cauchy–Schwarz inequality* $\left(\sum_{j\in\mathbf{B}} \beta_j\right)^2 \le \sum_{j\in\mathbf{B}} \beta_j^2 \cdot |\mathbf{B}| = \tau_{\boldsymbol{B}}^2 |\mathbf{B}| = O(1)$, *where the last equality holds since we assume that* $\mathbf{B} \subset \{1, ..., p\}$ *be a fixed set of some indices such that* $|\mathbf{B}| \ll p$. *Now, By (56) we have*

$$\sum_{j=1}^{p} \beta_j \theta_j = \left(\sum_{j=1}^{p} \beta_j\right)^2 - \tau^2 = \left(\sum_{j\in\mathbf{B}}^{p} \beta_j + \overbrace{\sum_{j\notin\mathbf{B}}^{p} \beta_j}^{0}\right)^2 - \tau_{\mathbf{B}}^2 \le |\mathbf{B}| - 1 = O(1).$$

Now, recall that $Var\left(\hat{\tau}^2\right) = \frac{20}{n} + O\left(\frac{1}{n^2}\right)$ *and* $Var(g_i) = p(p-1)/2$ *by (48) and (57) respectively. Therefore, we have*

$$Var(T_{c^*}) = Var\left(\hat{\tau}^2\right) - \frac{\left[2\sum_{j=1}^{p} \beta_j \theta_j\right]^2}{n\, Var(g_i)} = \frac{20}{n} + O\left(\frac{1}{n^2}\right) - O\left(\frac{1}{np^2}\right) = \frac{20}{n} + O\left(\frac{1}{n^2}\right),$$

Hence, in this scenario, T_{c^*} *and the naive estimator have the same asymptotic variance.*

Lastly, recall that in Example 2 we already showed that, asymptotically, the variance of T_{oracle} *(i.e., the optimal oracle estimator) is* 40% *lower than the naive variance (without any assumptions about the structure of the coefficient vector* β*).*

Proof of Proposition 5.

In order to prove that $\sqrt{n}(T_\gamma - T_{\mathbf{B}}) \xrightarrow{p} 0$, it is enough to show that

$$E\left\{\sqrt{n}(T_\gamma - T_{\mathbf{B}})\right\} \to 0, \tag{70}$$

$$\mathrm{Var}\left\{\sqrt{n}\left(T_\gamma - T_{\mathbf{B}}\right)\right\} \to 0. \tag{71}$$

We start with the first equation. Let A denote the event that the selection algorithm γ perfectly identifies the set of large coefficients, i.e., $A = \{\mathbf{B}_\gamma = \mathbf{B}\}$. Let $p_A \equiv P(A)$ denote the probability that A occurs, and let $\mathbb{1}_A$ denote the indicator of A. Notice that $E(T_{\mathbf{B}}) = \tau^2$ and $T_\gamma \mathbb{1}_A = T_{\mathrm{B}} \mathbb{1}_A$. Thus,

$$\begin{aligned} &E\left\{\sqrt{n}\left(T_\gamma - T_{\mathbf{B}}\right)\right\} = \sqrt{n}\left[E\left(T_\gamma\right) - \tau^2\right] \\ &= \sqrt{n}\left(E\left[T_\gamma(1-\mathbb{1}_A)\right] + E\left[T_\gamma \mathbb{1}_A\right] - \tau^2\right) \\ &= \sqrt{n}E\left[T_\gamma(1-\mathbb{1}_A)\right] + \sqrt{n}\left[E\left(T_{\mathbf{B}}\mathbb{1}_A\right) - \tau^2\right], \end{aligned} \tag{72}$$

where the last equality holds since $T_\gamma \mathbb{1}_A = T_{\mathrm{B}} \mathbb{1}_A$. For the convenience of notation, let C be an upper bound of the maximum over all first four moments of T_γ and $T_{\mathbf{B}}$, and consider the first term of (72). By the Cauchy–Schwarz inequality,

$$\begin{aligned} \sqrt{n}E\left[T_\gamma\left(1-\mathbb{1}_A\right)\right)] &\le \sqrt{n}\left\{E\left[T_\gamma^2\right]\right\}^{1/2}\left\{E\left[\left(1-\mathbb{1}_A\right)^2\right]\right\}^{1/2} \\ &\le \sqrt{n}C^{1/2}\left\{1-p_A\right\}^{1/2} \underset{n\to\infty}{\to} 0, \end{aligned} \tag{73}$$

where the last inequality holds since $\lim_{n\to\infty} n\left(1-p_A\right)^{1/2} = 0$ by assumption. We now consider the second term of (72). Write,

$$\sqrt{n}\left[E\left(T_{\mathbf{B}}\mathbb{1}_A\right) - \tau^2\right] = \sqrt{n}E\left(T_{\mathbf{B}}\mathbb{1}_A - T_{\mathbf{B}}\right) = -\sqrt{n}E\left[T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right],$$

and notice that by the same type of argument as in (73) we have

$$\sqrt{n}E\left[T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right] \underset{n\to\infty}{\to} 0.$$

This completes the proof of (70).
We now move to show that $\mathrm{Var}\left\{\sqrt{n}\left(T_\gamma - T_{\mathbf{B}}\right)\right\} \to 0$. Write,

$$\begin{aligned} &\mathrm{Var}\left\{\sqrt{n}\left(T_\gamma - T_{\mathbf{B}}\right)\right\} = n\mathrm{Var}\left(T_\gamma - T_{\mathbf{B}}\right) \\ &= n\left[\mathrm{Var}\left(T_\gamma\right) + \mathrm{Var}\left(T_{\mathbf{B}}\right) - 2\mathrm{Cov}\left(T_\gamma, T_{\mathbf{B}}\right)\right] \\ &= n\left\{E\left(T_\gamma^2\right) - \left[E\left(T_\gamma\right)\right]^2 + E\left(T_{\mathbf{B}}^2\right) - \tau^4 - 2\left[E\left(T_\gamma T_{\mathbf{B}}\right) - E\left(T_\gamma\right)\tau^2\right]\right\} \\ &= n\{E\left(T_\gamma^2\right) - E\left(T_\gamma T_{\mathbf{B}}\right) + E\left(T_{\mathbf{B}}^2\right) - E\left(T_\gamma T_{\mathbf{B}}\right) \\ &\qquad + E\left(T_\gamma\right)\left[\tau^2 - E\left(T_\gamma\right)\right] - \tau^2\left[\tau^2 - E\left(T_\gamma\right)\right]\} \\ &= n\left\{\underbrace{E\left(T_\gamma^2\right) - E\left(T_\gamma T_{\mathbf{B}}\right)}_{\theta_1} + \underbrace{E\left(T_{\mathbf{B}}^2\right) - E\left(T_\gamma T_{\mathbf{B}}\right)}_{\theta_2} - \underbrace{\left[\tau^2 - E\left(T_\gamma\right)\right]^2}_{\theta_3}\right\} \end{aligned}$$

Thus, it is enough to show that $n\theta_1 \to 0$, $n\theta_2 \to 0$ and $n\theta_3 \to 0$.

We start with showing that $n\theta_1 \to 0$. Notice that $T_{\mathbf{B}}^2 \mathbb{1}_A = T_{\mathbf{B}} T_\gamma \mathbb{1}_A = T_\gamma^2 \mathbb{1}_A$. Thus,

$$n\theta_1 = n\left\{E\left(T_\gamma^2\right) - E\left(T_\gamma T_{\mathbf{B}}\right)\right\}$$

$$
\begin{aligned}
&= n\left\{E\left(T_\gamma^2\right) - E\left[T_\gamma T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right] - E\left(T_\gamma T_{\mathbf{B}}\mathbb{1}_A\right)\right\} \\
&= n\left\{E\left(T_\gamma^2\right) - E\left[T_\gamma T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right] - E\left(T_\gamma^2\mathbb{1}_A\right)\right\} \\
&= n\left\{E\left[T_\gamma^2\left(1-\mathbb{1}_A\right)\right] - E\left[T_\gamma T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right]\right\}.
\end{aligned}
$$

Now, notice that $n(E\left[T_\gamma^2\left(1-\mathbb{1}_A\right)\right]) \to 0$ by similar arguments as in (73), with a slight modification of using the existence of the fourth moments of T_γ and $T_{\mathbf{B}}$, rather than the second moments. Also, by Cauchy–Schwarz inequality we have,

$$
\begin{aligned}
nE\left[T_\gamma T_{\mathbf{B}}\left(1-1_A\right)\right] &\le n\left\{E\left(T_\gamma^2 T_{\mathbf{B}}^2\right)\right\}^{1/2}\left\{E\left[\left(1-1_A\right)^2\right]\right\}^{1/2} \\
&\le n\left\{E\left(T_\gamma^4\right)E\left(T_{\mathbf{B}}^4\right)\right\}^{1/4}\left\{1-p_A\right\}^{1/2} \\
&\le nC^{1/2}\left\{1-p_A\right\}^{1/2} \to 0,
\end{aligned}
$$

where C is an upper bound of the maximum over all first four moments of T_γ and $T_{\mathbf{B}}$. Therefore, $n\theta_1 \to 0$.

Consider now $n\theta_2$. Write,

$$
\begin{aligned}
n\theta_2 &= n\left\{E\left(T_{\mathbf{B}}^2\right) - E\left(T_\gamma T_{\mathbf{B}}\right)\right\} \\
&= n\left\{E\left(T_{\mathbf{B}}^2\right) - E\left[T_\gamma T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right] - E\left(T_\gamma T_{\mathbf{B}}\mathbb{1}_A\right)\right\} \\
&= n\left\{E\left(T_{\mathbf{B}}^2\right) - E\left[T_\gamma T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right] - E\left(T_{\mathbf{B}}^2\mathbb{1}_A\right)\right\} \\
&= n\left\{E\left[T_{\mathbf{B}}^2\left(1-\mathbb{1}_A\right)\right] - E\left[T_\gamma T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right]\right\} \to 0,
\end{aligned}
$$

and notice that the last equation follows by similar arguments.

Consider now $n\theta_3$. Write,

$$
\begin{aligned}
n\theta_3 &= n\left[\tau^2 - E\left(T_\gamma\right)\right] \\
&= n\left[E\left(T_{\mathbf{B}}\right) - E\left(T_\gamma\right)\right] \\
&= n\left[E\left[T_{\mathbf{B}}\left(1-\mathbb{1}_A\right) + T_{\mathbf{B}}\mathbb{1}_A\right] - E\left(T_\gamma\right)\right] \\
&= n\left\{E\left[T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right] + E\left(T_{\mathbf{B}}\mathbb{1}_A - T_\gamma\right)\right\} \\
&= n\left\{E\left[T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right] + E\left(T_\gamma\mathbb{1}_A - T_\gamma\right)\right\} \\
&= n\left\{E\left[T_{\mathbf{B}}\left(1-\mathbb{1}_A\right)\right] - E\left[T_\gamma\left(1-\mathbb{1}_A\right)\right]\right\} \to 0,
\end{aligned}
$$

where the last equation follows by similar arguments as in (73). This completes the proof of (71) and we conclude that $\sqrt{n}\left(T_\gamma - T_{\mathbf{B}}\right) \xrightarrow{p} 0$. □

Proof of Proposition 6.
We wish to prove that

$$
n\left[\widehat{\mathrm{Var}\left(\hat{\tau}^2\right)} - \mathrm{Var}\left(\hat{\tau}^2\right)\right] \xrightarrow{p} 0. \tag{74}
$$

Recall by (6) that

$$
\mathrm{Var}\left(\hat{\tau}^2\right) = \frac{4\left(n-2\right)}{n\left(n-1\right)}\left[\beta^T\mathbf{A}\beta - \|\beta\|^4\right] + \frac{2}{n\left(n-1\right)}\left[\|\mathbf{A}\|_F^2 - \|\beta\|^4\right].
$$

Now, when we assume standard Gaussian covariates, one can verify that $\beta^T\mathbf{A}\beta - \|\beta\|^4 = \sigma_Y^2\tau^2 + \tau^4$ and $\|\mathbf{A}\|_F^2 - \|\beta\|^4 = p\sigma_Y^4 + 4\sigma_Y^2\tau^2 + 3\tau^4$, where $\sigma_Y^2 = \sigma^2 + \tau^2$. Thus, in this case we can write

$$\mathrm{Var}\left(\hat{\tau}^2\right) = \frac{4}{n}\left[\frac{(n-2)}{(n-1)}\left[\sigma_Y^2\tau^2 + \tau^4\right] + \frac{1}{2\,(n-1)}\left(p\sigma_Y^4 + 4\sigma_Y^2\tau^2 + 3\tau^4\right)\right]. \quad (75)$$

In order to prove that (74) holds, it is enough to prove the consistency of $\hat{\tau}^2$ and $\hat{\sigma}_Y^2$. Consistency of the sample variance $\hat{\sigma}_Y^2$ is a standard result, and since $\hat{\tau}^2$ is an unbiased estimator, it is enough to show that its variance converges to zero as $n \to \infty$. Since we assume $\hat{\tau}^2 + \sigma^2 = O(1)$ and $p/n = O(1)$, we have by (47) that $\mathrm{Var}(\hat{\tau}^2) \underset{n\to\infty}{\to} 0$, and (74) follows. □

Proof of Proposition 7.
We now move to prove that

$$n\left[\widehat{\mathrm{Var}\left(T_\gamma\right)} - \mathrm{Var}\left(T_\gamma\right)\right] \xrightarrow{p} 0, \quad (76)$$

Recall that by Proposition 5 we have $\lim_{n\to\infty} n\left[\mathrm{Var}\left(T_{\mathbf{B}}\right) - \mathrm{Var}\left(T_\gamma\right)\right] = 0$. Hence, it is enough to show that

$$n\left[\widehat{\mathrm{Var}\left(T_\gamma\right)} - \mathrm{Var}\left(T_{\mathbf{B}}\right)\right] \xrightarrow{p} 0.$$

Since we assume $X_i \overset{i.i.d}{\sim} N\left(\mathbf{0}, \mathbf{I}\right)$ then by (21) we have

$$\mathrm{Var}\left(T_{\mathbf{B}}\right) = \mathrm{Var}\left(\hat{\tau}^2\right) - \frac{8}{n}\tau_{\mathbf{B}}^4 + O(n^{-2}).$$

Recall that by definition we have $\widehat{\mathrm{Var}\left(T_\gamma\right)} = \widehat{\mathrm{Var}\left(\hat{\tau}^2\right)} - \frac{8}{n}\hat{\tau}_{\mathbf{B}_\gamma}^4$. Also recall that $\widehat{\mathrm{Var}(\hat{\tau}^2)}$ is consistent by Proposition 6. Thus, it is enough to prove that $\hat{\tau}_{\mathbf{B}_\gamma}^2 - \tau_{\mathbf{B}}^2 \xrightarrow{p} 0$. Now, since we assumed that $n\left[P\left(\{\mathbf{B}_\gamma \neq \mathbf{B}\}\right)\right]^{1/2} \xrightarrow[n\to\infty]{} 0$ then clearly $P\left(\mathbf{B}_\gamma = \mathbf{B}\right) \xrightarrow[n\to\infty]{} 1$. Thus, it is enough to show that $\hat{\tau}_{\mathbf{B}}^2 - \tau_{\mathbf{B}}^2 \xrightarrow{p} 0$. Recall that $E(\hat{\beta}_j^2) = \beta_j^2$ for $j = 1, ..., p$ and notice that $\mathrm{Var}(\hat{\beta}_j^2) \underset{n\to\infty}{\to} 0$ by similar arguments that were used to derive (6). Hence, we have $\hat{\beta}_j^2 - \beta_j^2 \xrightarrow{p} 0$. Since we assumed that $\mathbf{B}$ is finite, we have

$$\hat{\tau}_{\mathbf{B}}^2 - \tau_{\mathbf{B}}^2 = \sum_{j\in\mathbf{B}}\left(\hat{\beta}_j^2 - \beta_j^2\right) \xrightarrow{p} 0,$$

and (76) follows. □

Remark 8. *We use the the following simple selection algorithm γ:*

Algorithm 3: Covariate selection γ

Input: A dataset $(\mathbf{X}_{n\times p}, \mathbf{Y}_{n\times 1})$.

1. Calculate $\hat{\beta}_1^2, ..., \hat{\beta}_p^2$ where $\hat{\beta}_j^2$ is given in (4) for $j = 1, ..., p$.
2. Calculate the differences $\lambda_j = \hat{\beta}_{(j)}^2 - \hat{\beta}_{(j-1)}^2$ for $j = 2, \ldots, p$ where $\hat{\beta}_{(1)}^2 < \hat{\beta}_{(2)}^2 < ... < \hat{\beta}_{(p)}^2$ denotes the order statistics.
3. Select the covariates $\mathbf{B}_\gamma = \left\{ j : \hat{\beta}_{(j)}^2 > \hat{\beta}_{(j^*)}^2 \right\}$, where $j^* = \arg\max_j \lambda_j$.

Result: Return $\mathbf{B}_\gamma$.

The algorithm above finds the largest gap between the ordered estimated squared coefficients and then uses this gap as a threshold to select a set of coefficients $\mathbf{B}_\gamma \subset \{\mathbf{1}, ..., \mathbf{p}\}$. The algorithm works well in scenarios where a relatively large gap truly separates between larger coefficients and the smaller coefficients of the vector β.

Remark 9. *The following algorithm is used to construct the Selection estimator that improves an initial estimator of τ^2, as presented in Table 3.*

Algorithm 4: Empirical Estimator

Input: A dataset $(\mathbf{X}_{n\times p}, Y_{n\times 1})$, an estimation procedure $\tilde{\tau}^2$, and a covariate-selection procedure δ.

1. Apply the procedure δ to the dataset $(\mathbf{X}_{n\times p}, Y_{n\times 1})$ to obtain $\mathbf{B}_\delta$.
2. Apply the procedure $\tilde{\tau}^2$ to the dataset $(\mathbf{X}_{n\times p}, Y_{n\times 1})$.
3. Calculate the zero-estimator $Z_h = \frac{1}{n}\sum_{i=1}^n h(X_i)$, where $h(X_i) = \sum_{j<j'\in \mathbf{B}_\delta} X_{ij}X_{ij'}$.
4. **Bootstrap step:**
   - Sample n observations at random from $(\mathbf{X}_{n\times p}, Y_{n\times 1})$, with replacement, to obtain a bootstrap dataset.
   - Repeat steps 2 and 3 based on the bootstrap dataset.

   The bootstrap step is repeated M times in order to produce $(\tilde{\tau}^2)^{*1}, ..., (\tilde{\tau}^2)^{*M}$ and $Z_h^{*1}, ..., Z_h^{*M}$.
5. Approximate the coefficient $\tilde{c}_h^* = \frac{\widehat{\mathrm{Cov}\left(\tilde{\tau}^2, Z_h\right)}}{\mathrm{Var}(Z_h)}$ where $\widehat{\mathrm{Cov}}(\cdot)$ denotes the empirical covariance from the bootstrap samples.

Output: Return the empirical estimator $T_{\tilde{h}} \equiv \tilde{\tau}^2 - \tilde{c}_h^* Z_h$.

References

[1] Bahadur, R. (1957). On unbiased estimates of uniformly minimum variance. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **18** 211–224. MR0092319

[2] Bonnet, A., Gassiat, E. and Lévy-Leduc, C. (2015). Heritability estimation in high dimensional sparse linear mixed models. *Electronic Journal of Statistics* **9** 2099–2129. MR3400534

[3] Borogovac, T. and Vakili, P. (2008). Control variate technique: A constructive approach. In *2008 Winter Simulation Conference* 320–327. IEEE.

[4] Candes, E., Fan, Y., Janson, L. and Lv, J. (2017). Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection. MR3798878

[5] Chen, H. Y. (2022). Statistical Inference on Explained Variation in High-dimensional Linear Model with Dense Effects. *arXiv preprint arXiv:2201.08723*.

[6] Collazos, J. A., Dias, R. and Zambom, A. Z. (2016). Consistent variable selection for functional regression models. *Journal of Multivariate Analysis* **146** 63–71. MR3477649

[7] de Los Campos, G., Sorensen, D. and Gianola, D. (2015). Genomic heritability: what is it? *PLoS Genetics* **11** e1005048.

[8] Deng, S., Ning, Y., Zhao, J. and Zhang, H. (2020). Optimal Semi-supervised Estimation and Inference for High-dimensional Linear Regression. *arXiv preprint arXiv:2011.14185*.

[9] Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269–284. MR3215347

[10] Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 37–65. MR2885839

[11] Fisher, N. (1982). Unbiased estimation for some non-parametric families of distributions. *The Annals of Statistics* **10** 603–615. MR0653535

[12] Glynn, P. W. and Szechtman, R. (2002). Some new perspectives on the method of control variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000* 27–49. Springer. MR1958845

[13] Guo, Z., Wang, W., Cai, T. T. and Li, H. (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association* **114** 358–369. MR3941260

[14] Hastie, T. and Qian, J. (2014). Glmnet vignette. *Retrieved June* **9** 1–30.

[15] Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with Sparsity. The Lasso and Generalizations*. Chapman and Hall. MR3616141

[16] Hoeffding, W. (1977). Some Incomplete and Boundedly Complete Families of Distributions. *The Annals of Statistics* **5** 278–291. MR0443200

[17] Janson, L., Barber, R. F. and Candes, E. (2017). EigenPrism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 1037–1065. MR3689308

[18] Kong, W. and Valiant, G. (2018). Estimating learnability in the sublinear data regime. *Advances in Neural Information Processing Systems* **31** 5455–5464.
[19] Lavenberg, S. S. and Welch, P. D. (1981). A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Science* **27** 322–335. MR0720496
[20] Lehmann, E. L. and Casella, G. (1998). Theory of point estimation. https://www.springer.com/us/book/9780387985022.
[21] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer Science & Business Media. MR1639875
[22] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.
[23] Nayak, T. K. and Sinha, B. (2012). Some aspects of minimum variance unbiased estimation in presence of ancillary statistics. *Statistics & Probability Letters* **82** 1129–1135. MR2915079
[24] Oda, R., Yanagihara, H. et al. (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electronic Journal of Statistics* **14** 1386–1412. MR4080281
[25] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166
[26] Taylor, J. and Tibshirani, R. (2018). Post-selection inference forpenalized likelihood models. *Canadian Journal of Statistics* **46** 41–61. MR3767165
[27] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242
[28] Tony Cai, T. and Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82** 391–419. MR4084169
[29] van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press. MR1652247
[30] Verzelen, N., Gassiat, E. et al. (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli* **24** 3683–3710. MR3788186
[31] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42** 565–569.
[32] Zambom, A. Z. and Kim, J. (2018). Consistent significance controlled variable selection in high-dimensional regression. *Stat* **7** e210. MR3905866
[33] Zhu, H. and Zhou, X. (2020). Statistical methods for SNP heritability estimation and partition: a review. *Computational and Structural Biotechnology Journal* **18** 1557–1568.